CADERNOS DO IME – Série Estatística

Universidade do Estado do Rio de Janeiro - UERJ
ISSN on-line 2317-4536 / ISSN impresso 1413-9022 - v.52, p.33-64, 2022
DOI:10.12957/cadest.2022.72568

O DESEMPENHO DOS ALUNOS DO ENSINO MÉDIO NO ENEM 2019 E A DESIGUALDADE SOCIAL: UM ESTUDO UTILIZANDO ANÁLISE EXPLORATÓRIA E TÉCNICAS DE AGRUPAMENTO

Helena Ferreira Paraiso Faillace
UERJ - Universidade do Estado do Rio de Janeiro
helenapfaillace@gmail.com

Isadora Lopes Maldonado Britto
UERJ - Universidade do Estado do Rio de Janeiro
isadora.lopes23@gmail.com

Fernanda da Serra Costa
UERJ - Universidade do Estado do Rio de Janeiro
fcosta@ime.uerj.com

Resumo

Este trabalho objetivou analisar o desempenho dos candidatos que realizaram o ENEM em 2019, de forma a identificar possíveis relações entre tais desempenhos e um conjunto de variáveis socioeconômicas, assim como, expor como as técnicas estatísticas são ferramentas valiosas para extração de informações de grandes bases de dados. Os resultados mostraram como a análise exploratória de dados e as técnicas de agrupamento permitiram identificar seis grupos de candidatos com desempenho e características socioeconômicas similares, indicando a influência socioeconômica na disputa por uma vaga em um Ensino Superior de qualidade e, consequentemente, no mercado de trabalho, indicando a importância das políticas públicas que visem deixar, ao menos um pouco, essa competição mais justa.

Palavras-chave: *Desempenho no ENEM; Análise Exploratória de Dados; Análise de Agrupamentos; Kmeans; Desigualdade social.*

1. Introdução

No contexto educacional brasileiro, o Exame Nacional do Ensino Médio (ENEM) tem fundamental papel quando se trata do ingresso no ensino superior. Desde 1998, o ENEM existe como instrumento de avaliação do desempenho dos estudantes no ensino médio (Ministério de Educação e Cultura, 2022) e hoje, tornou-se a maior porta de entrada para o ensino superior no país. Apesar de algumas instituições, como a Universidade do Estado do Rio de Janeiro (UERJ), não o utilizar para a seleção de seus estudantes.

O Ministério da Educação disponibiliza de maneira aberta, microdados de todas as provas do ENEM desde 1998. Os microdados são o menor nível de desagregação de dados recolhidos por meio do exame. Eles atendem à demanda por informações específicas ao disponibilizar as provas, os gabaritos, as informações sobre os itens, as notas e o questionário respondido pelos inscritos no ENEM contendo informações socioeconômicas dos mesmos (INEP, 2019).

Através da análise dessa base de dados é possível investigar a influência das condições socioeconômicas no desempenho dos candidatos. O resultado deste tipo de investigação pode, por exemplo, ajudar na identificação de pontos de atuação para melhoria no ensino médio brasileiro.

Presente em todas as áreas do conhecimento humano, a estatística é um campo da matemática responsável por desenvolver diferentes métodos para análise e interpretação de dados. Através dela, é possível obter informações de conjuntos de dados e observar padrões, até a tomada de decisão através da inferência aplicada sob os mesmos.

Assim, neste trabalho, foi realizada uma análise estatística do desempenho dos candidatos da região Sudeste no ENEM 2019, última edição do ENEM disponível na época do desenvolvimento deste trabalho, de forma a mostrar como as técnicas estatísticas são ferramentas valiosas para extração de informações de grandes bases de dados e verificar a existência de influência da situação socioeconômica no desempenho dos candidatos do ENEM 2019.

2. O ENEM

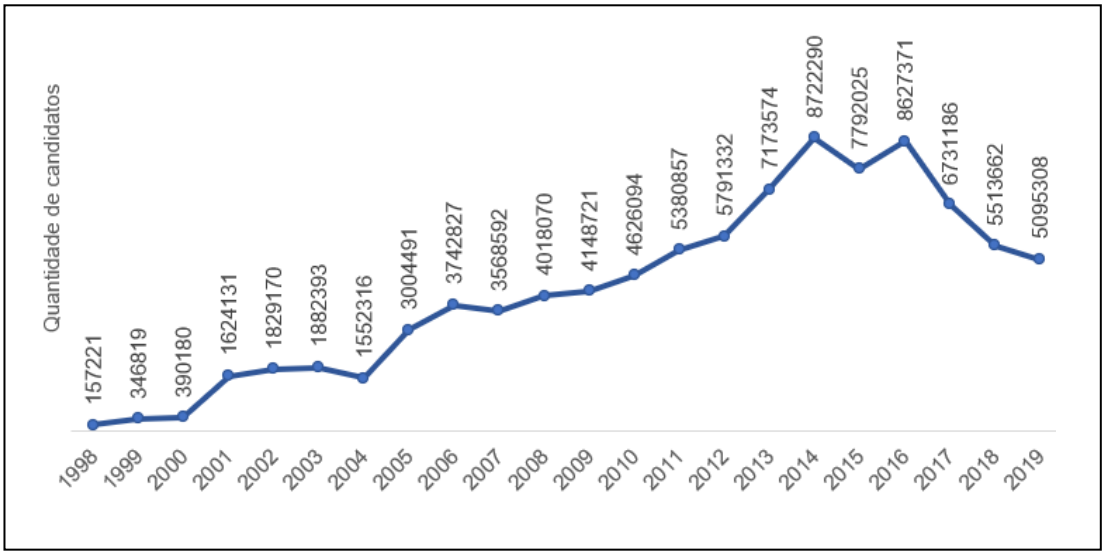
O ENEM foi instituído em 1998, com o objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica e, hoje em dia, tornou-se a maior porta de entrada do ensino superior do Brasil. As notas do ENEM podem ser usadas no Sistema de Seleção Unificada (SISU) para acesso a universidades que aderem ao ENEM e ao

Programa Universidade para Todos (ProUni) para concessão de bolsas de estudos integrais e parciais aos estudantes. Além disso, os participantes do ENEM podem pleitear financiamento estudantil em programas do governo, como o Fundo de Financiamento Estudantil (Fies). A Figura 1 mostra a linha do tempo dos principais acontecimentos ao longo da história do ENEM. Na Figura 2, pode-se observar a evolução do número de inscritos no ENEM de 1998-2019, onde se observa o número crescente de inscrições até 2014.

Figura 1 – Linha do tempo dos principais acontecimentos ENEM

Fonte: autoria própria

Figura 2 – Histórico de inscritos no ENEM



Fonte: autoria própria

O Exame é dividido em 5 provas: Linguagens, Códigos e suas Tecnologias (LC), Matemática e suas Tecnologias (MT), Ciências da Natureza e suas Tecnologias (CN), Ciências Humanas e suas Tecnologias (CH) e Redação. Cada prova possui suas competências a serem avaliadas de acordo com o conteúdo abordado no ensino médio (EM).

A nota das provas objetivas do ENEM, desde 2009, se baseia na Teoria de Resposta ao Item (TRI), utilizada em diversos países, como por exemplo nos Estados Unidos para a aplicação do SAT (*Scholastic Aptitude Test*) (BATISTA, 2018), onde a quantidade de acertos não define a nota do estudante, mas sim a consistência nos níveis de questões acertadas.

Pela TRI, se sai melhor o candidato que acertar mais questões fáceis e médias do que difíceis, pois o sistema entende que o participante apresentou um comportamento coerente, ou seja, não houve uma tentativa de acerto por sorte.

Diferentemente das outras provas, a redação possui um sistema de pontuação diferenciado onde cada uma das 5 competências de sua estrutura vale de 0-200 pontos, totalizando uma nota máxima de 1000 pontos. Por conta desse método de avaliação, um candidato pode obter uma nota zero, o que não ocorre nas outras provas – tema este que será discorrido mais adiante (INEP,2020).

3. Métodos

Para a análise estatística do desempenho dos candidatos, da região Sudeste do Brasil no ENEM 2019, e identificação de possíveis relações com as condições socioeconômicas, adotou-se as técnicas estatísticas de Análise Exploratória de Dados (AED). Segundo John W. Tukey, em *The Future of Data Analysis* em 1962, a AED define-se como:

> *"procedimentos para analisar dados, técnicas para interpretar os resultados de tais procedimentos, formas de planejar a reunião dos dados para tornar sua análise mais fácil, mais precisa ou mais exata e toda a maquinaria e os resultados da estatística (matemática) que se aplicam a análise de dados.".*

A AED emprega uma grande variedade de técnicas gráficas e quantitativas, visando maximizar a obtenção de informações ocultas na sua estrutura, descobrir variáveis importantes e suas tendências, detectar comportamentos anômalos, testar se são válidas as hipóteses assumidas, escolher modelos e determinar o número ótimo de variáveis, sendo, portanto, uma ferramenta valiosa para inferência estatística. Neste trabalho utilizou-se, em especial, gráficos de barras, gráficos boxplots, gráficos de dispersão, histogramas, análise de correlação e regressões lineares.

Na busca por identificar padrões de comportamento, utilizou-se também técnicas de mineração de dados (*Data Mining*), tais como árvores de decisão e análise de agrupamentos, técnica que permite agrupar dados semelhantes entre si.

Ao longo da pesquisa fez-se necessário o uso de ferramentas de apoio estatístico e gráfico. Para a parte estatística, utilizou-se a linguagem de programação *Python* e o *Software Rstudio* versão 1.4.1717 - Programa R versão 4.1.0 (2021-05-18). Foram utilizados os pacotes do R, *tidyverse, ggplot2, psych, corrplot, rpart, rpart.plot, caret, partykit e dplyr*. No *Python*, foram utilizadas as bibliotecas *pandas, numpy, matplotlib.plot, sklearn.decomposition (PCA), seaborn, sklearn.cluster (KMeans), sklearn.preprocessing (StandardScaler), sklearn.metrics (silhouette_samples e silhouette_score)*. Para apoio visual utilizou-se, além das linguagens de programação citadas anteriormente, o *Software Microsoft Power BI* que é um serviço de análise de dados com objetivo de fornecer visualizações interativas de forma simples e prática.

A seguir são descritas sucintamente algumas das técnicas utilizadas.

3.1 Correlação de Pearson

O coeficiente de correlação de Pearson mede o grau de ajustamento dos valores em torno de uma reta e é determinado por:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum(x_i - \bar{x})^2] * [\sum(y_i - \bar{y})^2]}} \quad (1)$$

Onde:

x_i,y_i representam respectivamente os valores das variáveis X e Y

$\bar{x}$,$\bar{y}$ representam respectivamente as médias dos valores x_i e y_i

Na Tabela 1 encontram-se os graus de correlação linear entre duas variáveis.

Tabela 1 – Grau de relação linear entre X e Y

Valor de R	Relação Linear
0	Nula
0 < R ≤ 0,30	Fraca
0,30 < R ≤ 0,60	Média
0,60 < R ≤ 0,90	Forte
0,90 < R ≤ 0,99	Fortíssima
1	Perfeita

Fonte: Hochhein, 2011 apud Fonseca, 1995

3.2 Regressão Linear Simples e Gráficos de Dispersão

Ao invés de explicar a força de correlação entre as variáveis, a regressão linear descreve a relação entre as variáveis analisadas através de uma equação. A regressão linear simples consiste em obtermos uma equação linear que melhor traduza a relação entre a variável dependente e a independente. Dada por:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad (2)$$

Onde:

Y_i é o valor observado para a variável Y no i-ésimo nível da variável X

β_0 é a constante de regressão, que representa o intercepto da reta com o eixo dos Y

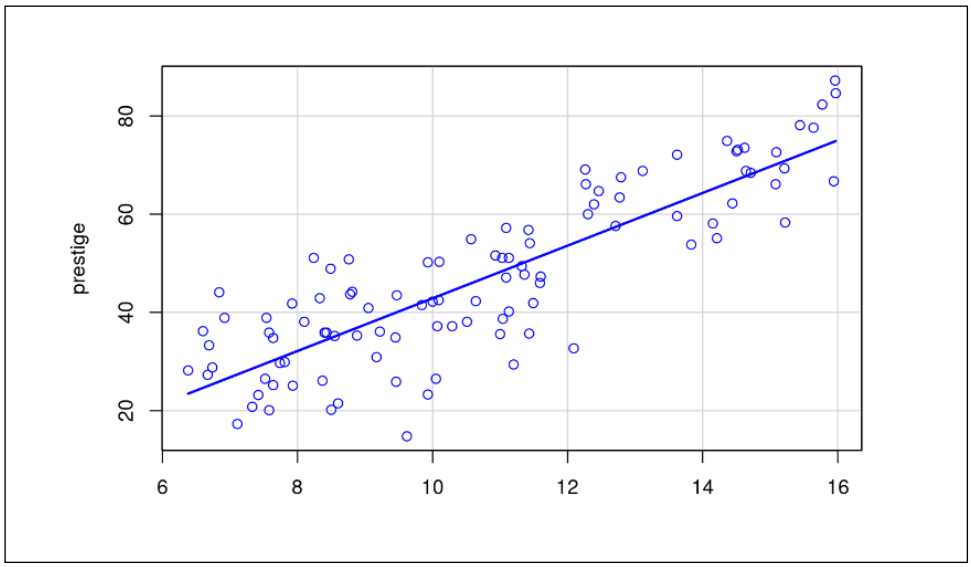
β_1 é o coeficiente de regressão, que representa a variação de Y em função da variação de uma unidade da variável X

X_i é o i-ésimo nível da variável independente X (i = 1,2…, n)

e_i é o erro que está associado à distância entre o valor observado Y_i e o correspondente ponto na curva para o mesmo nível i de X

A Figura 3 apresenta um exemplo de diagrama de dispersão dos valores da variável dependente (Y) em função da variação da variável independente (X), e a reta de regressão.

