Análise dos Dados Socioeconômicos na Investigação do Impacto da Pandemia da Covid-19 no Enem nas capitais da Região Sudeste

Pedro Gabriel Cruz[1], **Cristiane N. Targa**[1], **Carlos A. Silva**[1]

[1]Departamento de Informática – Instituto Federal de Minas Gerais (IFMG)
CEP – 34590-390 – Sabará – MG – Brasil

pedrogabrielcruz00@gmail.com, {cristiane.targa,carlos.silva}@ifmg.edu.br

Abstract. *The present work aims to identify the impact of the Covid-19 pandemic on Enem in its socioeconomic scope for students who took the exam from 2019 to 2022 by applying Data Mining techniques. The use of the K-means clustering algorithm made it possible to map 3 well-defined groups of students, and thus understand how each group was affected throughout the pandemic through its determining variables: family income, type of school administration, access to computers and internet . The results show that students belonging to the group with lower socioeconomic indicators were those who showed the greatest drop in exam participation and an increase in the participation of students in the group with better socioeconomic indicators, therefore a growth in inequality between participants.*

Resumo. *O presente trabalho tem como objetivo identificar o impacto da pandemia da Covid-19 no Enem em seu âmbito socioeconômico para os estudantes que prestaram o exame nos anos de 2019 a 2022 aplicando técnicas de Mineração de Dados. O uso do algoritmo de clusterização K-means permitiu mapear 3 grupos bem definidos de estudantes, e assim entender como cada grupo foi afetado ao longo da pandemia por meio de suas variáveis determinantes: renda familiar, tipo de administração escolar, acesso a computador e internet. Os resultados mostram que os alunos pertencentes ao grupo com indicativos socioeconômicos mais baixos foram os que apresentaram a maior queda de participação do exame e um aumento na participação de estudantes do grupo com melhores indicativos socioeconômicos, portanto um crescimento da desigualdade entre os participantes.*

1. Introdução

No contexto brasileiro, o Exame Nacional do Ensino Médio (Enem) tem como objetivo avaliar o desempenho dos estudantes ao fim do ensino médio, além de ser utilizado como forma de ingresso à educação superior na maioria das instituições nacionais, também em mais de 50 instituições de educação superior portuguesas e países como Reino Unido, França, Canadá, Estados Unidos e Irlanda [INEP 2023].

A base de microdados utilizada neste trabalho está acessível por meio do portal do INEP[1] (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), onde

[1]https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

é possível acessar o conjunto de informações detalhadas relacionadas às pesquisas, aos exames e avaliações do Instituto. Entre os conjuntos de informação disponíveis, estão os microdados do Enem, que disponibilizam informações específicas de provas, gabaritos, itens, notas e o questionário respondido pelos inscritos no Enem.

Dada essa grande relevância do exame, é importante realizar análises eficientes nos dados obtidos anualmente sobre o conhecimento técnico e de informações socioeconômicas, para permitir medir os níveis de conhecimento desses participantes em seus diferentes contextos. E assim, possibilitar avaliar a eficácia do sistema educacional, identificar desigualdades educacionais e orientar o desenvolvimento de políticas públicas voltadas para a melhoria da qualidade da educação e a promoção da igualdade de oportunidades.

A pandemia da Covid-19 trouxe inúmeros desafios para diversos setores, no Brasil e no mundo. Na tentativa de reduzir a ampla disseminação do novo Coronavírus, medidas de distanciamento social foram adotadas pelos países. Na Educação, tais medidas resultaram no fechamento de escolas públicas e particulares, com interrupção de aulas presenciais. Apesar das atividades educacionais à distancia assumirem um caráter essencial, são evidentes as suas limitações e impactos na experiência escolar [Educação 2020].

Neste estudo foram analisados os dados do Enem dos anos de 2019 a 2022, referentes ao período de pré-pandemia e de pandemia da Covid-19, das capitais da região sudeste do Brasil, Belo Horizonte (MG), Rio de Janeiro (RJ), São Paulo (SP) e Vitória (ES). O objetivo deste trabalho é, por meio de um processo de clusterização de dados, entender como a pandemia impactou os alunos participantes do exame em seus diferentes grupos socioeconômicos.

A partir das análises realizadas foi possível verificar que a pandemia impactou na participação dos estudantes, sendo que alunos mais vulneráveis socioeconomicamente tiveram queda nas participações e piores resultados no exame, enquanto alunos com melhores condições socioeconômicas foram menos afetados e obtiveram melhores resultados.

O trabalho está organizado da seguinte maneira: a Seção 2 destaca publicações relevantes sobre o tema de mineração de dados e desempenho educacional; a Seção 3 apresenta a metodologia empregada, bem como os detalhes dos passos de desenvolvimento utilizados nesta pesquisa. Os resultados alcançados são discutidos na Seção 4 e, por fim, as conclusões e possíveis trabalhos futuros são abordados na Seção 5.

2. Trabalhos Relacionados e Fundamentação Teórica

É possível encontrar na literatura diversos trabalhos que apresentam estudos em relação ao uso da mineração de dados sobre os dados do Enem. A abordagem destes trabalhos permite compreender como o desempenho dos alunos é influenciado e identificar os principais fatores que exercem essa influência.

A pesquisa de [Dutra et al. 2023] fez uma revisão na literatura e identificou estudos publicados nos últimos 10 anos que avaliaram o desempenho dos estudantes no Enem. Por meio desta pesquisa vemos que para realização da mineração e análise dos dados há diversas técnicas, que permitiram concluir que o desempenho dos alunos é influenciado principalmente por questões socioeconômicas, como localização (contexto geográfico) e tipo de administração escolar (pública ou privada).

Algumas pesquisas revisadas por [Dutra et al. 2023] utilizaram técnicas exploratórias, que realizam a análise por meio de recursos de visualização e exploração dos dados. Enquanto outras fizeram o uso de técnicas mais elaboradas, aplicando inteligência artificial e aprendizado de máquina para a realização das análises dos dados.

De acordo com [Sousa 2023], técnicas de Aprendizado de Máquina podem ser divididas em aprendizagem supervisionada e não supervisionada. A aprendizagem supervisionada permite que entradas e saídas de dados sejam definidas por um supervisor, já a não supervisionada apenas os dados de entrada são definidos, enquanto os dados de saída são obtidos por meio da máquina sem a atuação de um supervisor, ou seja, sem a necessidade de intervenção humana por meio de dados rotulados.

Os autores na pesquisa de [Banni et al. 2021] aplicaram técnicas de aprendizado de máquina supervisionado e comparação de algoritmos, para identificar os atributos mais relacionados ao desempenho dos estudantes participantes do Enem 2018. No estudo foi possível observar uma alta relação entre o resultado obtido no Enem com os seguintes atributos: escolaridade dos pais, renda per capita familiar, se o estudante é recém formado, cor e raça, além da faixa etária.

Através de técnicas de aprendizado de máquina não supervisionado, por meio do algoritmo *K-Means* para realizar o agrupamento dos dados, é possível identificar características relevantes nos dados. Assim como realizado no estudo de [Barcellos et al. 2018], que através dos dados do Enem de 2018 mostrou que aspectos como a escolaridade do pai, o acesso à internet e a renda da família, tem indícios de afetarem a média do estudante.

A pesquisa de [Lima et al. 2020] analisou dados do Enem entre 2012 e 2017, e obteve por meio do algoritmo de agrupamento *K-Means*, grupos caracterizados como Baixo Desempenho, Desempenho Mediano e Alto Desempenho para cada região do Brasil, assim como os atributos mais presentes em cada grupo: tipo de escola, faixa de nota em cada grupo, presença de alunos deficientes.

