https://doi.org/10.35685/revintera.v7i1.4077

MINERAÇÃO DE MICRODADOS PARA ANÁLISE SOCIOECONÔMICA COM PERFORMANCES DO EXAME NACIONAL DO ENSINO MÉDIO - ENEM

MINING MICRODATA FOR SOCIOECONOMIC ANALYSIS USING NATIONAL HIGH SCHOOL EXAM (ENEM) PERFORMANCE

Hernandes Carneiro Macedo[1]

Anna Paula de Sousa Parente Rodrigues[2]

Resumo: O Exame Nacional do Ensino Médio (Enem) é o principal instrumento de acesso ao ensino superior no Brasil e tem papel fundamental na democratização e distribuição mais justa de oportunidades, sendo muitas vezes considerado um verdadeiro símbolo de meritocracia pelos cidadãos brasileiros. Nesse contexto, este trabalho objetivou a investigação da existência de influências socioeconômicas sobre o desempenho no exame de participantes no Norte do Brasil. Para isso, a mineração baseada no processo Crisp-DM dos microdados oficiais do Enem 2021 foi realizada, com utilização do algoritmo de aprendizado de máquina k-means e aplicação de análise descritiva. Como resultado, obteve-se uma clusterização com coeficiente de silhueta avaliado em 0,427 e a documentação dos padrões identificados durante o pós-processamento através da utilização de técnicas de estatística descritiva e de visualização dos dados modelados. Os objetivos propostos foram alcançados e os resultados representam descoberta de conhecimento relevante, atual e de grande valor social.

Palavras-chave: Mineração de Dados. Inteligência Artificial. Aprendizado Não Supervisionado. K-means. Enem.

Abstract: The high school national exam Enem (from Portuguese: Exame Nacional do Ensino Médio) is the main instrument of admission to higher education in Brazil and plays a fundamental role in democratization and fairer distribution of opportunities, it is often considered a real symbol of meritocracy by Brazilian citizens. In this context, this research aimed to investigate the existence of socioeconomic influences on the exam performance of participants from northern Brazil. To do that, mining based on the Crisp-DM process of the official Enem 2021 data has been performed, using the k-means machine learning algorithm and applying descriptive analysis. As a result, a clustering has been obtained with a silhouette coefficient evaluated at 0.427 and the documentation of patterns identified during the post-processing through the use of descriptive statistics and modeled data visualization techniques. The proposed objectives were achieved and the results represent the discovery of relevant, current knowledge of great social value.

Keywords: Data Mining. Artificial Intelligence. Unsupervised Learning. K-means. Enem.

1 Graduado em Ciência da Computação (UFT). Analista de Tecnologia no SERPRO.

2 Docente da Universidade Federal do Tocantins (UFT) - Ciência da Computação. Doutora em Ciências Mecânicas (UnB).

INTRODUÇÃO

O Exame Nacional do Ensino Médio (Enem) é utilizado como mecanismo de admissão para o ensino superior por milhares de brasileiros desde sua reformulação em 2009. Seu aproveitamento como prova que complementaria ou substituiria os tradicionais

https://doi.org/10.35685/revintera.v7i1.4077

vestibulares surge, entre outros motivos, a partir da expectativa de padronizar os métodos de ingresso ao ensino superior, representando uma maior democratização do acesso às instituições para qualquer cidadão que estivesse apto a candidatar-se.

Ao se inscrever para realização das provas, os candidatos respondem um questionário, de caráter pessoal elaborado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), mecanismo utilizado para conhecer melhor os perfis dos participantes. Ao fim de cada edição e com as avaliações concluídas, as respostas do questionário são associadas aos dados de desempenho de cada inscrito e, então, disponibilizadas oficialmente pelo Inep (Inep, 2022). Com isso, apesar do exame não objetivar diagnóstico socioeconômico nacional, o grande conjunto de dados do Enem, que associa perfis sociais do país inteiro a seus desempenhos, viabiliza análises técnicas que podem proporcionar conhecimento imparcial e valioso sobre a situação socioeconômica do Brasil e sua distribuição de oportunidades.