Figura 3 – Exemplo de gráfico de dispersão e reta de regressão entre as variáveis X e Y



Fonte: Miranda Freire (2021)

3.3 Árvore de Regressão

A árvore de decisão é uma ferramenta de suporte à tomada de decisão que usa um gráfico em formato de árvore e demonstra de forma visual as condições e probabilidades para se chegar aos resultados desejados. As árvores de decisão são modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de dados, i.e., são construídas utilizando um conjunto de treino formado por entradas e saídas, onde estas últimas, são as classes.

De forma geral, estes modelos buscam decompor problemas complexos em subproblemas mais simples e assim, recursivamente, esta técnica é aplicada a cada subproblema encontrado. Uma árvore de decisão geralmente começa com um nó único dividindo-se em n possíveis resultados. Cada um destes leva a nós adicionais, que se ramificam em outras possibilidades e assim, cria-se o formato de uma árvore (LAURETTO, 2010).

3.4 Análise de Agrupamento

Análise de agrupamentos, ou *clustering*, é o grupo de técnicas cujo objetivo consiste em separar objetos em grupos, com base nas características que estes possuem. A ideia básica visa colocar, em um mesmo grupo (cluster), objetos que sejam similares de acordo com algum critério estabelecido. Este, baseia-se normalmente em uma função de similaridade, que recebe dois objetos e retorna a distância entre eles. Os grupos encontrados devem apresentar alta homogeneidade interna e alta heterogeneidade externa (WUNSCH, 2008).

A análise de agrupamento é uma técnica de aprendizado não supervisionado, pois tem por objetivo agrupar um conjunto de dados apenas de acordo com suas características, sem nenhum rótulo prévio na base que nos indique qualquer tipo de separação entre eles. Assim, é possível extrair características escondidas dos dados e desenvolver as hipóteses a respeito de sua natureza.

Dentro as técnicas de aprendizado não supervisionado e de *clustering*, foi adotado o algoritmo *KMeans* (K-Médias). O nome do algoritmo deve-se a forma como os k clusters são formados a partir do conjunto de dados em que o centro do cluster é a média aritmética de todos os objetos do cluster. Podendo ser descrito de forma geral pelos seguintes passos:

a) Passo 1 - Definir o número de agrupamentos (clusters)

Neste trabalho foi adotado o método Cotovelo (*Elbow Method*). Sua ideia é executar o *KMeans* para várias quantidades de clusters. Com o aumento da quantidade de clusters testados, as diferenças entre os clusters se tornam menores, e as diferenças das observações intra-clusters aumentam. Busca-se, então, o equilíbrio para que as informações em cada agrupamento sejam o mais homogêneas possíveis e que os agrupamentos formados sejam os mais diferentes uns dos outros. Ou seja, busca-se uma quantidade de clusters em que a soma dos quadrados intra-cluster seja a menor possível, sendo zero o resultado ótimo. Esse cálculo é dado por:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \qquad (3)$$

Onde: x representa cada uma das observações que temos μ_i como média de cada cluster i (ou centro do cluster i, centroide), C_i conjunto de observações do clusters i e k o número de clusters.

b) Passo 2 - Inicialização dos centroides

Os K centroides recebem valores iniciais (de forma aleatória do conjunto de pontos) para que o algoritmo possa começar a trabalhar de forma iterativa até encontrarmos os melhores agrupamentos.

c) Passo 3 - Cálculo de distâncias dos N pontos até os K centroides

Todas as observações são associadas ao seu centroide mais próximo, ou seja, de menor distância, calculada como a distância Euclidiana entre as observações e os centroides. Assim, os centroides 'incorporam' os pontos que estiverem mais próximos deles e são formados os primeiros clusters.

d) Passo 4 - Recalculando os centroides e geração novos clusters

Neste momento, as coordenadas dos centroides são refinadas. Para cada um dos K clusters formados, são calculadas médias utilizando todos os pontos que a eles pertencem, fazendo com que suas coordenadas sejam atualizadas.

e) Passo 5 - Repetir o processo até a convergência

A convergência é atingida quando ao realizada a média dos pontos para recálculo das coordenadas dos centroides, a mesma não apresentar diferença significativa no resultado, fazendo com que os centroides não se movimentem mais.

4. Base de Dados

O banco de dados utilizado foi disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Juntamente com o *dataset*, são oferecidos materiais complementares como o dicionário das variáveis, gabarito/itens das provas e edital do concurso, matriz de referência de cada prova - que compreende os eixos cognitivos, as competências e as habilidades avaliadas em cada área de conhecimento do EM e, o manual de redação - que expõe a metodologia de avaliação da redação bem como o que se espera dos participantes em cada uma de suas competências avaliadas. A estrutura do banco de dados é organizada sendo cada linha representando um participante do concurso e, nas colunas, estão as informações sobre seu desempenho nas 5 provas - LC, CH, MT, CN e Redação, assim como as variáveis que traduzem seu perfil socioeconômico. O banco de dados do ENEM 2019 é composto por 5.095.270 registros e 136 variáveis, como por exemplo idade, sexo, nível de instrução dos pais dos candidatos

e no que eles atuavam no ano em questão e, se possuíam aparelhos eletrônicos ou não em suas residências.

Cabe destacar que este estudo não teve como objetivo inferir sobre a evolução temporal do ENEM, mas sim traçar um perfil socioeconômico com base no desempenho dos participantes em um ano específico, no caso 2019. Este ano foi escolhido por ser o mais recente quando do início do desenvolvimento deste trabalho. Entretanto, os métodos adotados podem ser aplicados aos dados de qualquer edição do ENEM.

4.1 Delimitação da base de dados

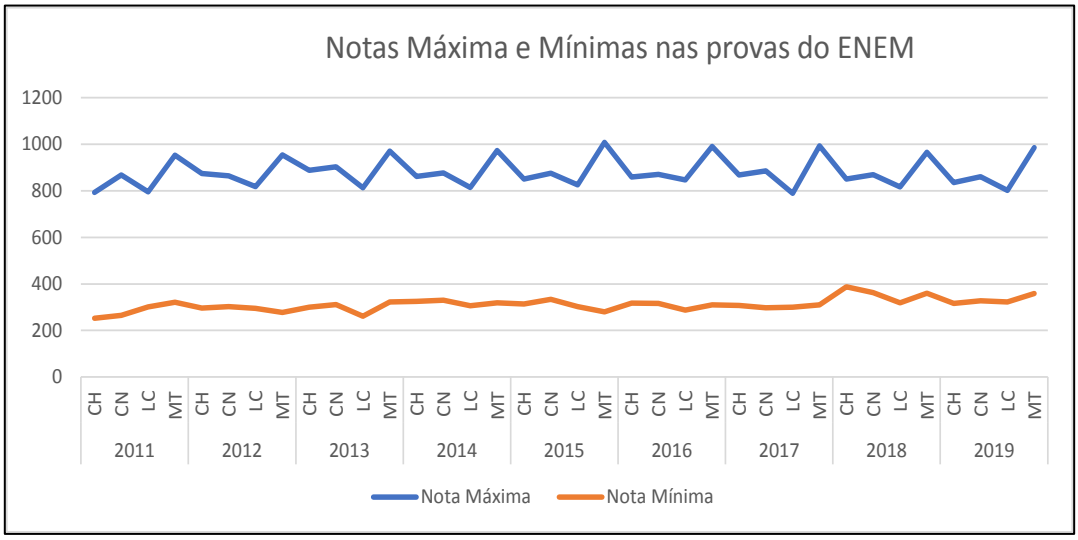
Neste trabalho optou-se por trabalhar com um recorte da base de dados do ENEM 2019 compreendendo a região sudeste do Brasil, uma vez que, segundo dados do IBGE, essa região concentra aproximadamente 42% da população do país e que no ENEM 2019, 35,2% dos inscritos pertenciam a essa região (IBGE,2019).

Na base de dados, é disponibilizada uma variável dicotômica que identifica se o candidato é treineiro ou não. Ser treineiro significa que foi identificado, no instante da inscrição, que o participante não estava cursando ou já teria cursado o ensino médio, portanto, apesar de poder realizar o ENEM não poderia utilizar o resultado para entrar em uma universidade ou se candidatar a algum programa de auxílio governamental, como o ProUni, o SISU e o FIES. Foram identificados 212.723 treineiros, correspondendo a 11.9% da base estudada.

Sendo o objetivo analisar o desempenho dos candidatos aptos a utilizarem o resultado do ENEM para ingressarem no ensino superior, os treineiros foram excluídos da base de dados do estudo. Também só foram considerados os candidatos que compareceram os dois dias do exame. Além disso, os candidatos que tiveram nota zero em alguma prova, também foram excluídos, pois pelo critério adotado na correção, não seria possível tirar zero, a menos que estivesse ausente (SOUZA CRUZ, 2014). A Figura 4 apresenta as notas máximas e mínimas das quatro provas objetivas nos ENEM 2011 a 2019, pode-se observar que as notas mínimas se encontram entorno de 300 pontos. Assim, foram excluídos os candidatos que tiveram alguma nota zero.

Dessa forma, a população estudada é composta por todos os participantes do ENEM do ano de 2019 da região Sudeste, que estiveram presentes em ambos os dias de prova, que não eram treineiros e, que obtiveram uma pontuação diferente de zero em suas provas, totalizando 1.085.096 candidatos.

Figura 4 – Notas máximas e mínimas das quatro provas objetivas nos ENEM 2011 a 2019



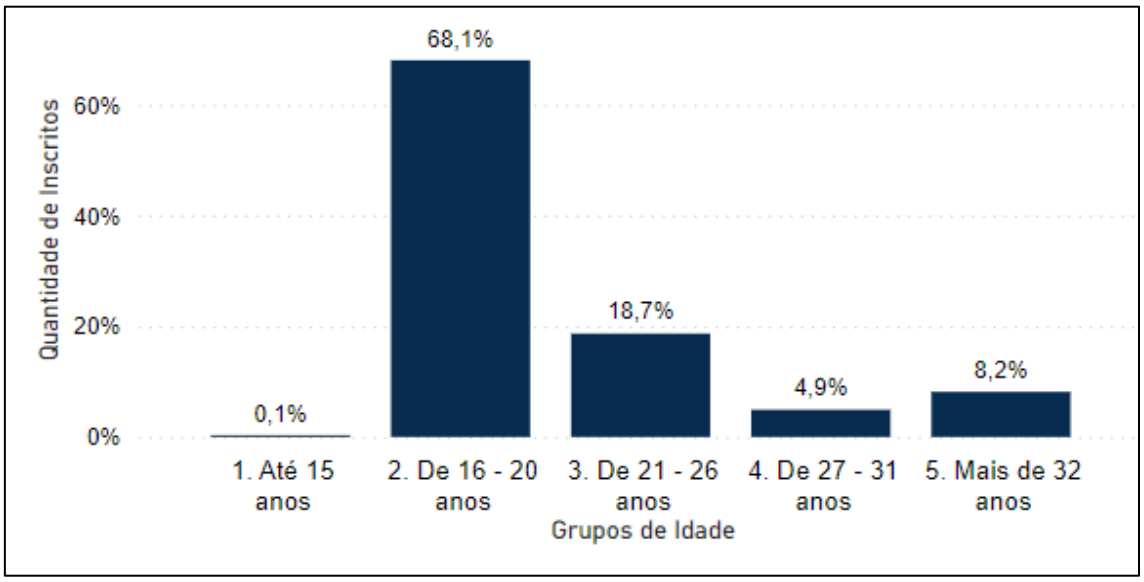
Fonte: autoria própria

5. Extração de Informação através da Análise Estatística

5.1 Características demográficas e socioeconômicas dos participantes do ENEM 2019 da região Sudeste

Identificou-se um equilíbrio entre os gêneros, sendo 59% do gênero feminino. As Figura 5, 6, 7 e 8 apresentam distribuição dos candidatos por faixa etária (5), raça (6), tipo de escola (7) e ano de término do ensino médio (8).