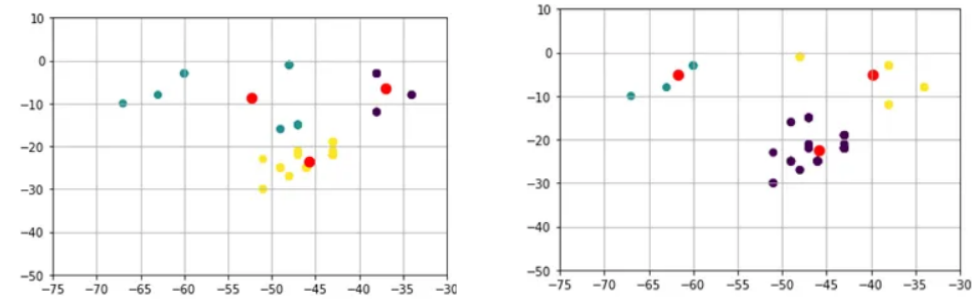
O *K-Means*, de acordo com [Sousa 2023], é um algoritmo de clusterização de dados que atua por meio de aprendizagem não supervisionada e é utilizado para particionar dados em agrupamentos distintos. Esse algoritmo tem o objetivo de identificar os elementos de um conjunto de dados que tendem a ser mais semelhantes entre si, por compartilhar características importantes e parecidas. A identificação das semelhanças ocorre por meio de cálculos iterativos para selecionar os elementos mais próximos de k centros.

De maneira simplificada, o *K-Means* possui o seguinte funcionamento:

1. O primeiro passo é definir um K número de grupos desejados;
2. Após essa definição, é atribuído de forma aleatória um centróide para cada *cluster*;
3. Cada ponto é associado ao centróide mais próximo comparados ao demais centroides dos agrupamentos;
4. Ao executar o passo anterior, diversos pontos são realocados entre os *cluster*, e por isso o centroide é reposicionado de acordo com a média da posição de todos os pontos de grupo;
5. Os dois passos anteriores são repetidos até que os centroides não sejam alterados, com isso obtém-se uma posição ótima dos centroides.

Um modelo gráfico de iteração para clusterização de elementos a partir da

definição dos centróides pode ser visto na Figura 1. No primeiro gráfico, Figura 1(a), é possível ver a ocorrência dos passos 1, 2 e 3. O passo 4 é exibido na Figura 1(b), onde os centróides foram reposicionados de acordo com a média da posição dos demais pontos.



(a) Distribuição dos elementos após a iteração.

(b) Reposicionamento dos centroides e clusterização final dos elementos após a iteração.

Figura 1. Iteração de clusterição do K-means
Fonte: [Anastacio 2020]

Dentre os algoritmos de clusterização, o *K-Means* é um dos precursores e mais conhecidos, apesar de ter mais de 50 anos de surgimento, ainda atualmente é um dos algoritmos mais utilizados devido a sua simplicidade de implementação e qualidade dos resultados alcançados com sua aplicação.

3. Metodologia

O método para descoberta de conhecimento em base de dados utilizado neste estudo foi o *Knowledge Discovery in Databases* (KDD). De acordo com [da Silva et al. 2020], esse método possui as etapas: seleção de dados, pré-processamento, transformação de dados, mineração de dados e análise dos resultados. A Figura 2 apresenta como foram realizadas cada etapa deste estudo.

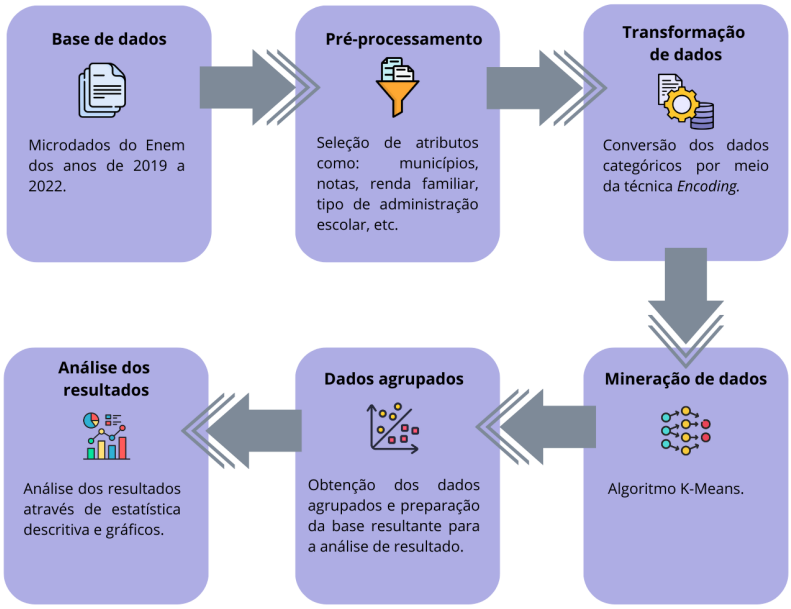


Figura 2. Diagrama de etapas KDD.
Fonte: Autores

3.1. Base de dados

A base de dados dos microdados do Enem dos anos de 2019, 2020, 2021 e 2022 estão disponíveis no formato .csv, totalizando aproximadamente 7 GB de dados. A Tabela 1 apresenta os números gerais desse conjunto de dados. Os detalhes dos procedimentos para filtrar os dados não relevantes para esse estudo são apresentados na subseção 3.2.

Ano	Antes do pré-processamento		Depois do pré-processamento	
	Num. de linhas	Num. de colunas	Num. de linhas	Num. de colunas
2019	5095171	76	80028	19
2020	5783109		50422	
2021	3389832		62165	
2022	3476105		70049	

Tabela 1. Números gerais do conjunto de dados.

3.2. Pré-processamento e tratamento de dados

A fase de pré-processamento é a etapa de preparação dos dados, portanto envolve selecionar, eliminar e transformar os dados para serem usados no projeto. Para realizar a

tarefa de pré-processamento, assim como a de processamento, foi utilizado a linguagem de programação Python, através da versão 3.10.12.

O primeiro passo foi filtrar o conjunto de dados de cada ano, inicialmente os dados apresentavam a dimensionalidade conforme Tabela 1. Em seguida foram filtrados os dados de apenas alunos pertencentes às capitais da região sudeste do Brasil (Belo Horizonte, Rio de Janeiro, São Paulo e Vitória). Os atributos foram selecionados a partir daqueles que apresentavam maior relevância com as questões investigadas nesta pesquisa, com base nos estudos dos trabalhos relacionados, como: as notas e principais características socioeconômicas (renda familiar, tipo de administração escolar, acesso a computador e internet, etc). Estes atributos estão listados na Tabela 2.

Após a seleção dos dados relevantes, os atributos que não interessavam a este trabalho foram descartados, assim como valores nulos. Com isso, obtivemos um decréscimo de cerca de mais de 98% nos números de linha de cada anos após o pré-processamento e de 75% no número de colunas.