Dessa forma, com a motivação de examinar a incidência da equidade social brasileira sobre a educação na região Norte do país, este trabalho propõe a extração de conhecimento a partir da aplicação de técnicas de aprendizado não supervisionado para mineração dos milhares de dados referentes aos participantes no Norte brasileiro no Enem 2021. O objetivo central do presente trabalho é demonstrar a existência de correlações frequentes entre características socioeconômicas e performances de participantes na edição do Enem 2021 no Norte do Brasil, produzindo assim, conhecimento de valor social, o qual pode ser relevante para instituições e áreas associadas ao Enem e à educação em geral uma vez que podem ser utilizados como base para elaborar estratégias de políticas educacionais públicas e colaborar com o aperfeiçoamento daquelas já existentes. A edição de 2021 foi escolhida devido à sua peculiaridade de ter sido realizada em meio à pandemia do coronavírus.

Para isso, a metodologia de Processo Padrão Inter-Indústrias para Mineração de Dados Crisp-DM (*Cross Industry Standard Process for Data Mining*) é aplicada como base para garantir estruturação e consistência em todo o fluxo do projeto (Provost; Fawcett, 2013), no qual o algoritmo de clusterização k-means é utilizado e avaliado para explorar padrões e associações não explícitas entre os perfis socioeconômicos dos participantes e suas performances na edição do ano de 2021 do exame.

Trabalhos relacionados

https://doi.org/10.35685/revintera.v7i1.4077

Desenvolvido por Silva et al. (2020), esse trabalho utilizou técnicas de clusterização, regras de associação e estatística descritiva analisando as relações entre variáveis socioeconômicas e o desempenho de concluintes do ensino médio do estado de Minas Gerias no Enem de 2019, totalizando 88.659 registros. Foi utilizado o algoritmo k-means para obtendo um Cluster A com 46.419 registros e um Cluster B com 40.240. Através do algoritmo Apriori, descobriu-se que fatores socioeconômicos como tipo de escola, cor e escolaridade dos pais influenciam o desempenho dos alunos. O estudo destacou a necessidade de discutir e abordar as desigualdades sociais evidenciadas.

Já Silva (2021) apresenta o uso de técnicas de aprendizado de máquina para identificar desigualdades sociais com base nos dados do ENEM de 2019. Utilizando o algoritmo de árvore de decisão para classificação dos registros em alto e baixo desempenho no exame, k-means para clusterização a partir das notas dos participantes nos cinco eixos avaliados e mineração de regras de associação para identificação de padrões e ocorrências frequentes entre os dados e seus clusters, o estudo ampliou as análises realizadas pelo trabalho Silva et al. (2020) e, ao final, considerou que os resultados socioeconômicos para os registros nacionais se comportaram de maneira semelhante aos registros de Minas Gerais daquele mesmo ano, validando a significância do primeiro trabalho.

Utilizando uma metodologia baseada em mineração de dados educacionais - MDE, Banni, Oliveira e Bernardini (2021) realizaram uma análise dos resultados dos participantes do Enem 2018. Para isso, foram utilizadas técnicas de visualização de dados e construção de modelos preditivos, assumindo que quanto melhor fosse a qualidade da predição do desempenho dos participantes, maior seria o indicativo de que os atributos de entrada teriam influência no rendimento dos mesmos. O trabalho aplicou os algoritmos de Decision Tree, Random Forest, Logistic Regression, Naive Bayes e K-Nearest Neighbors (KNN). Para verificação de qualidade, foram utilizadas acurácia, precisão, recall e F1-Score como métricas de avaliação. Ainda, o conjunto de dados foi separado aleatoriamente em dois subconjuntos disjuntos, sendo 70% para treino dos modelos preditivos e 30% para testes. Como resultado, todos os algoritmos utilizados, com exceção do Naive Bayes, alcançaram acurácia e precisão superiores a 70% na predição de desempenho dos participantes, destacando-se o algoritmo Logistic Regression que superou 80% de precisão. Além disso, foi apontado que atributos socioeconômicos apresentaram relação significativa com o

resultado dos participantes no Enem de 2018. Alguns dos principais atributos destacados foram a renda familiar per capita, raça/cor, nível de escolaridade dos pais e o estado de habitação do participante.