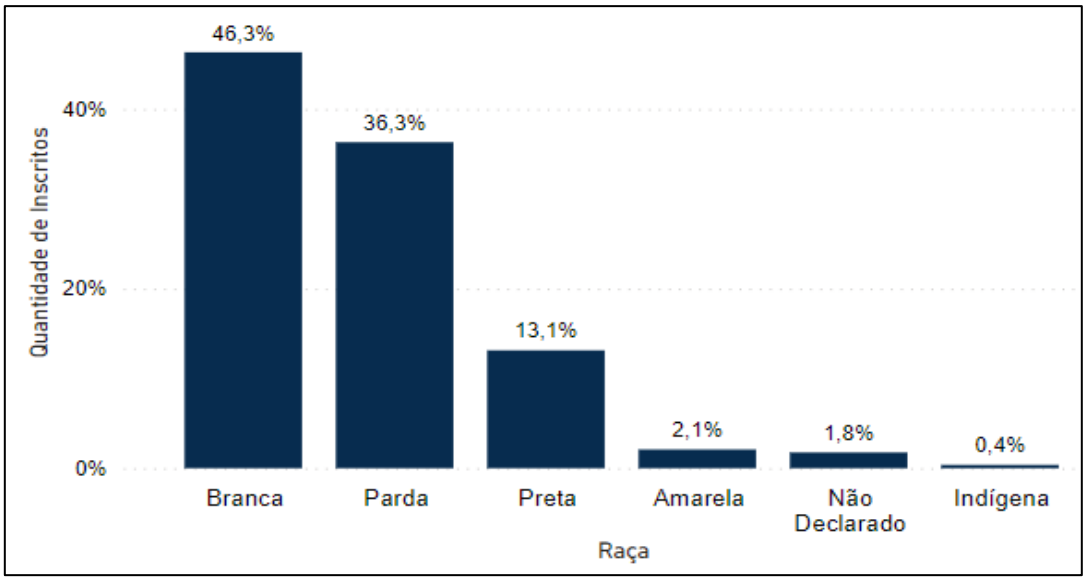
Figura 5 – Perfil demográfico dos participantes do ENEM 2019 da região Sudeste - faixa etaria.



Fonte: autoria própria

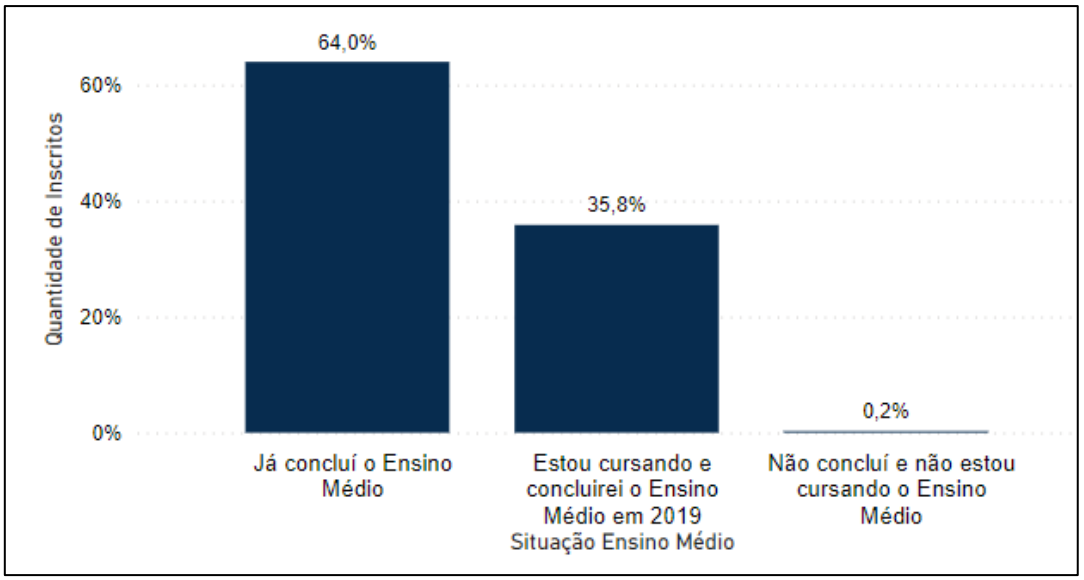
Interessante notar que, a maioria dos candidatos são jovens (faixsa etárias de 16 a 20 e 21 a 26 anos), os pardos e pretos somados ultrapassam em muito pouco os brancos e, ainda chama atenção o percentual de 58% dos candidatos que não responderam o tipo de escola, sendo esta uma informação que seria interessante ser analisada.

Figura 6 – Perfil demográfico dos participantes do ENEM 2019 da região Sudeste – raça.



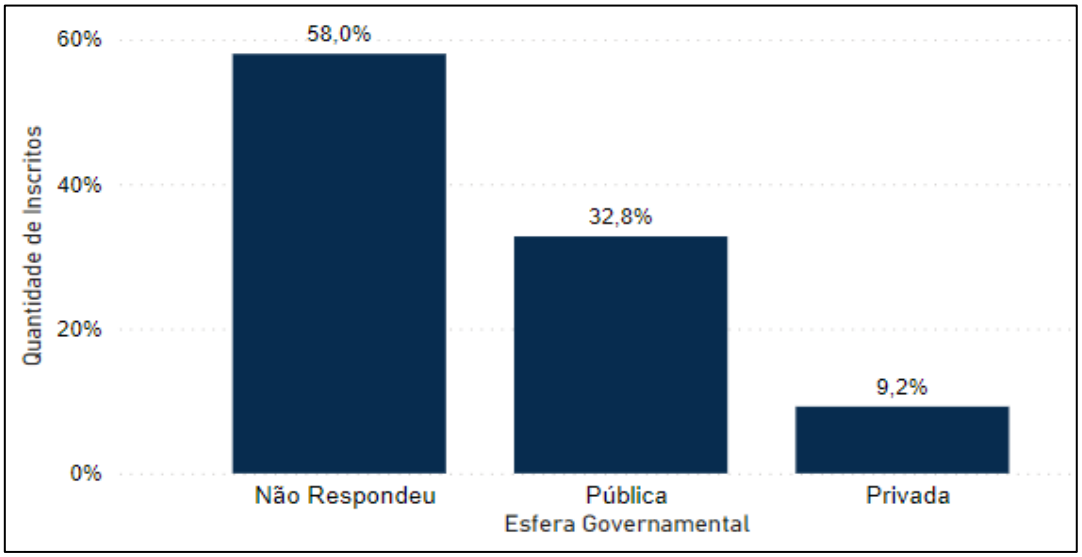
Fonte: autoria própria

Figura 7 – Perfil demográfico dos participantes do ENEM 2019 da região Sudeste - ano de termino do ensino médio.



Fonte: autoria própria

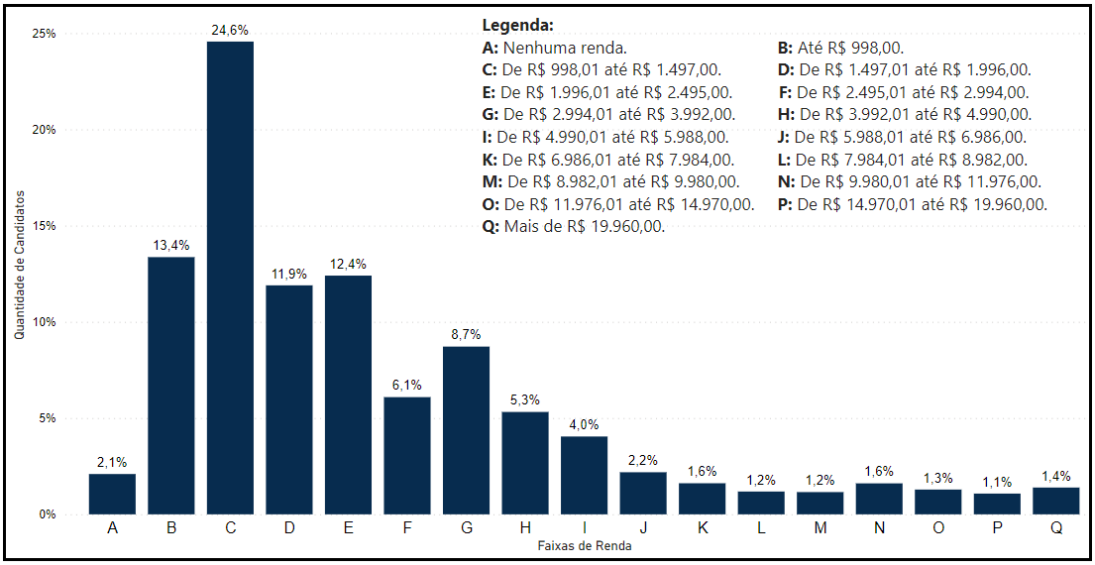
Figura 8 – Perfil demográfico dos participantes do ENEM 2019 da região Sudeste - tipo de escola.



Fonte: autoria própria

As condições socioeconômicas podem ser aferidas por meio de, por exemplo, indicadores de renda, escolaridade e ocupação dos pais, nos permitindo ter uma visão mais ampla acerca das condições de vida dos candidatos. Na Figura 9 pode-se ver a distribuição dos candidatos pela renda familiar que, no questionário do ENEM, estava dividida em 17 classes (A a P). Observa-se que mais de 65% dos participantes encontravam-se concentrados nas faixas A até E, que constituem valores um pouco superiores a dois salários mínimos que, no ano de 2019, correspondia a R$998,00.

Figura 9 – Distribuição do percentual de candidatos por faixa de renda familiar

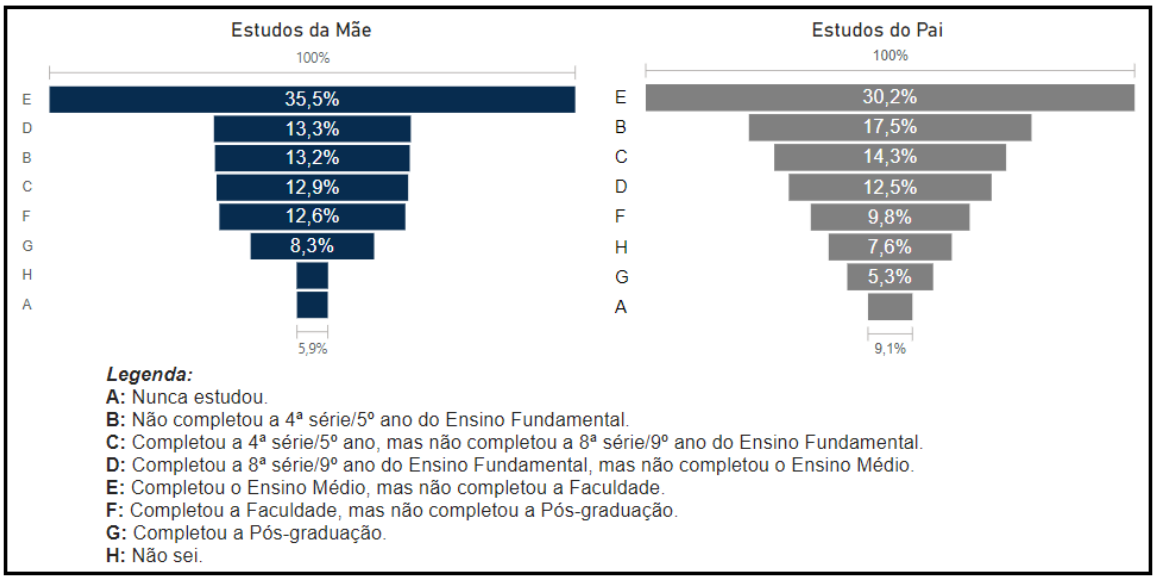


Fonte: autoria própria

Atrelado à renda, temos o nível de estudo e ocupação dos familiares/responsáveis dos participantes cujas categorias estão no questionário do ENEM. Como o nível de estudo e a ocupação são fortemente correlacionados, apresentamos, na Figura 10, apenas a distribuição do percentual de candidatos por faixa de estudo do pai e estudo da mãe respectivamente.

Pode-se observar que o maior percentual se concentra na faixa E que corresponde a ter completado o ensino médio, mas não ter concluído a graduação. Por outro lado, o percentual cujos pais não nunca estudaram é muito baixo, inferior a 3%.