Coluna	Tipo de dado	Descrição
NU_INSCRICAO	Numérico	Número de inscrição do candidato mascarado por ano, não permite identificar o mesmo participante em anos diferentes.
NU_ANO	Numérico	Ano do Enem
TP_FAIXA_ETARIA	Numérico	Faixa etária do candidato
TP_SEXO	Alfanumérico	Sexo
NO_MUNICIPIO_ESC	Alfanumérico	Nome do município da escola
SG_UF_ESC	Alfanumérico	Sigla da Unidade da Federação da escola
TP_DEPENDENCIA_ADM_ESC	Numérico	Dependência administrativa da escola (Federal, Estadual, Municipal e Privada)
NU_NOTA_CN	Numérico	Nota da prova de Ciências da Natureza
NU_NOTA_CH	Numérico	Nota da prova de Ciências Humanas
NU_NOTA_LC	Numérico	Nota da prova de Linguagens e Códigos
NU_NOTA_MT	Numérico	Nota da prova de Matemática
TP_LINGUA	Numérico	Língua Estrangeira
NU_NOTA_REDACAO	Numérico	Nota da prova de redação
Q001	Alfanumérico	Até que série seu pai, ou o homem responsável por você, estudou?
Q002	Alfanumérico	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q005	Alfanumérico	Incluindo você, quantas pessoas moram atualmente em sua residência?
Q006	Alfanumérico	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
Q024	Alfanumérico	Na sua residência tem computador?
Q025	Alfanumérico	Na sua residência tem acesso à Internet?

Tabela 2. Atributos selecionados.

Os dados que eram categóricos, ou seja, aqueles do tipo alfanuméricos apresentados na Tabela 2, precisaram ser convertidos para numéricos por meio das técnicas de *Encoding*: *one-hot-encoding* e *label encoding*. Essas técnicas recompõem as informações sem descaracterizar os dados originais conforme exemplos apresentado na Tabela 3, permitindo submeter os dados em um algoritmo de agrupamento, como o *K-Means*.

Coluna	Dados originais	Dados Convertidos
Q024	A	1
	B	2
	C	3

Coluna Original	TP_SEXO	M	F
Coluna Conversão	TP_SEXO_M	1	0
	TP_SEXO_F	0	1

Tabela 3. Exemplos de conversão de dados categóricos para numéricos.

3.3. Mineração de dados: Agrupamento

A técnica de Agrupamento ou *Clustering* consiste em agrupar os dados que são semelhantes entre si, e dessa forma são reunidos em subconjuntos, ou *clusters* [Sousa 2023]. Para a realização da tarefa de clusterização desta pesquisa foi utilizado o *K-Means*, que é um algoritmo de aprendizagem de máquina não supervisionado, onde os dados são divididos em k agrupamentos distintos, onde k é um número a ser definido de agrupamentos.

O Método do Cotovelo discutido inicialmente por [Thorndike 1953], auxilia na definição de k, ou seja o número ótimo de *cluster* a ser utilizado. Por meio da soma da distância euclidiana ao quadrado de cada ponto ao centro do *cluster* (SSE - *Sum Square Error*) é possível verificar a distorção do agrupamento realizado. Portanto, o método consiste em incrementar k até que ocorra uma grande redução do SSE, pois o melhor valor para k é aquele que apresenta a menor redução de SSE [Lima et al. 2020]. A forma de análise e aplicação do método é ilustrado na Figura 3.

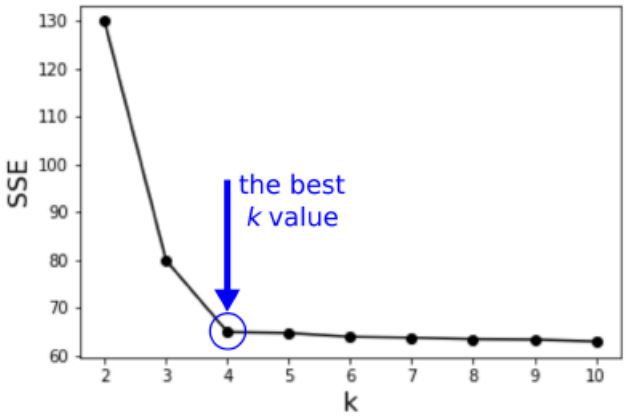


Figura 3. Gráfico de SSE
Fonte: [Lima et al. 2020]

Aplicando o Método do Cotovelo aos dados do Enem a serem utilizados, é possível verificar na Figura 4 que com 3 agrupamentos conseguimos uma grande redução na distorção, e também com base na análise semântica dos resultados de agrupamento. Sendo assim, utilizaremos $K = 3$ no algoritmo *K-means* para os dados em questão.

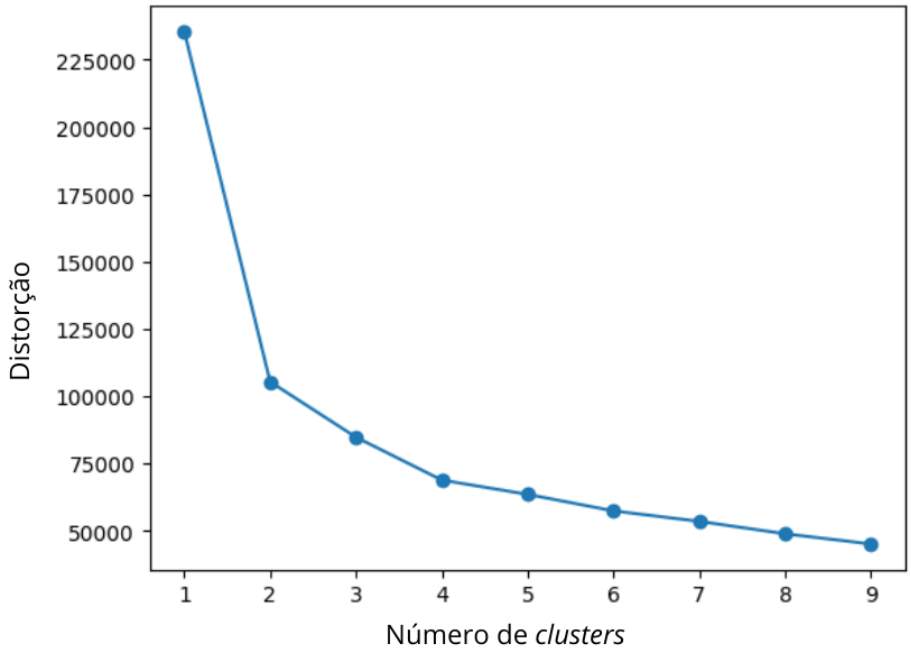


Figura 4. Gráfico de SSE por número de *cluster* obtido do algoritmo *K-Means*
Fonte: Autor

Para realizar o agrupamento de dados o atributo de nota considerado foi o de nota média, obtido por meio do cálculo da média aritmética das avaliações. Devido ao atributo de nota possuir uma escala variante de 0 a 1000 pontos, foi necessário o uso da técnica de normalização de dados Min Max, para que os dados fiquem em um intervalo entre 0 e 1.

Também foram desconsiderados no algoritmo de clusterização os dados de município dos alunos, que caracterizavam cada capital da região sudeste (Belo Horizonte, São Paulo, Rio de Janeiro e Vitória). Pois, durante os testes iniciais com a utilização dos dados de localização, para o algoritmo *K-Means* esses dados estavam representando maior significância para o agrupamento do que os atributos socioeconômicos. Portanto, as análises de agrupamento não foram consideradas por cada capital.

4. Análise dos resultados

Nesta seção são apresentados os resultados obtidos de acordo com as análises descritivas ou exploratória dos dados, para uma percepção geral dos atributos, e também a análise resultante do processo de clusterização aplicado.

4.1. Análise Exploratória

Como ponto de partida da análise foi realizada uma exploração inicial dos dados para descobrir padrões, identificar diferenças, testar e validar hipóteses com a ajuda de estatística descritiva e representações gráficas. O objetivo dessa etapa do estudo foi explorar dados para a extração de informações relevantes, e assim compreender o impacto da pandemia no Enem.

Para conduzir as análises foram levantadas algumas questões de pesquisa, que buscam compreender a dinâmica socioeconômica dos alunos participantes do Enem durante

a pandemia do Coronavírus. As questões elaboradas para a etapa exploratória foram:

1. No período considerado, qual foi a variação do número de alunos participantes nos dois dias de prova do Enem?
2. Como ficou a distribuição da quantidade de estudantes por tipo de instituição do ensino médio (federal, municipal, estadual e privada)?
3. A participação dos alunos foi afetada de acordo com a sua renda familiar?

Para responder à primeira questão, a Figura 5 mostra a quantidade de alunos participantes nos dois dias de prova em cada ano considerado do exame, relacionados às capitais da região sudeste do Brasil.

É possível ver que houve uma redução significativa na participação de alunos, principalmente em 2020 que foi o primeiro ano da pandemia da Covid-19. Também é notório, que até o ano de 2022 o número de alunos que realizaram o Enem ainda era consideravelmente inferior ao último ano antes da pandemia.

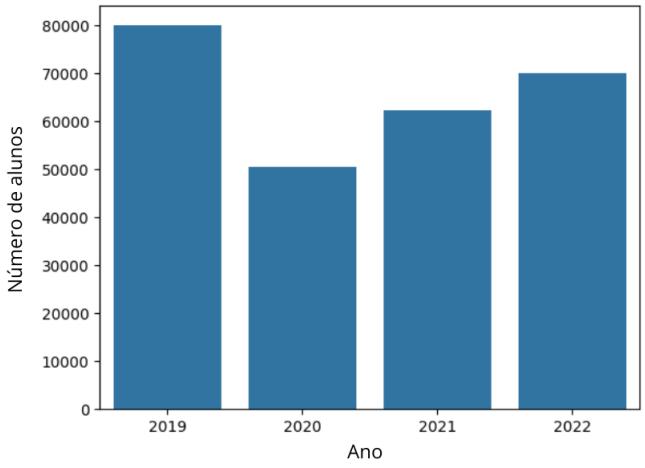


Figura 5. Número de alunos por edição do Enem
Fonte: Autores

Com relação ao número de estudantes que realizaram o exame, nas capitais de modo geral tivemos um comportamento semelhante no decorrer dos anos analisados, como é possível ver na Figura 6.

A distribuição de alunos por tipo de escola em cada ano está representada na Figura 7, e permite responder à segunda questão.

No geral, as escolas municipais oferecem ensino fundamental I e II enquanto que as escolas estaduais oferecem ensino médio e assim possuem um maior número de alunos participantes. Portanto, os alunos de escolas públicas, principalmente da administração estadual, foram os maiores impactados durante a pandemia do Coronavírus. Enquanto, as escolas da rede privada tiveram um aumento na quantidade de alunos participantes do Enem considerando os mesmos períodos.

A questão 3 é respondida pela Figura 8, que permite avaliar a evolução da distribuição da renda familiar dos alunos.

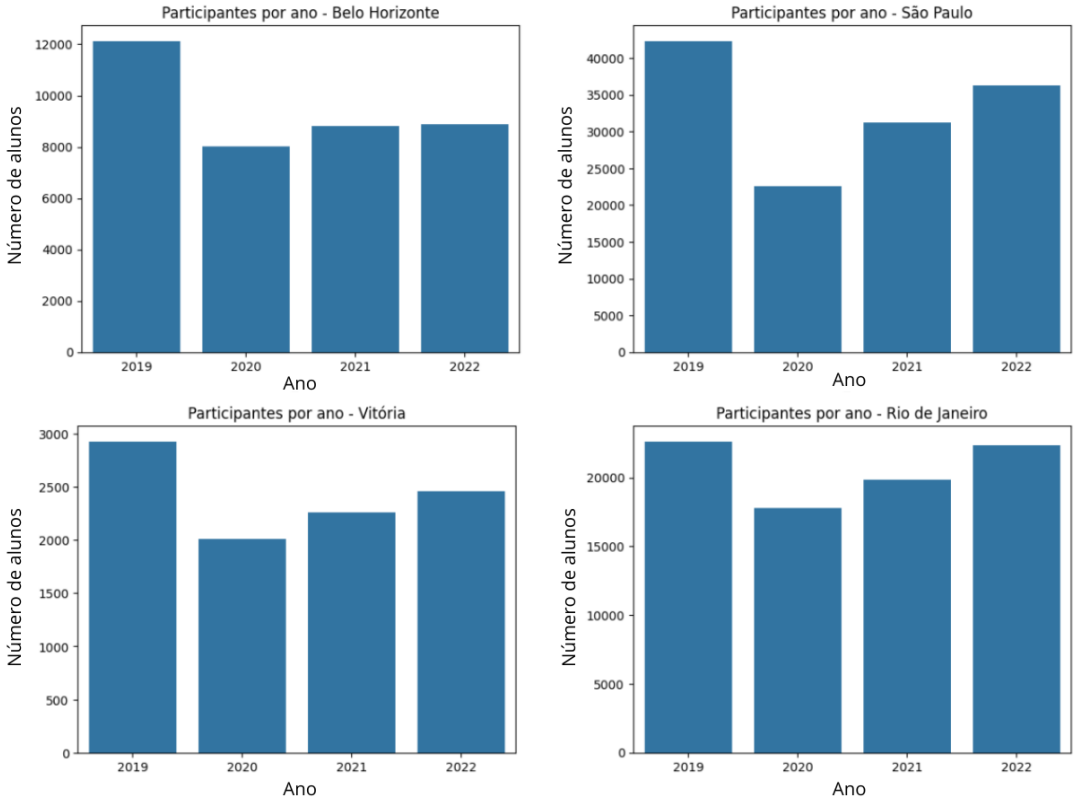


Figura 6. Número de alunos em cada capital da região sudeste por edição do Enem
Fonte: Autores

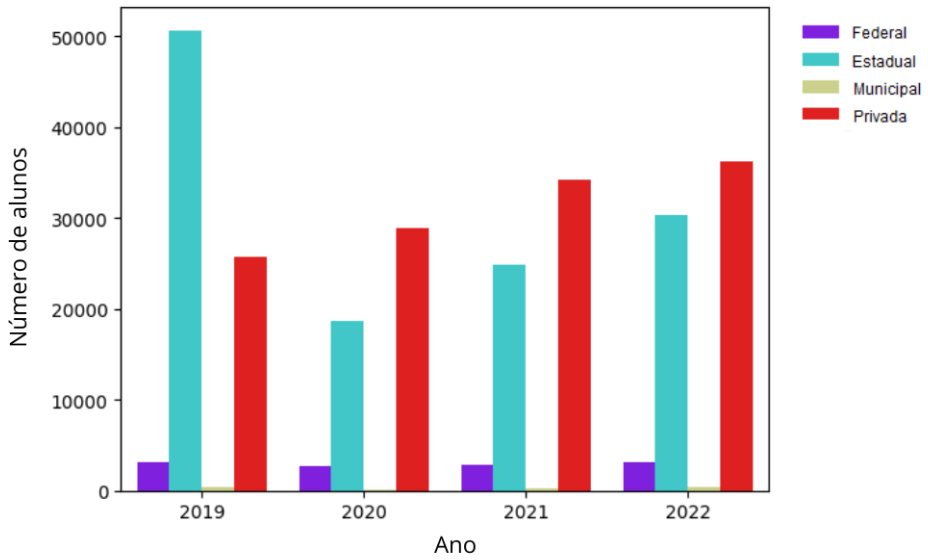


Figura 7. Quantidade de aluno por tipo de administração escolar
Fonte: Autores

Quando observada a mediana dos gráficos da Figura 8, percebe-se que 50% dos alunos participantes do Enem durante os anos da pandemia possuíam uma renda familiar de até 4 salários mínimo.