Figura 10 – Distribuição do percentual de candidatos por faixa de estudo do pai e da mãe

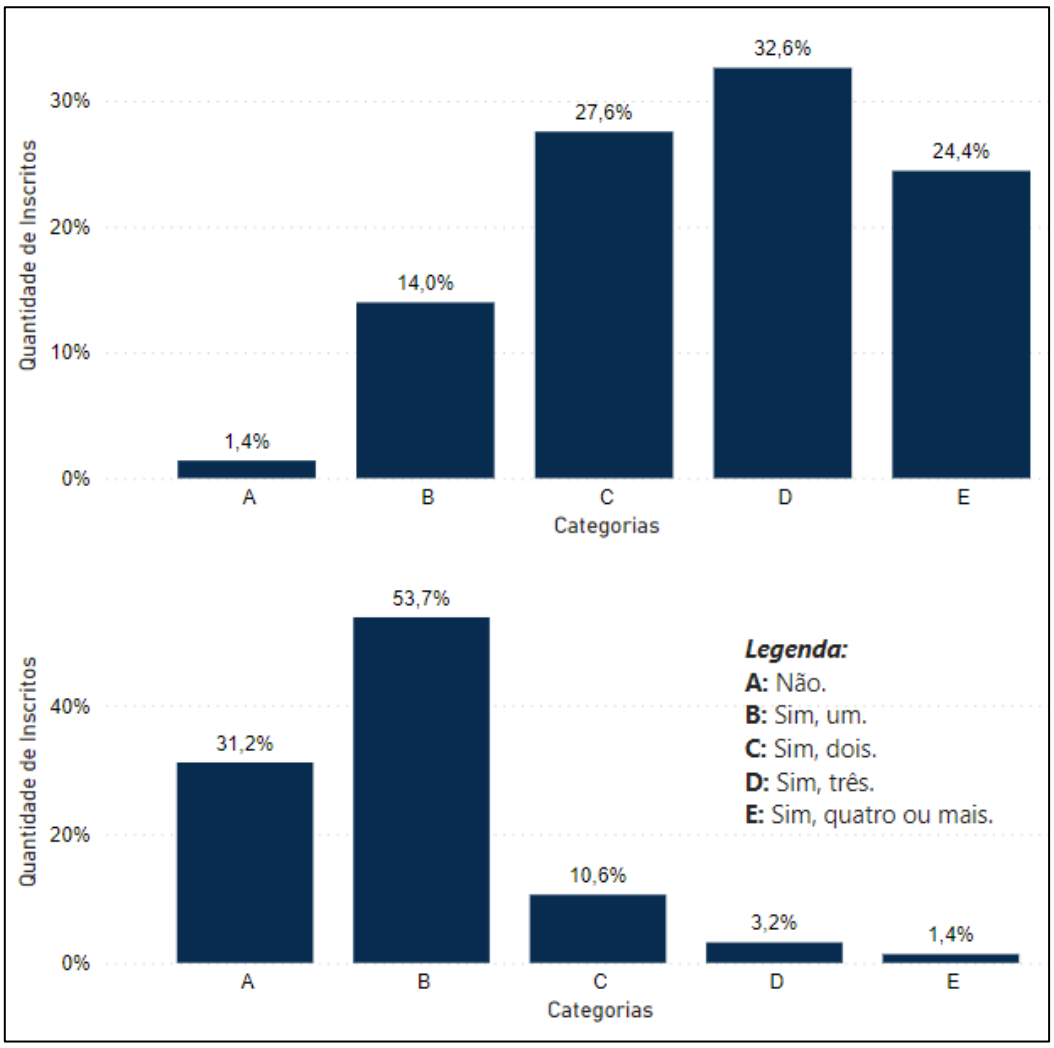


Fonte: autoria própria

Em relação ao acesso a tecnologia, no questionário respondido pelos candidatos existiam questões sobre o número de telefones celulares e computadores. A Figura 11 apresenta os gráficos das distribuições dos percentuais de candidatos por quantidade de computadores e telefones celulares em suas residências.

Pode-se observar o formato inverso das distribuições, enquanto 53,7% possuem apenas um computador na residência e apenas 1,45% possuem quatro ou mais computadores, o percentual de candidatos que possuem quatro ou mais celulares na residência é de 24,4%.

Figura 11 – Distribuição percentual de candidatos por quantidade de, respectivamente, telefones celulares e de computadores em suas residências

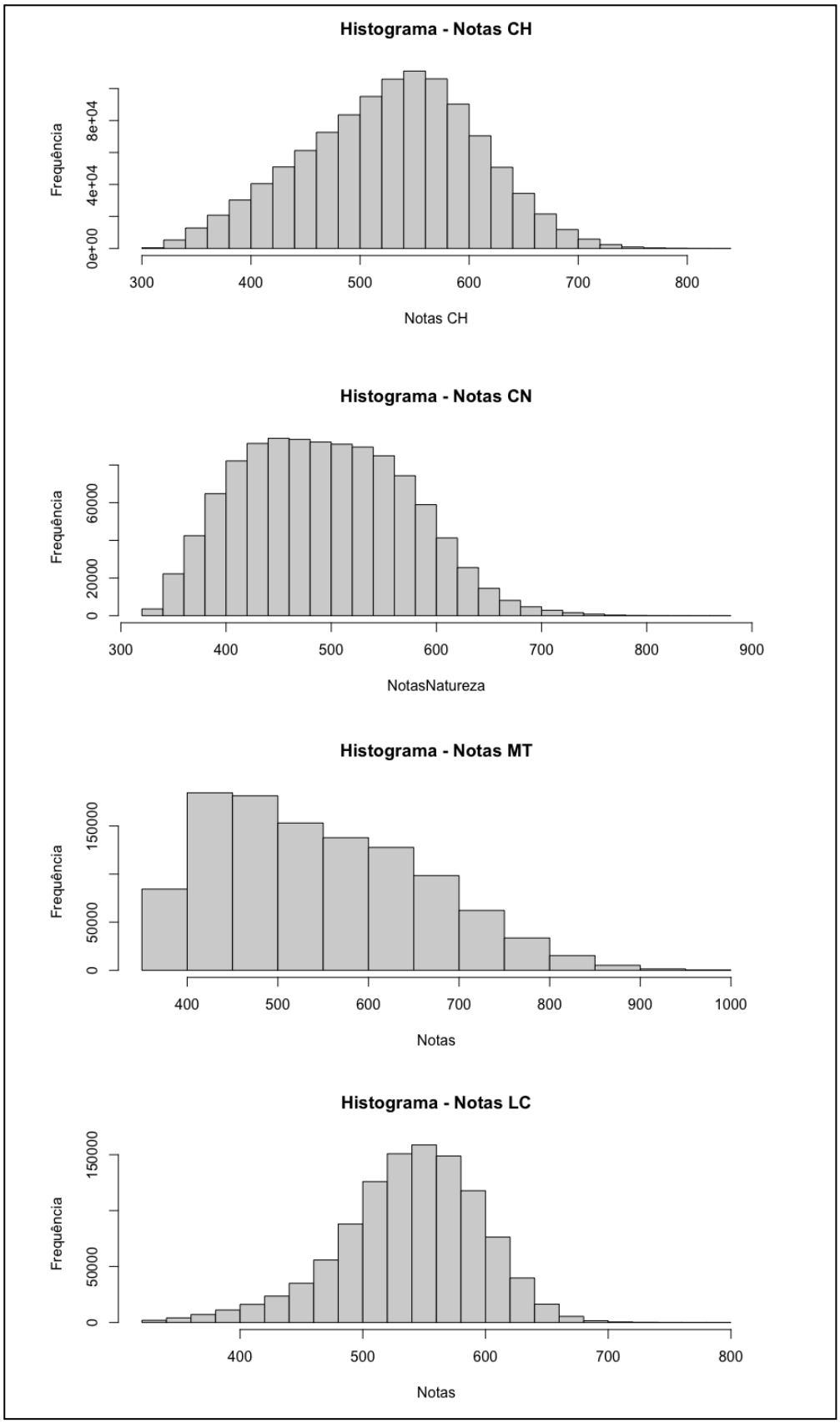


Fonte: autoria própria

5.2 Desempenho nas áreas de conhecimento

Na Figura 12 são apresentadas as distribuições das notas dos candidatos em cada uma das cinco provas através dos respectivos histogramas, nos quais pode-se observar os comportamentos distintos dos candidatos nas cinco provas, enquanto as notas das provas CN e MT apresentam assimetria a direita, as notas de CH e LC apresentam comportamento inverso e menor assimetria. Na Tabela 2 encontram-se as principais estatísticas das notas de cada prova. Entre as provas objetivas a maior variabilidade (desvio padrão) ocorre em MT, onde também se vê o maior valor máximo, apesar da mediana ser inferior à das provas CH e LC.

Figura 12 – Distribuição das notas dos candidatos em cada uma das cinco provas



Fonte: autoria própria

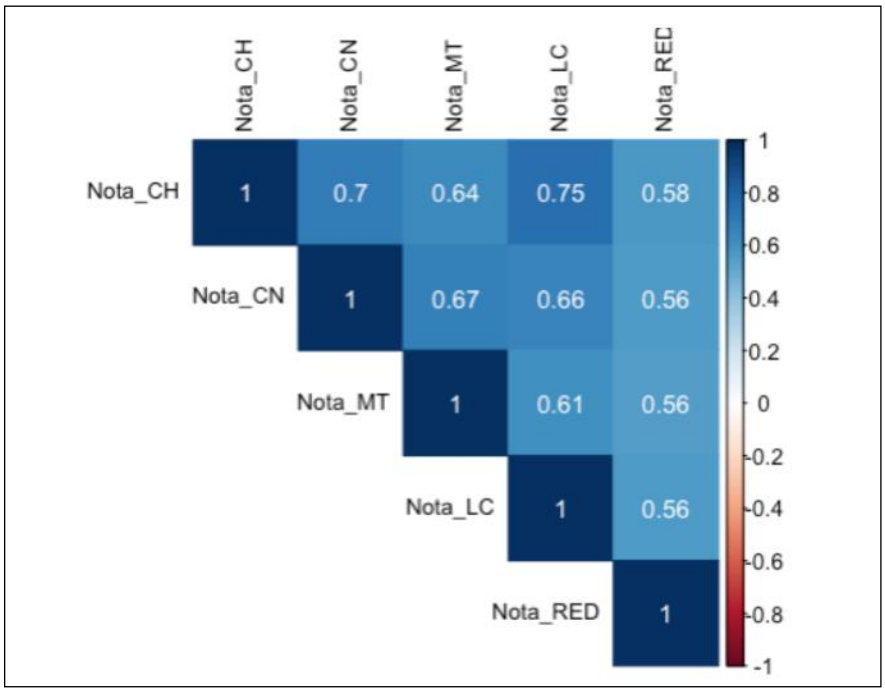
Tabela 2 - Resumo das principais estatísticas de cada prova

Prova	Média	DP	Máx.	Min.	Mediana	25%	75%
CH	528.2	78.7	835.1	315.9	533.3	473.9	583.6
LC	538.4	57.5	787.3	322.0	543.0	505.1	578.1
CN	493.7	77.8	860.9	327.9	490.4	432.4	550.8
MT	545.5	114.5	985.5	359.0	529.3	450.7	627.7
RED	611.7	158.81	1000.0	40.0	600.0	520.0	700.0

Fonte: autoria própria

A Figura 13 contem a matriz de correlação entre as notas das cinco provas. Todas as correlações são positivas, sendo a maior delas entre CH e LC - o que talvez seria esperado visto que são provas que possuem conteúdo mais relacionados e, são realizadas no mesmo dia de concurso.

Figura 13 - Matriz de correlação entre as notas dos candidatos de cada prova

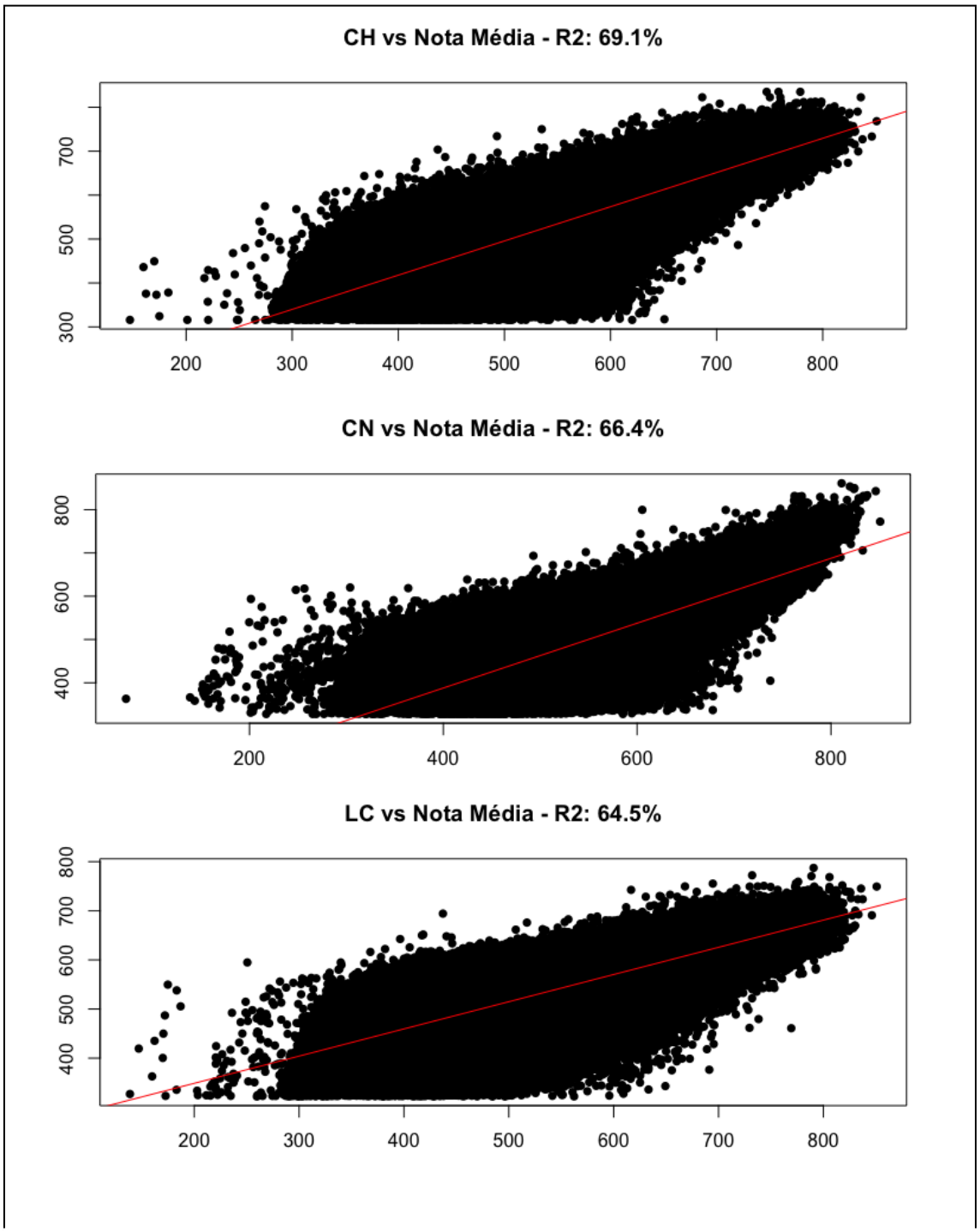


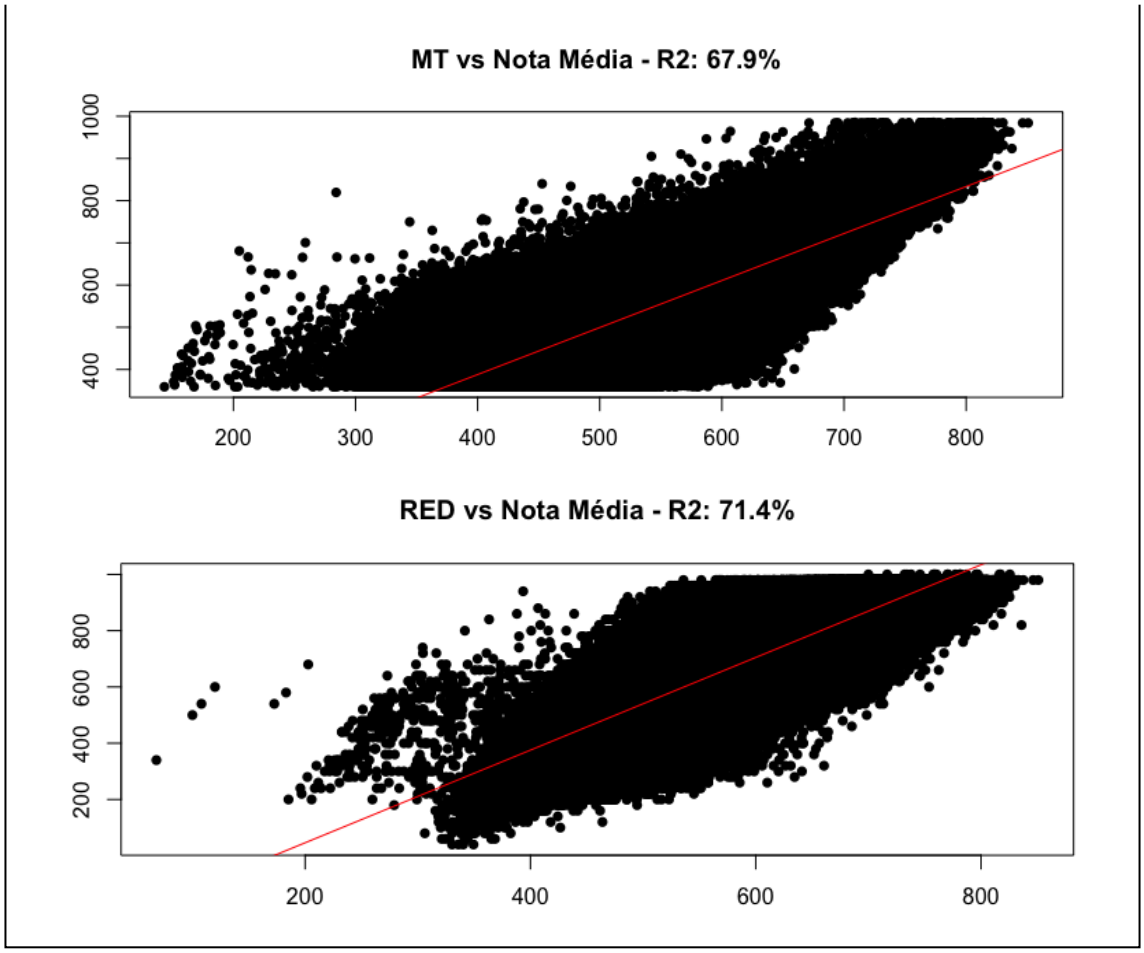
Fonte: autoria própria

No ENEM os candidatos não possuem uma nota final única e sim, notas individuais para cada prova. Uma possível motivação para a não consolidação dos resultados das provas em uma única nota esteja associada a forma de utilização do resultado do ENEM para acesso ao ensino superior. Cada universidade tem liberdade de adotar seu próprio critério, em geral, as notas das provas associadas à área de conhecimento do curso superior escolhido têm maior peso.

O objetivo do trabalho não é investigar o acesso ao ensino superior, mas traçar um perfil socioeconômico do desempenho acadêmico dos candidatos. Assim, decidiu-se trabalhar com a média aritimética simples das cinco notas para representar o desempenho dos candidatos. Os gráficos de dispersão juntamente com a reta de regressão das notas das provas e a média das cinco provas encontram-se na Figura 14.

Figura 14 – Gráficos de dispersão das notas das cinco provas em relação a média





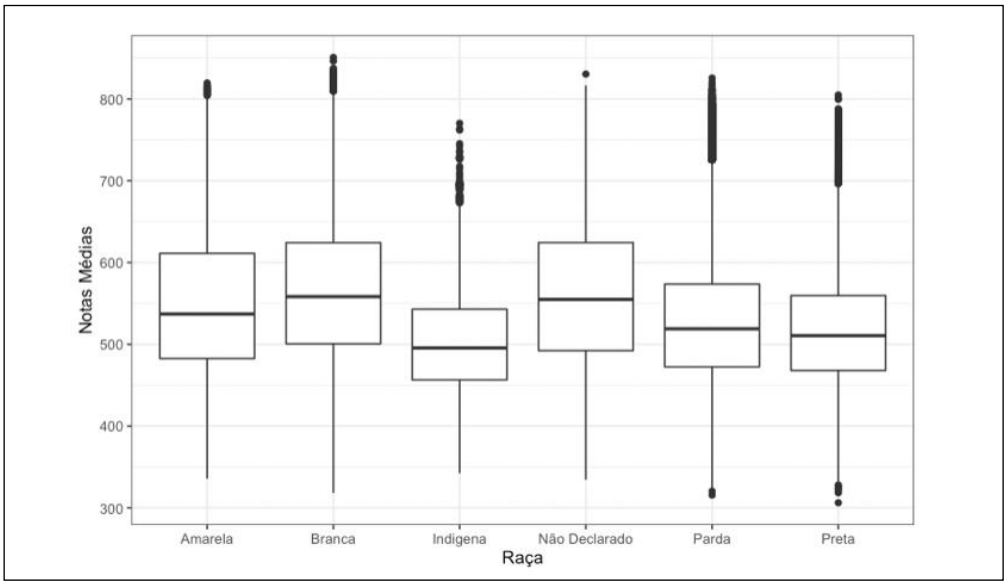
Fonte: autoria própria

Apesar dos coeficientes R2 das regressões estarem entre 65% e 71%, podemos ver que a média explica a nota individual das provas uma vez que seguem a mesma tendência apesar da dispersão.

5.3 Investigação da influência socioeconômica no desempenho no ENEM

Analisando o desempenho dos candidatos em relação a cada uma das variáveis socioeconômicas disponíveis no questionário do ENEM, observa-se na Figura 15 o desempenho dos candidatos por raça. Os gráficos Box-Plot de cada raça mostram que as raças Amarela, Branca e Não Declarado têm maior variabilidade e apresentam valores medianos mais altos quando comparado às outras raças, que por sua vez, apresentam mais *outliers*. As notas dos candidatos que se declararam indígenas apresentaram menor dispersão e menor valor mediano, valor inferior ao primeiro quartil daqueles que se declararam brancos.

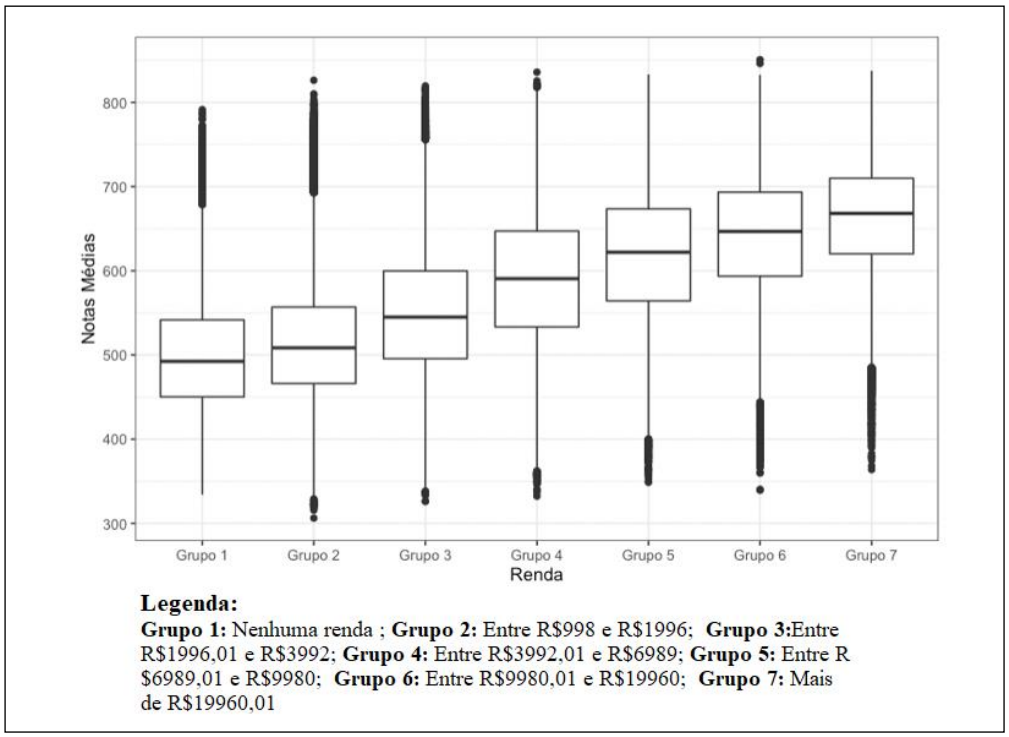
Figura 15 – Desempenho por raça



Fonte: autoria própria

Já o desempenho dos candidatos em relação as faixas de renda familiar é apresentado na Figura 16, onde observa-se a forte relação das duas variáveis, com nítida tendência de aumento do desempenho médio para as faixas de maiores rendas familiares e a inversão dos *outliers*.