É possível verificar que houve um aumento nos valores de todos os *quartis* durante os anos da pandemia na Figura 8, isso significa que alunos com rendas mais baixas passaram a ter menor frequência na participação do exame.

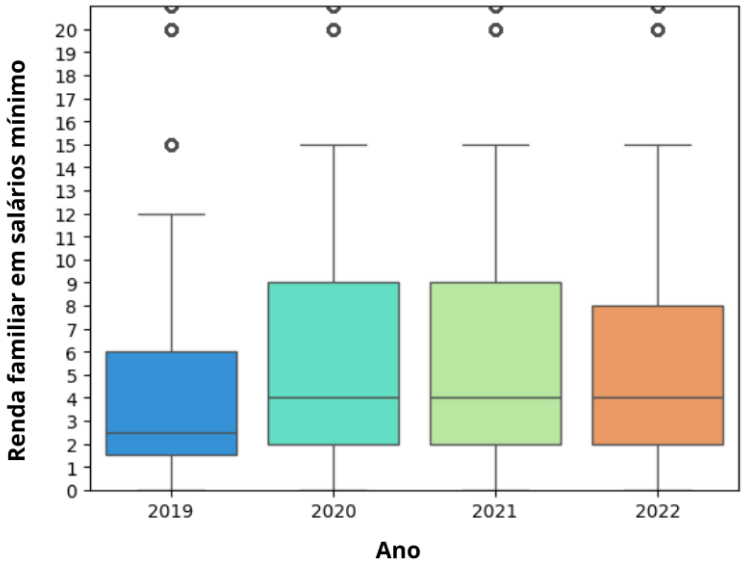


Figura 8. Evolução da renda familiar dos alunos por ano
Fonte: Autores

4.2. Análise do Agrupamento

Por meio do algoritmo de clusterização de dados *K-Means*, foi realizado o agrupamento dos dados semelhantes, para identificar as características específicas de cada grupo de alunos participantes do Enem antes da pandemia, no ano de 2019, e durante a pandemia, nos anos de 2020, 2021 e 2022.

Assim como na análise exploratória, alguns questionamentos foram levantados para compreender como a pandemia de Covid-19 impactou os alunos por meio de uma perspectiva socioeconômica. As questões elaboradas para guiar a análise foram as seguintes:

1. Qual foi a distribuição de nota média por aluno em cada *cluster* para cada ano do período considerado?
2. Quais são as características de cada grupo com relação à renda familiar?
3. Quais são as características de cada grupo com relação ao tipo de administração escolar?
4. Quais são as características de cada grupo com relação ao acesso a computador e internet?

A primeira questão é respondida por meio dos gráficos de distribuição de notas médias por *cluster* apresentados nas Figuras 9, 10, 11 e 12. Que permitem verificar por

meio dos *quartis* de cada gráfico, que há uma forte relação entre a nota média e o agrupamento realizado, sendo que os *clusters* ficaram distribuídos da seguinte maneira: o *cluster* 0 com a concentração do grupo de alunos com as notas mais baixas, o *cluster* 1 concentram-se os alunos com as notas intermediárias e o *cluster* 2 aqueles com notas mais elevadas, ou seja, superiores.

Com relação aos anos, o Enem de 2019 apresentou uma quantidade próxima de alunos presentes no *cluster* 0 e 1, sendo o 1 o grupo com o maior número de alunos. Enquanto, durante os anos de pandemia do coronavírus, observou-se que os alunos do *cluster* 0 foram aqueles que mais sofreram queda no número de participantes no exame, já o *cluster* 2 foi o agrupamento que obteve um número crescente de estudantes em todo o período analisado.

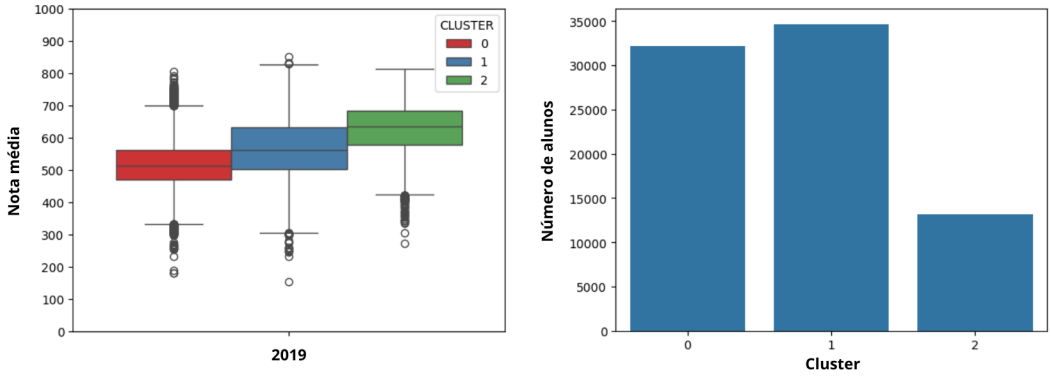


Figura 9. Distribuição de nota média dos alunos por *cluster* no ano de 2019
Fonte: Autores

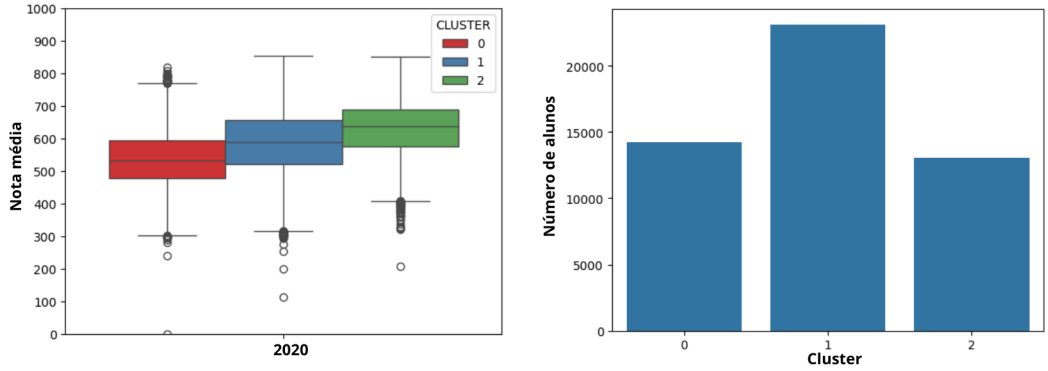


Figura 10. Distribuição de nota média dos alunos por *cluster* no ano de 2020
Fonte: Autores