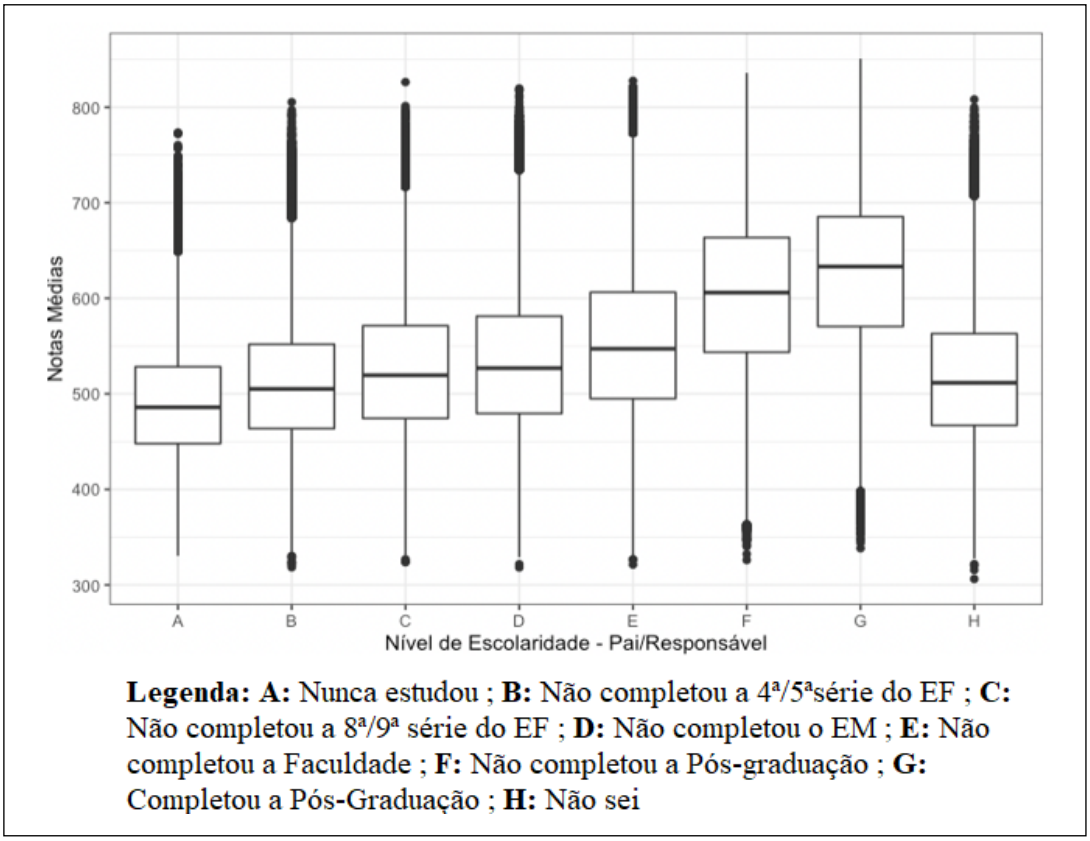
Figura 16 – Desempenho por faixa de renda familiar



Fonte: autoria própria

Analogamente ao comportamento encontrado entre as notas médias dos participantes e suas respectivas faixas de renda familiares, a análise em relação do nível de estudo dos pais/responsáveis apresentada na Figura 17 mostra que as maiores notas correspondem aos candidatos cujos responsáveis possuíam maior nível de instrução, como podemos ver se compararmos as classes A e G, onde a A representa que eles nunca estudaram e G, que possuem pós-graduação completa. O mesmo comportamento dos *outliers* é observado na análise por faixa de renda. Como o comportamento considerando o nível de escolaridade dos pais e mães são similares, na Figura 17 é apresentado apenas o desempenho em relação ao nível de escolaridade do pai/reponsável.

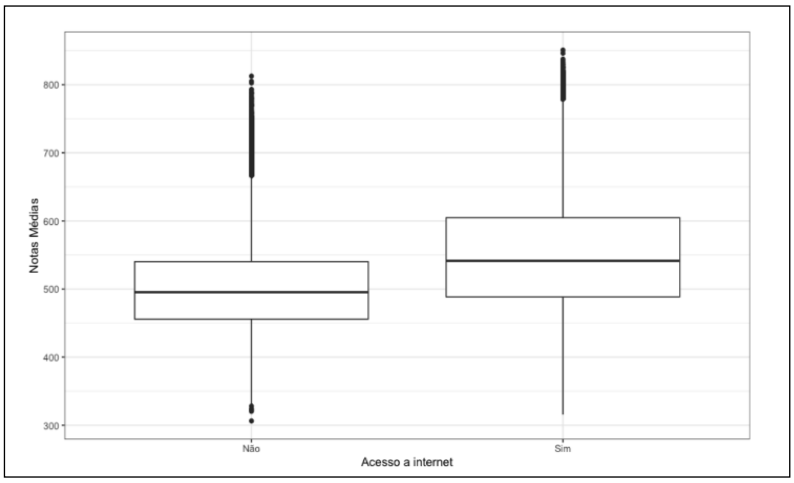
Figura 17– Desempenho por nível de estudo do pai



Fonte: autoria própria

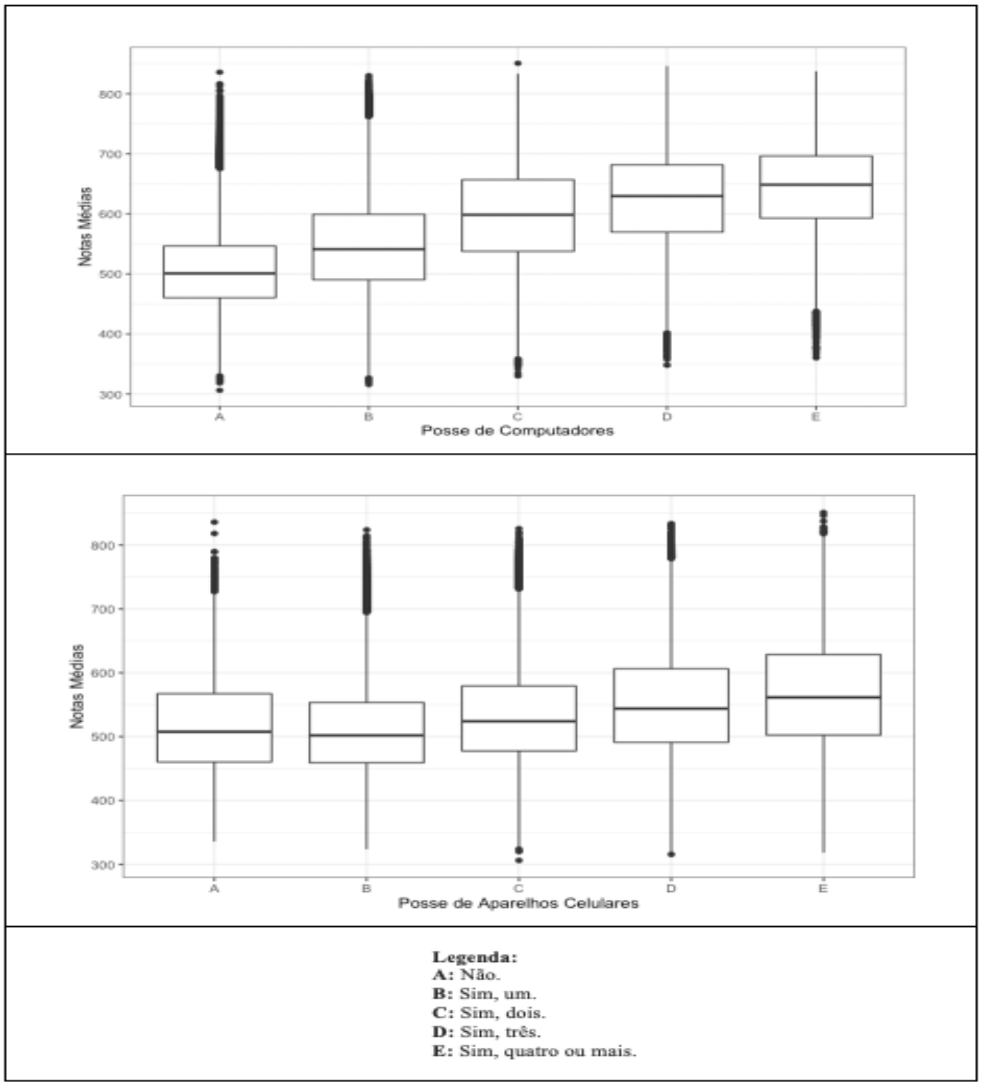
Outro ponto que pode influenciar o desempenho dos candidatos é o acesso a internet, celulares e computadores. Na Figura 18, observa-se que os candidatos que possuíam acesso à internet obtiveram um melhor desempenho na prova. Assim como quanto maior o número de computadores e celulares melhor é o desempenho (Figura 19). Esta diferença é significativamente maior em relação à posse de computadores.

Figura 18 – Média de nota por acesso à internet



Fonte: autoria própria

Figura 19 – Média de nota por número de computadores e aparelhos celulares na residência



Fonte: autoria própria

Para investigar a influencia conjunta de características socioeconômicas no desempenho dos candidatos no ENEM, foi utilizada a árvore de decisão que nos permite realizar uma análise multivariada (HAIR, 1998). Considerando todas as variáveis analisadas individualmente anteriormente, a árvore resultante é apresentada na Figura 20, onde vemos que a primeira variável com menor entropia, i.e., a que gera a primeira divisão nos dados, é a variável renda que inicialmente, representando 100% dos candidatos analisados, temos uma média geral de 543 pontos para estes. Assim a renda define a primeira divisão dos candidatos. À esquerda da árvore, temos os candidatos que possuíam renda entre R$0,00 e R$2.994,00 (Q006 = A – F) – representando 70% dos candidatos com 522 pontos de nota média, enquanto à direita, os 30% restantes com renda acima de R$2.994,01 (Q006 = G – Q), com média superior (595 pontos).

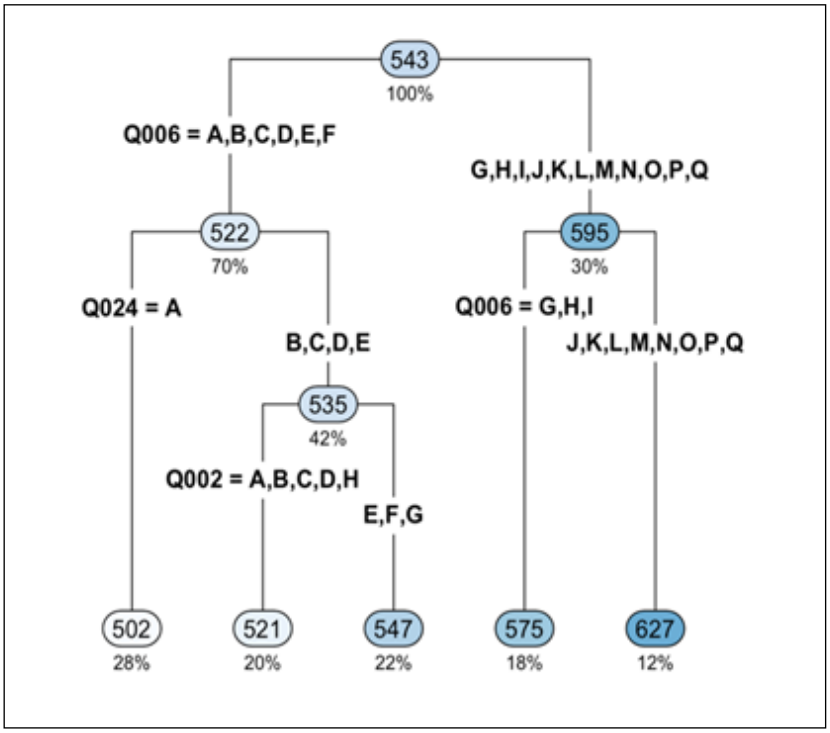
Analisando o lado esquerdo da árvore, temos que a próxima divisão foi quanto à posse de computadores e, a partir desta, já observamos a primeira regra explicitada pela árvore: se o candidato possui uma renda de até R$2.994,00 (Q006 = A – F) e não possui computador (Q024 = A), sua nota média está em 502 pontos (28% dos candidatos), evidenciando um pior desempenho. Mas, caso o candidato tivesse pelo menos 1 computador (Q024 = B – E), sua média está em 535 pontos – superior à média dos candidatos que não possuem computador – representando 42% dos candidatos. Este resultado indica a influência da disponibilidade de computador para o desempenho dos candidatos. A próxima divisão foi baseada no estudo da mãe/responsável dos candidatos. Caso a responsável possuísse até o ensino fundamental completo ou não soubesse responder a essa questão (Q002 = A, B, C, D, H), encontramos dos 42% que tinham pelo menos 1 computador em casa (Q024 = B – E), 20% com uma média de 521 pontos. Caso contrário (EM completo até pós-graduação, Q002 = E – G), temos 22% dos candidatos com uma média de 547 pontos. Sendo assim, as últimas duas regras encontradas do lado esquerdo da árvore mostram que candidatos cujas responsáveis possuíam maior grau de escolaridade, apresentaram maiores médias e vice-versa.

Analisando o lado direito da árvore – candidatos com renda a partir de R$2.994,01 (Q006 = G – Q), representando 30% dos candidatos e, com uma média de 595 pontos – encontramos novamente a renda fazendo uma divisão de dados e nos evidenciando as últimas 2 regras da árvore: caso os candidatos possuam uma renda entre R$2.994,01 e R$5.988,00 (Q006 = G – I), encontramos uma nota média de 575 pontos (18% dos

candidatos) e, caso possuam as maiores categorias de renda – acima de R$5.988,01 (Q006 = J – Q)- encontramos uma média mais alta, de 627 pontos (12% dos candidatos).