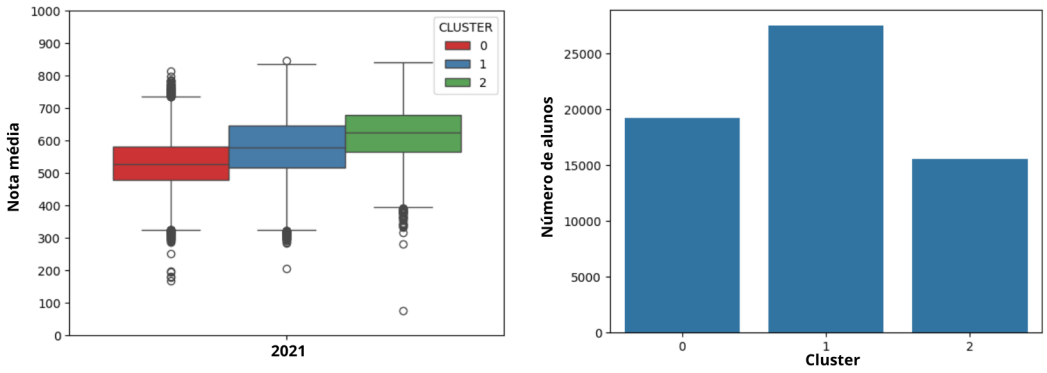


Figura 11. Distribuição de nota média dos alunos por *cluster* no ano de 2021
Fonte: Autores

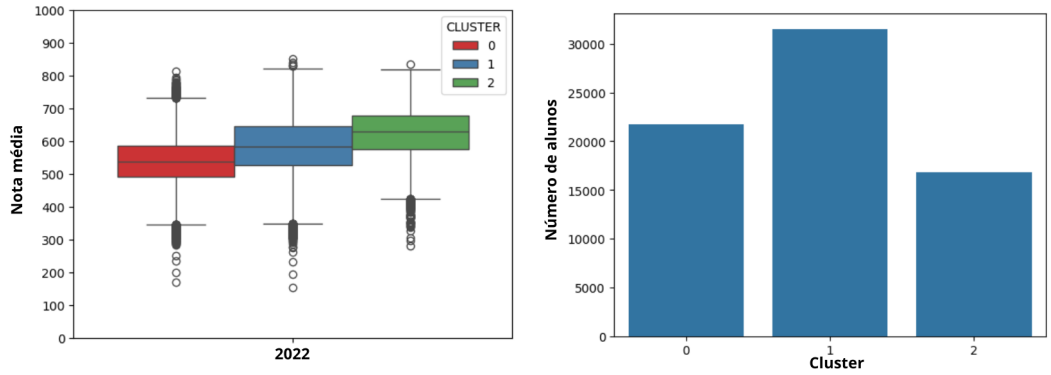


Figura 12. Distribuição de nota média dos alunos por *cluster* no ano de 2022
Fonte: Autores

A questão 2, que aborda sobre as características de renda familiar de cada *cluster* do agrupamento, é respondida por meio dos gráficos apresentados na Figura 13. Esses gráficos mostram uma relação direta entre o agrupamento realizado pelo algoritmo e a renda familiar dos alunos no período considerado, conforme detalhado adiante.

Por meio da análise dos gráficos da Figura 13 é evidente que a maior parte dos alunos que compõe o *cluster* 0 apresentam uma renda mais baixa em comparação aos outros agrupamentos. Através da análise histórica, é possível identificar que o *cluster* 0 sofreu uma variação significativa no período analisado, onde após o ano de 2019 houve aumento na mediana e 3º quartil, e também aumento do limite superior com relação à renda familiar dos alunos do grupo 0.

No *cluster* 1, de acordo com o gráfico da Figura 13 está agrupado os alunos com uma renda familiar intermediária, representado pela cor azul. Por meio da análise histórica do período considerado, observa-se um aumento da mediana e terceiro quartil da renda familiar dos estudantes participantes do Enem. Ou seja, alunos de rendas um pouco maiores passaram a ser a maioria no *cluster* 1, portanto alunos de rendas menores passaram a ser menos frequentes no exame durante o período de pandemia.

Os alunos que apresentam uma renda familiar mais alta estão concentrados no agrupamento do *cluster* 2, representado pela cor verde na Figura 13. Através da análise

de todo o período considerado, verificou-se que o *cluster* 2 foi o que manteve-se mais constante, ou seja, sofreu pouca variação com relação a renda familiar dos alunos presentes neste agrupamento. Isso significa que alunos de rendas mais altas foram os menos impactados com relação à participação no Enem durante o período de pandemia.

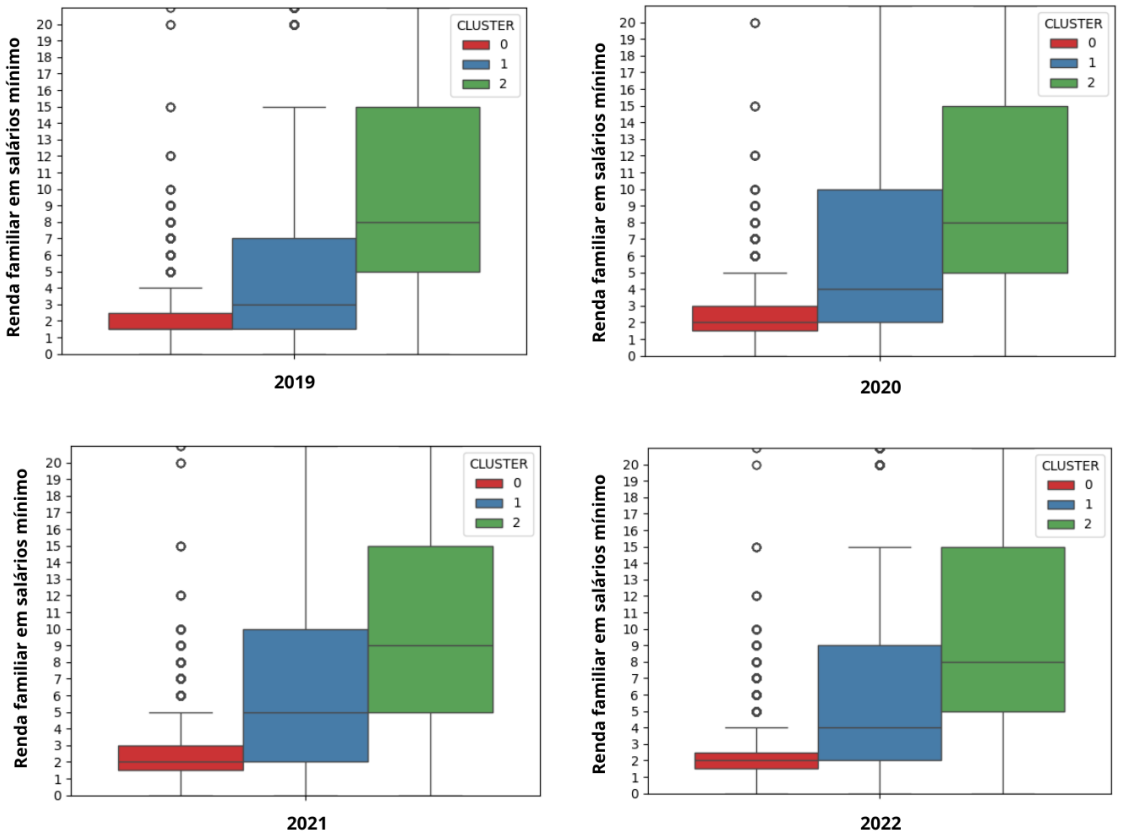


Figura 13. Renda familiar por *cluster* em cada ano
Fonte: Autores

Para o tipo de administração escolar, tema da terceira questão desta seção do estudo, assim como nota média e renda, observa-se uma relação direta com os agrupamentos obtidos por meio do *K-Means*. Os gráficos da Figura 14 ilustram a quantidade de alunos para cada tipo de administração escolar em cada *cluster* para os anos analisados.

De modo geral, o *cluster* 0 contém a maioria dos alunos do tipo de administração escolar da rede estadual em todos os anos considerados. Além disso, foi o *cluster* que apresentou a maior queda no número de participantes durante a pandemia com relação ao número de alunos presente no grupo em comparação ao ano de 2019.

No grupo do *cluster* 2 estão a maioria dos estudantes da rede privada, com base nos gráficos é possível verificar um aumento significativo na quantidade de estudantes deste grupo em todo o período analisado.

Durante a pandemia, nos anos de 2020, 2021 e 2022, os alunos de instituições privadas ultrapassaram o número de alunos de escolas estaduais no *cluster* 1, enquanto em 2019 os estudantes da rede estadual eram o maior número no mesmo *cluster*. Portanto, é uma evidência de que os alunos de escolas estaduais foram os mais afetados na participação do Enem durante a pandemia.

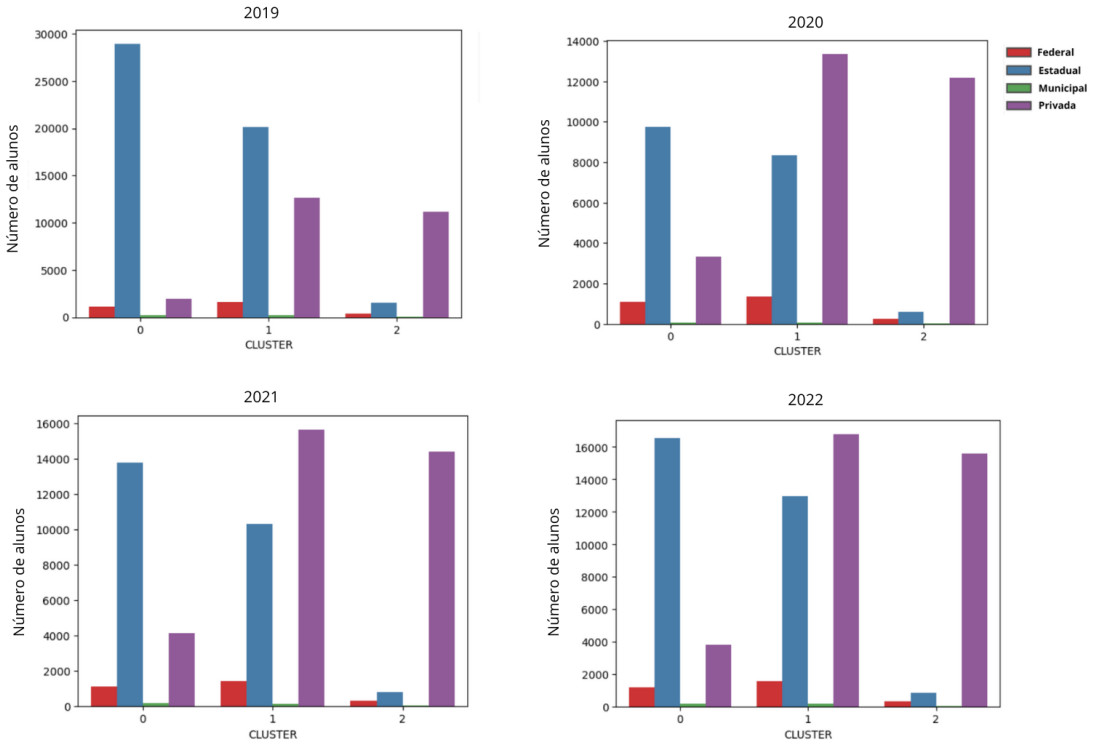


Figura 14. Número de alunos por dependência administrativa escolar por *cluster* em cada ano.
Fonte: Autores

Devido a necessidade do isolamento social durante a pandemia de Covid-19, as instituições de ensino de todo o Brasil precisaram adotar o modelo remoto de ensino [Educação 2020]. Portanto, computador e principalmente acesso a internet, se tornaram ferramentas fundamentais para os alunos, e por isso esta característica foi pautada na questão 4.

De acordo com os gráficos das Figuras 15, 16, 17 e 18 houve um aumento de modo geral nos acessos a computador e internet, em comparação entre o período de pandemia e pré-pandemia (ano de 2019). O grupo que apresentou uma maior defasagem nos acessos no decorrer dos anos foi o cluster 0, seguido do *cluster* 1.

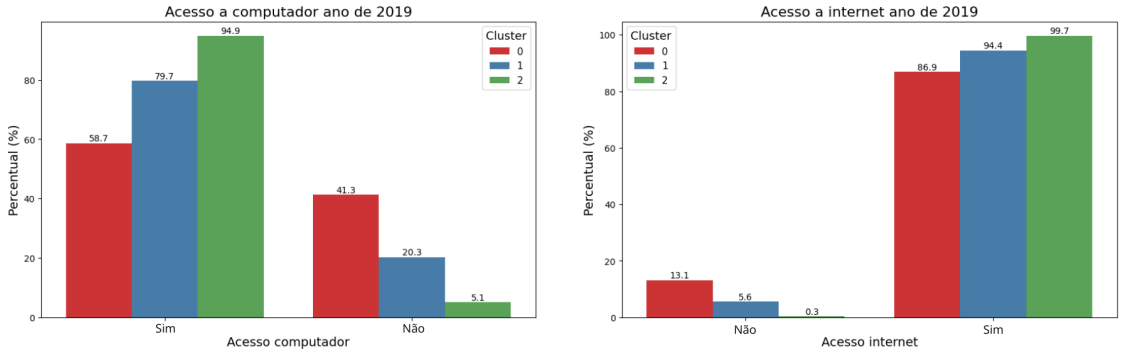


Figura 15. Acesso dos alunos a computador e internet em cada cluster no ano de 2019.
Fonte: Autores

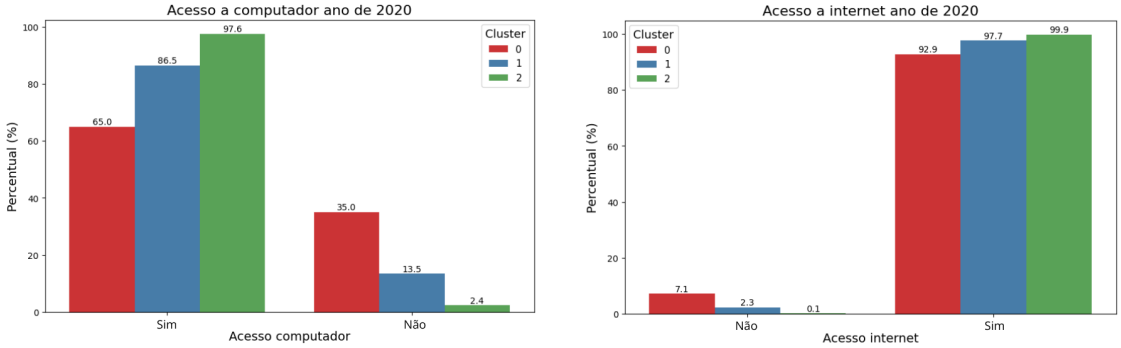


Figura 16. Acesso dos alunos a computador e internet em cada *cluster* no ano de 2020.
Fonte: Autores