Cabe destacar que, no caso renda de até R$2.994,00 (renda mais baixa) a existência de computador e a escolaridade do responsável (mãe) são fatores que influenciam no desempenho dos candidatos. Já para a parcela com renda superior a R$2.994,00, a renda foi o fator dominante no desempenho dos mesmos.

Figura 20 - Árvore de decisão



Fonte: autoria própria

De forma a verificar como conjuntos de características socioeconômicas podem influenciar o desempenho dos candidatos no ENEM, foi utilizada a clusterização através de um algoritmo de *Machine Learning* não-supervisionado, o *KMeans*, para identificar grupos com desempenhos similares e, a partir dos agrupamentos definidos, verificar o conjunto de características socioeconômicas de cada agrupamento para formar os perfis socioeconômicos relacionados ao desempenho dos candidatos. O procedimento de clusterização foi aplicado à média dos candidatos.

A aplicação do Método Cotovelo resultou em 6 clusters. Como a variável adotada para clusterização foi a média dos candidatos, as demais foram analisadas para perfilar os clusters buscando traçar um perfil socioeconômico associados aos desempenhos. Na

Tabela 3 encontram-se características dos clusters (tamanho, média e desvio padrão das médias dos candidatos. Observa-se boa distribuição dos candidatos nos clusters, com maior concentração nos clusters intermediários. As maiores variabilidades estão nos clusters extremos. Os clusters foram numerados de 0 a 5, sendo a numeração crescente com o desempenho médio.

Tabela 3- Distribuição dos clusters com seus respectivos desempenhos médios e suas variabilidades

Cluster	Tamanho do Cluster	Desempenho Médio	Variabilidade
0	12.5%	424 pontos	21.3 pontos
1	22.7%	479 pontos	15 pontos
2	24%	529 pontos	14.6 pontos
3	19.7%	581 pontos	15.9 pontos
4	13.8%	639 pontos	18.3 pontos
5	7.2%	709 pontos	27 pontos

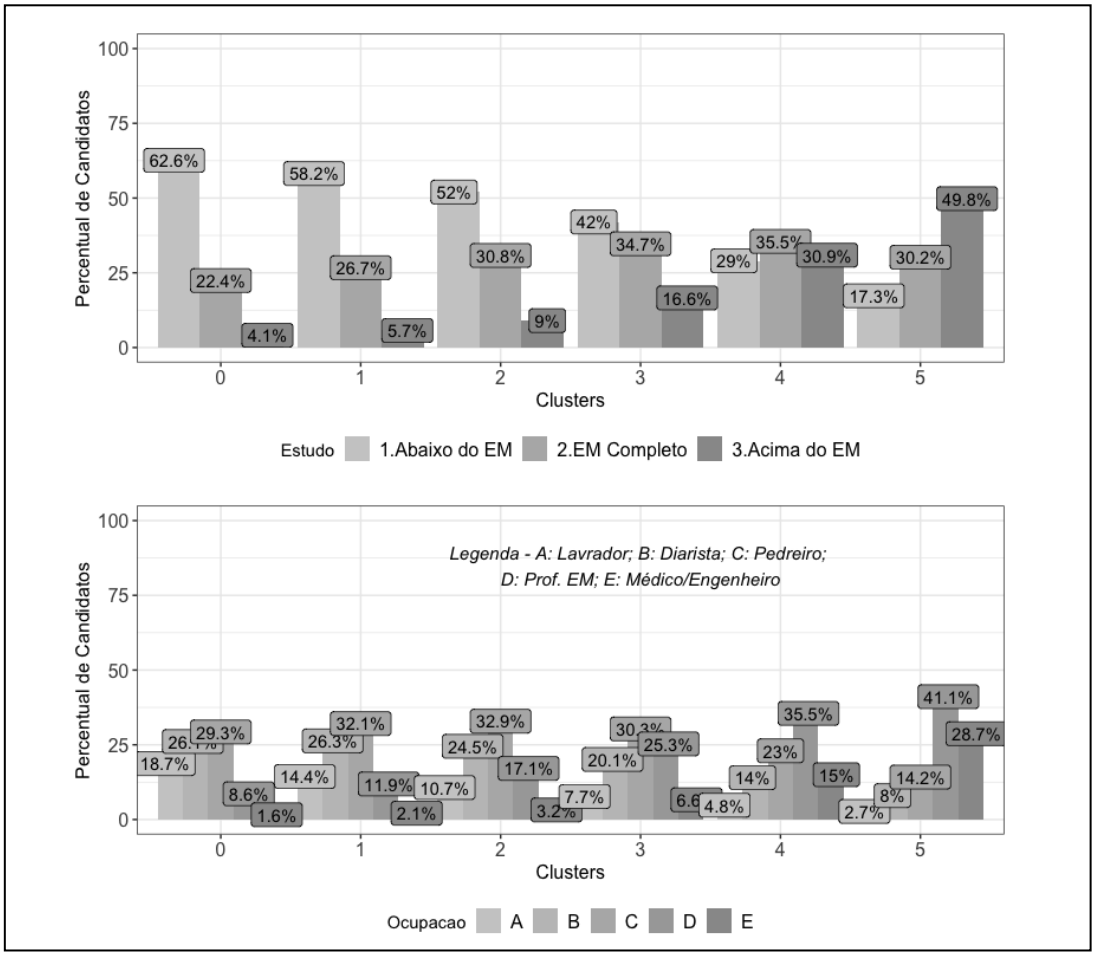
Fonte: autoria própria

As Figuras 21, 22 e 23 apresentam o comportamento das variáveis escolaridade e ocupação dos responsáveis (representadas pela escolaridade e ocupação do pai/responsável, já que o mesmo comportamento foi observado em relação à mãe/responsável), renda familiar, disponibilidade de internet, computador e celular nos seis clusters.

Na Figura 21 observa-se uma inversão do histograma da escolaridade do responsável, enquanto 62% tem escolaridade abaixo do EM no cluster 0, cujo desempenho dos candidatos é o mais baixo, no cluster 5 a escolaridade do responsável acima do EM apresenta o maior percentual (49,8%). Conforme esperado, um comportamento similar ocorre com a ocupação do responsável, já que está diretamente relacionada com a sua escolaridade.

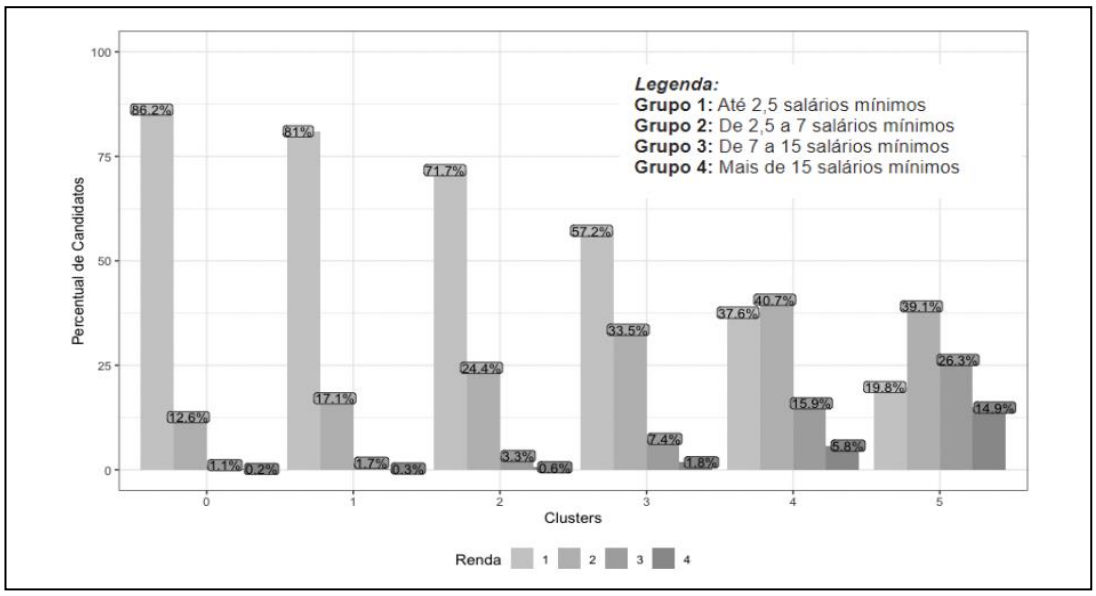
Na Figura 22, para os clusters iniciais - de menor desempenho - observa-se uma maior concentração de candidatos na primeira faixa de renda, isto é, com uma renda familiar de até 2,5 salários mínimos da época. Contrariamente a esse comportamento, os últimos clusters - de melhor desempenho - apresentam candidatos, além de mais dispersos entre as categorias, mais concentrados nas faixas mais altas e menos concentrados nas faixas mais baixas.

Figura 21 – Perfil dos cluster em relação à escolaridade e ocupação do responsável



Fonte: autoria própria

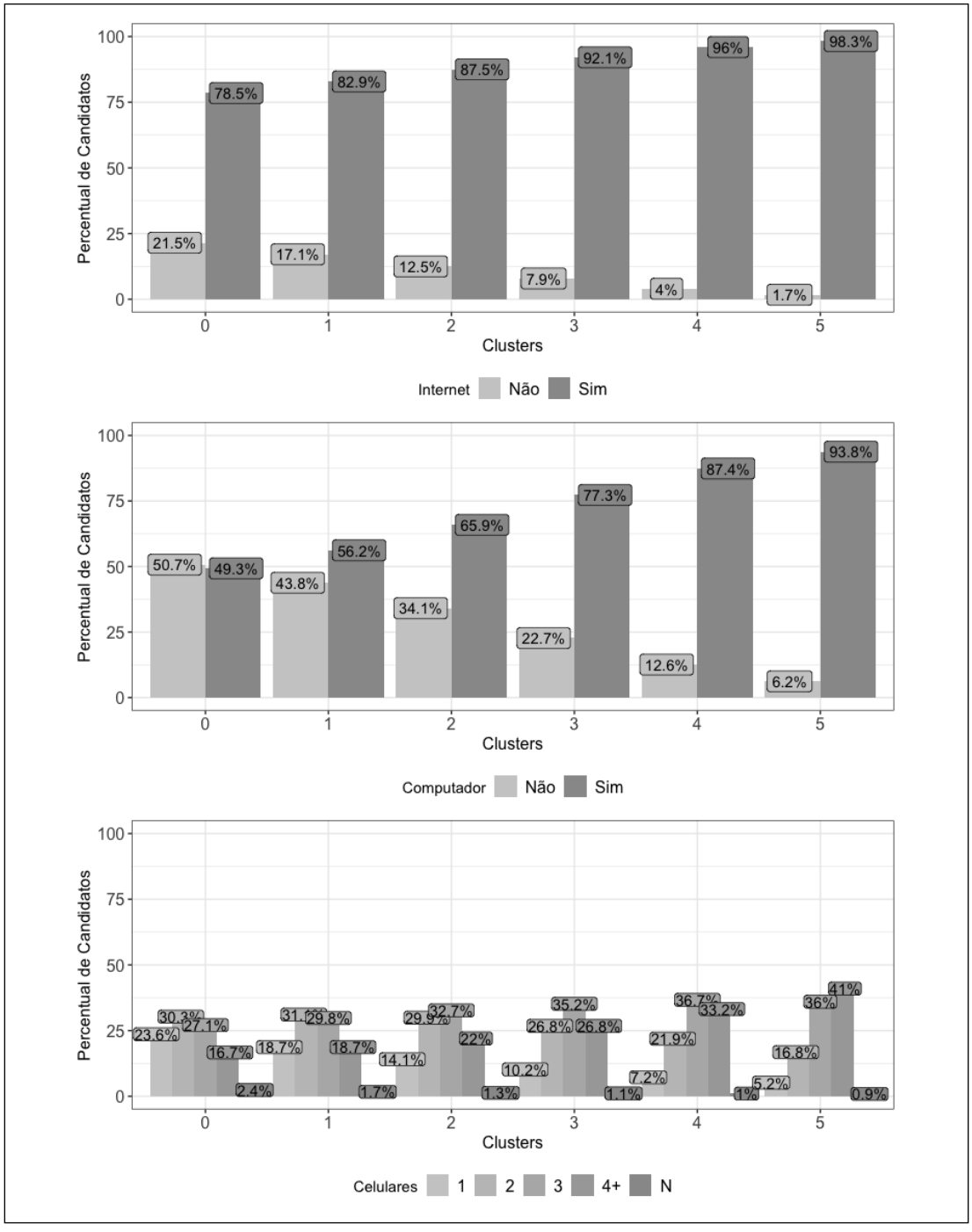
Figura 22 – Perfil dos cluster em relação à renda familiar



Fonte: autoria própria

A Figura 23 apresenta o comportamento quanto ao acesso à internet, posse de computador e posse de celular nos seis clusters.

Figura 23 – Perfil dos cluster em relação à acesso a internet, posse de computadores e posse de celulares



Fonte: autoria própria

Enquanto uma inversão é observada em relação a posse de computador, onde metade dos candidatos do cluster 0 não têm computador em casa e mais de 90% do cluster

5 têm, em relação a internet, apesar de 21,5% não possuir acesso a internet no cluster 0, em todos os clusters o percentual dos que têm acesso a internet é muito grande. Em relação a posse de celulares, nota-se um aumento do percentual das famílias que possuem mais de quatro celulares nos clusters de maior desempenho dos candidatos.

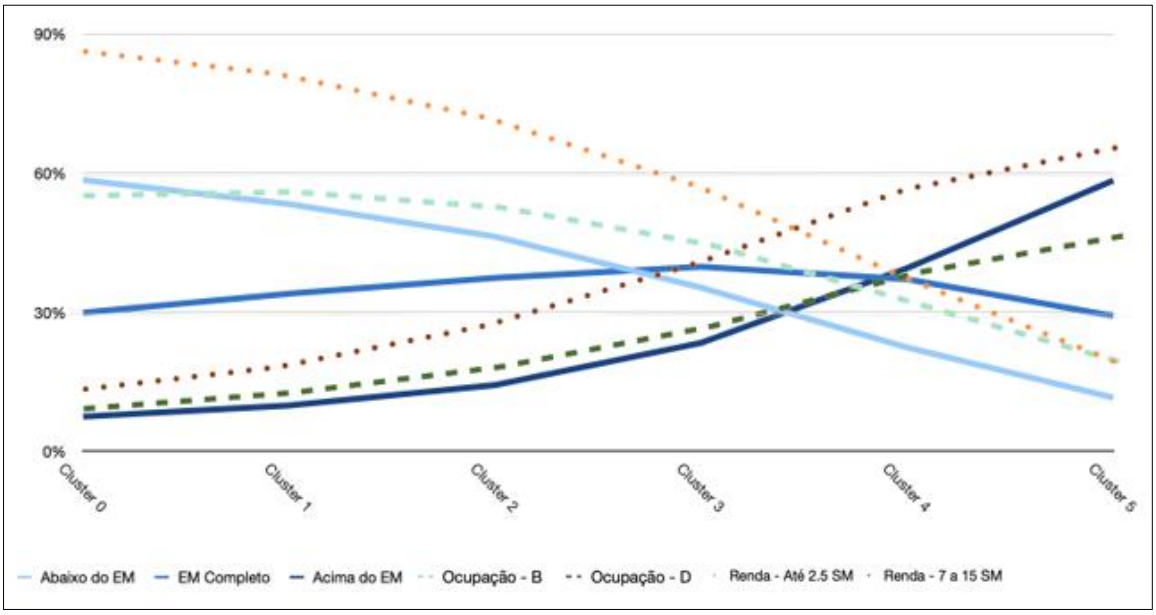
Na Tabela 4 procura-se resumir os perfis socioeconômico dos clusters, onde pode-se observar a redução do percentual de responsáveis cuja escolaridade é abaixo do ensino médio (EM) a medida que o desempenho dos candidatos aumenta (coluna 2) e no sentido inverso observa-se o aumento do percentual cuja escolaridade é superior ao EM (coluna 4). A coluna 5 mostra a ocupação dos responsáveis que apresentou o maior percentual em cada cluster, podendo-se observar que os clusters 0 a 3 tem maior concentração no tipo de ocupação representada por "pedreiro", enquanto os clusters 4 e 5, a maior concentração está um nível acima, representada por "professor de EM". Em alinhamento com as características socioeconômicas anteriores, a renda familiar (coluna 6) mostra que a maior concentração encontra-se na faixa de até 2,5 salários mínimos (SM) nos clusters 0 a 3, enquanto que no cluster 4 a maior concentração está na faixa de até 7 SM e, no cluster 5, na faixa de 7 a 15 SM. Apesar de não apresentar um comportamento tão acentuado como as variáveis anteriormente analisadas, a posse de computadores também segue comportamento similar, onde no cluster 0 menos de 50% possuem computador e no cluster 5 esse percentual é de 94% (coluna 8). Quanto ao acesso a internet e a posse de celular (colunas 7 e 9) observa-se variação bem baixa, indicando pouca influencia no desempenho dos candidatos. A Figura 24 ilustra o perfil socioeconômico dos clusters.

Tabela 4 – Resumo do perfil socioeconômico dos clusters

Cluster	Abaixo EM (%)	EM completo (%)	Acima EM (%)	Ocupação (%)	Renda (%)	Internet SIM	Computador SIM	Celular SIM
0	63	22	4	**C** 29	Até 2,5 SM 86	79	49	98
1	58	27	6	**C** 32	Até 2,5 SM 81	83	56	99
2	51	31	9	**C** 33	Até 2,5 SM 72	88	66	99
3	42	35	17	**C** 30	Até 2,5 SM 57	92	77	99
4	29	36	31	**D** 36	Até 7 SM 78	96	87	99
5	17	30	50	**D** 41	De 7 a 15 SM 65	98	94	99

Fonte: autoria própria

Figura 24 – Sintese do perfil socioeconômico dos cluster



Fonte: autoria própria

Podemos concluir que a escolaridade e ocupação do responsável, a renda e a posse de computador são características socioeconômicas que influenciam no desempenho dos candidatos mas o acesso à internet e a posse de celular, pela sua grande penetração na sociedade, não possui grande influencia em seus desempenhos.

6. Conclusões

Este trabalho analisou o perfil dos candidatos da região sudeste do Brasil que realizaram o ENEM em 2019 buscando identificar a influencia das condições socioeconômicas nos seus desempenhos. Para tal, foram utilizadas técnicas estatísticas de análise exploratória de dados, destacando-se aquelas que permitiram a análise conjunta das variáveis socioeconômicas, como árvore de decisão e clusterização com o algoritmo *K-means*. O ano de 2019 foi selecionado para a análise por ser o concurso mais recente na data de início do desenvolvimento do presente trabalho.

A base de dados analisada continha mais de um milhão de candidatos, mais de 68% estavam na faixa etária entre 16 e 20 anos, 59% eram do gênero feminino, 46% se declararam como brancos e 36% como pardos e apenas 0,4% se declararam como indígena. 58% dos candidatos não responderam se estudaram em escola pública ou privada e a motivação deste percentual deveria ser aprofundada. Mais da metade dos candidatos (52%) tinham renda familiar até R$1.996,00 e apenas 2,5% superior a

R$9.980,00. A maior concentração de escolaridade do responsável encontrava-se na faixa E (completou o EM, mas não concluiu a graduação) (>30%). Quanto à ocupação do pai/mãe/responsável a maior concentração estava nas faixas C (representada por pedreiro) 29% e B (representada por diarista) 47%. Apenas 1,4% não tinha celular e 31% não tinham computador.

O desempenho dos candidatos foi representado pela média aritmética das cinco provas. A raça com maior desempenho mediano foi a branca e a pior a indígena. O desempenho se mostrou fortemente crescente com o aumento da renda familiar, assim como com a escolaridade e ocupação do responsável e, o número de aparelhos celulares e computadores.

Foi possível identificar a formação de seis clusters quanto ao desempenho dos candidatos. A análise das características socioeconômicas destes clusters permitiu traçar os seus perfis socioeconômicos. Por exemplo, o cluster de melhor desempenho é formado por candidatos cujos maiores percentuais estão na renda familiar entre 7 e 15 salários mínimos, ocupação do responsável nível D (ex: prof. EM), escolaridade acima do EM e possuem mais de 4 computadores. Já o cluster com o pior desempenho é formado por candidatos cujos maiores percentuais estão na renda familiar até 2,5 salários mínimos, ocupação do responsável nível B (ex: diarista), escolaridade abaixo do EM e não possuem computadores (51%). O que mostra a importância do perfil socioeconômico no desempenho no ENEM, mostrando a importância de políticas sociais.

Finalmente, é importante destacar que a variável "Tipo de escola no ensino média" não pode ser avaliada neste trabalho devido ao grande percentual de candidatos que não responderam esta pergunta (58%). Investigar a motivação deste comportamento de forma a reduzir este percentual e, assim, poder incluir esta variável em análises futuras nos parece bastante relevante, uma vez que, um dos objetivos do ENEM é avaliar o ensino médio e portanto, identificar possíveis diferenças associadas aos tipos de escola.

Referências

BATISTA, R. Enem 20 anos: a transformação da maior prova do Brasil. Vestibular Brasil Escola, 2018. Disponível em: https://vestibular.brasilescola.uol.com.br/enem/enem-20-anos-transformacao-maior-prova-brasil.html . Acessado em: 01/07/2021.

BATISTA, R. Teoria de Resposta ao Item (TRI) no Enem. Vestibular Brasil Escola, 2018. Disponível em: https://vestibular.brasilescola.uol.com.br/enem/teoria-resposta-ao-item-tri-no-enem.htm Acessado em: 21/07/2021.

BATISTA, R. Treineiros no Enem. Disponível em: https://www. https://vestibular.brasilescola.uol.com.br/enem/treineiros-no-enem.htm Acessado em: 31/07/2021.

C. WUNSCH, D.; XU, R.. **Clustering**. Ed. Wiley-IEEE Press, California, 2008.

ENEM - EXAME NACIONAL DO ENSINO MÉDIO. Wikipédia, 2021. Disponível em: https://pt.wikipedia.org/wiki/Exame_Nacional_do_Ensino_Médio Acessado em: 01/07/2021.

HAIR, J. **Multivariate Data Analysis**. Editora Prentice Hall, New Jersey, 1998.

HOCHHEIM, N. **Engenharia de Avaliações II**: modelos de regressão linear para avaliação de imóveis. GEAP - Universidade Federal de Santa Catarina. Florianópolis. 2011.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Estimativas da População. IBGE, 2019. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?edicao=25272&t=resultados Acessado em: 02/02/2021.

INEP – INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Govbr, 2020. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem Acessado em 01/07/2021.

INEP – INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Gov br, 2020. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_do_enem_2020_-_cartilha_do_participante.pdf Acessado em: 21/07/2021.

LAURETTO S., M. Árvores de decisão, 2020. Disponível em: https://edisciplinas.usp.br/pluginfile.php/4469825/mod_resource/content/1/ArvoresDecisao_normalsize.pdf.

MICROSOFT POWER BI. Microsoft Power BI, 2021. Disponível em: https://learn.microsoft.com/pt-br/power-bi/. Acessado em: 05/06/2021.

MINISTÉRIO DE EDUCAÇÃO E CULTURA - MEC. ENEM – Disponível em: http://portal.mec.gov.br/enemsp2094708791#:~:text=Criado%20em%201998%2C%20o%20Exame,ensino%20m%C3%A9dio%20em%20anos%20anteriores. Acessado em: 15/02/2022.

MIRANDA FREIRE, S. Regressão Linear, 2021. Disponivel em: https://www.lampada.uerj.br/arquivosdb/_book/regress%c3%a3o-linear.html Acessado em: 31/07/2021.

PYTHON SOFTWARE FOUNDATION. Python Language Site: Documentation, 2020. Página de documentação. Disponível em: <https://www.python.org/doc/>. Acessado em: 06/11/2021.

R CORE TEAM (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: https://www.R-project.org/. Acessado em: 06/11/2021.

SOUZA CRUZ, B. Você nunca vai tirar zero na prova do Enem, saiba por quê. Educação UOL, 2014. Disponível em: https://educacao.uol.com.br/noticias/2014/10/31/voce-nunca-vai-tirar-zero-na-prova-do-enem-saiba-por-que.html Acessado em: 17/08/2021.

TUKEY J. W. The Future of Data Analysis. **Ann. Math. Statist**. 33 (1) 1 - 67, March, 1962. Disponível em: https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-33/issue-1/The-Future-of-Data-Analysis/10.1214/aoms/1177704711.full?tab=ArticleLink Acessado em: 30/08/2021

THE PERFORMANCE OF HIGH SCHOOL STUDENTS AT ENEM 2019 AND SOCIAL INEQUALITY: A STUDY USING EXPLORATORY ANALYSIS AND CLUSTERING TECHNIQUES

Abstract

This study aimed to analyze the performance of candidates who performed the ENEM in 2019, to identify possible relationships between such performances and a set of socioeconomic variables, as well as to expose how statistical techniques are valuable tools for extracting information from large databases. The results showed how exploratory data analysis and clustering techniques allowed the identification of six groups of candidates with similar performance socioeconomic characteristics, indicating the socioeconomic influence in the dispute for a vacancy in a quality Higher Education and, consequently, in the labor market, indicating the importance of public policies aimed at leaving, at least a little, this fairer competition.

Keywords: *Performance in ENEM; Exploratory Data Analysis; Cluster Analysis; Kmeans; Social inequality.*