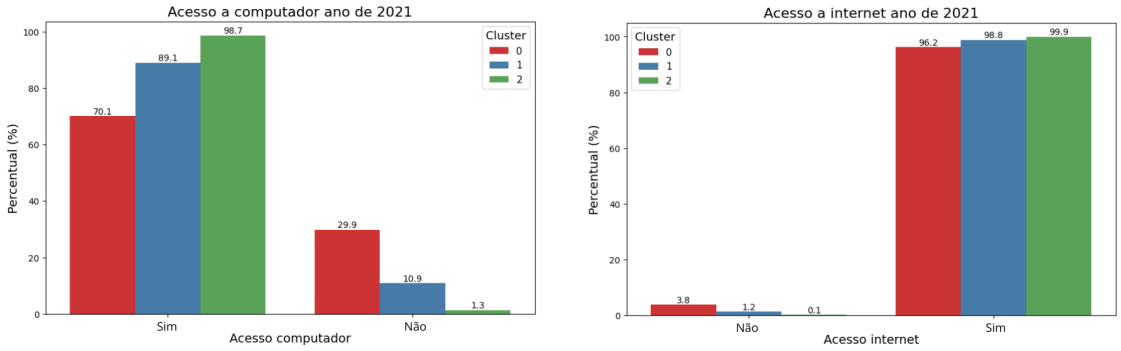


Figura 17. Acesso dos alunos a computador e internet em cada *cluster* no ano de 2021.
Fonte: Autores

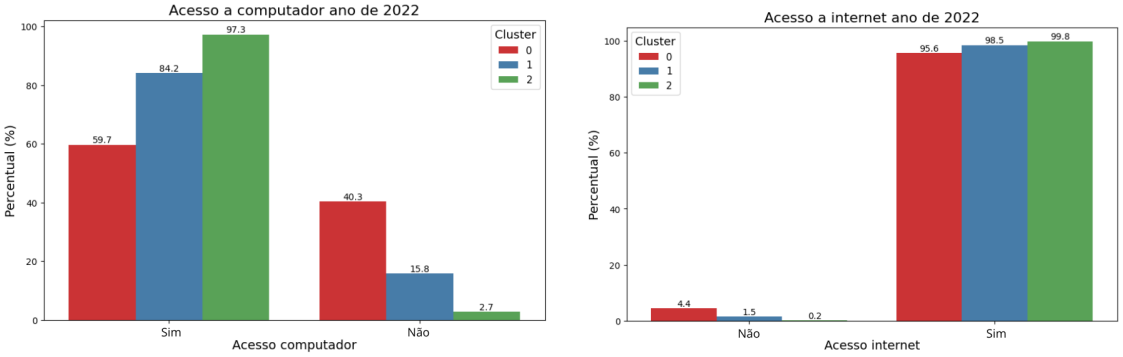


Figura 18. Acesso dos alunos a computador e internet em cada *cluster* no ano de 2022.
Fonte: Autores

4.3. Discussões

De acordo com a análise exploratória, tivemos queda no número de participantes do Enem para as capitais da região sudeste durante os anos de pandemia em relação ao ano de pré-pandemia.

A partir da análise de agrupamento foi possível identificar que o *cluster* 0 foi o grupo que mais sofreu queda no número de participantes, as principais características deste grupo são: nota média mais baixa, menores valores de renda familiar, estudantes da rede estadual de ensino, menor taxa de acesso a internet e computador.

Em contraponto, conforme apresentado na Figura 7 a participação de alunos de escola da rede de ensino privada apresentou um crescimento durante os anos de pandemia, assim como o ocorrido para o grupo do *cluster* 2 na análise de agrupamento. Este grupo de alunos foram caracterizados pelos seguintes atributos: nota média mais alta, maiores valores de renda familiar, estudantes da rede privada de ensino, maior taxa de acesso a internet e computador.

5. Conclusão

O acesso a educação de qualidade é um fator primordial para a redução das desigualdades sociais de um país. Portanto, a análise do desempenho escolar por meio da perspectiva socioeconômica é um instrumento muito útil para apoiar discussões sobre os desequilíbrios que permeiam as desigualdades sociais [da Silva et al. 2020].

Este trabalho utilizou análise de dados educacionais, e identificou informações relevantes para responder se a pandemia de Covid-19 impactou os estudantes participantes do Enem das cidades de Belo Horizonte, Rio de Janeiro, São Paulo e Vitória, ou seja, as capitais da região sudeste do Brasil. Para isso, a pesquisa apresentou resultados de análise exploratória e análise de agrupamento, utilizando o algoritmos *K-Means* para obtenção da clusterização dos dados do Enem de 2019, 2020, 2021 e 2022.

Foi constatado que a presença dos estudantes no exame foi afetada durante o período de pandemia. Em relação aos grupos, identificou-se que há uma relação direta entre os parâmetros socieconômicos (renda familiar, tipo de administração escolar e acesso a computador e internet) e o agrupamento obtido. Com isso, foi possível identificar que alunos com melhores condições socioeconômicas foram menos afetados durante a pandemia, além de possuírem melhores resultados no Enem. Enquanto, alunos mais vulneráveis socioeconomicamente tiveram uma queda bem considerável na participação do exame, assim como piores resultado no Enem.

Trabalhos futuros, a partir deste estudo, podem analisar também o cenário pós-pandemia através da análise dos dados dos alunos participantes do Enem de 2023 e verificar se houve um cenário de reequilíbrio com relação a participação dos estudantes. É possível também, analisar as características dos alunos que estão acessando o ensino público superior, afim de entender como melhorar políticas públicas educacionais para trazer um equilíbrio de acessos à educação de qualidade no país.

Referências

Anastacio, B. (2020). K-means: o que é, como funciona, aplicações e exemplo em python. https://medium.com/programadores-ajudando-programadores/k-means-o-que-%C3%A9-como-funciona-aplica%C3%A7%C3%B5es-e-exemplo-em-python-6021df6e2572 Acessado em 13 de novembro de 2023.

Banni, M. R., Oliveira, M. V. d. P., and Bernardini, F. C. (2021). Uma análise experimental usando mineração de dados educacionais sobre os dados do enem para identificação de causas do desempenho dos estudantes. In *Anais do II Workshop sobre as Implicações da Computação na Sociedade*, pages 57–66. SBC.

Barcellos, A. A., Isotani, S., Diego, C., and Damasceno, N. (2018). Mineraçao de dados abertos-enem 2018. *Anais dos Trabalhos de Conclusao de Curso da Pós-Graduaçao em Computaçao Aplicadaa Educaçao*.

da Silva, V. A. A., Moreno, L. L. O., Gonçalves, L. B., Soares, S. S. R. F., and Júnior, R. R. S. (2020). Identificação de desigualdades sociais a partir do desempenho dos alunos do ensino médio no enem 2019 utilizando mineração de dados. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 72–81. SBC.

Dutra, J. F., Júnior, J. B. F., and de Souza Fernandes, D. Y. (2023). Fatores que podem interferir no desempenho de estudantes no enem: uma revisão sistemática da literatura. *Revista Brasileira de Informática na Educação*, 31(1):323–351.

Educação, T. P. (2020). Ensino a distância na educação básica frente à pandemia da covid-19. *Nota Técnica*, page 15.

INEP (2023). Inep. https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem Acessado em 20 de Agosto de 2023.

Lima, A., Florez, A., Lescano, A., Novaes, J., Martins, N., Junior, C. T., Sousa, E., Júnior, J. R., and Cordeiro, R. (2020). Analysis of enem's attendants between 2012 and 2017 using a clustering approach. *Journal of Information and Data Management*, 11(2).

Sousa, M. C. C. (2023). Uma análise do algoritmo k-means como introdução ao aprendizado de máquinas.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.