Os estudos mencionados anteriormente contribuíram para a pesquisa atual, oferecendo uma compreensão valiosa sobre as metodologias e técnicas de análise de dados que foram empregadas. Eles destacaram a relevância dos fatores socioeconômicos no desempenho educacional, fornecendo uma base sólida que orientou e fundamentou a abordagem metodológica adotada neste trabalho.

METODOLOGIA

As etapas de desenvolvimento desse trabalho foram estruturadas considerando o modelo de processo Crisp-DM. Assim como ocorre em várias áreas de estudo, para mineração de dados também foram propostas várias metodologias de fluxo que guiam e estruturam projetos, possibilitando a obtenção de melhores resultados a partir de uma sequência consistente de etapas. Uma dessas metodologias é o Crisp-DM, que apoia boas práticas ao definir uma rota completa de etapas a fim de obter melhor entendimento e condução de tais projetos, sendo idealizado em 1996 por quatro líderes de empresas pioneiras na área: Daimler-Benz, Integral Solutions Ltd, NCR e OHRA (Shearer, 2000). Tal metodologia é composta pelas seguintes etapas:

Entendimento do negócio

Nessa fase foi necessário entender a área do projeto e quais eram os seus objetivos, seja por uma perspectiva de negócio ou instituição, possibilitando uma definição mais inteligível do problema a ser abordado. Para tanto, leitura de artigos e pesquisas sobre o ENEM foram realizadas. Com essas informações, foi possível estabelecer os objetivos a serem alcançados, escolher o modelo base de fluxo a ser seguido e, consequentemente, definir quais eram o dados necessários para a mineração.

Entendimento dos dados

Os dados adquiridos foram digitais e estruturados obtidos da fonte do INEP (Inep, 2022). Tais dados foram compostos por 76 atributos e 3.389.832 registros. De acordo com o dicionário de dados disponibilizado pelo INEP (*Dicionário_Microdados_Enem 2021.xlsx*), os atributos foram classificados em seis categorias: Dados dos participantes; Dados da escola participante; Dados do local de aplicação da prova; Dados da prova objetiva; Dados da redação e Dados do questionário socioeconômico.

https://doi.org/10.35685/revintera.v7i1.4077

Na segunda parte do entendimento de dados, leituras e análises exploratórias do conjunto de dados referentes à região norte do Brasil foram realizadas, sendo os atributos, seus valores e frequência de ocorrência no conjunto de dados observados. Para isso, a linguagem de programação Python e a biblioteca de análise e gerenciamento de dados Pandas foram utilizadas.

Preparação dos dados

Essa etapa trata do pré-processamento, ou seja, a preparação do conjunto para transformá-lo no que, de fato, será aplicado ao algoritmo selecionado. Para essa atividade, a linguagem de programação Python e as bibliotecas Pandas e NumPy (Harris et al., 2020) foram utilizadas. Assim, após realizar a limpeza e tratamento dos dados foram gerados os atributos para o processamento do k-means. Os mesmos são apresentados na Tabela 1, com suas respectivas descrições e tipos.

Tabela 1: Atributos considerados para processamento do k-means

Atributo original	Pós-preparação	Descrição	Tipo
NU_NOTA_CN	NOTA_TRANS	Média aritmética reescalada dos participantes	Número decimal, intervalo [0,1]
NU_NOTA_CH			
NU_NOTA_LC			
NU_NOTA_MT			
NU_NOTA_REDACAO			
Q001	MAX_ESC	Máximo nível de escolaridade entre os responsáveis do participante	
Q002			
Q003	MAX_OCUP	Máximo grupo de ocupação profissional entre os responsáveis do participante	
Q004			
Q005	RENDA_PERCAP_TRANS	Renda familiar per capita reescalada do participante	
Q006			
Q007	EMPREG_DOM	Residência com serviço de empregado(a) doméstico(a)	Número binário, 0 ou 1
Q009	QUARTO_DORM	Residência com quarto(s) para dormir	
Q019	TV_CORES	Residência com televisão(ões) em cores	
Q022	CELULAR	Residência com telefone(s) celular(es)	
Q024	COMPUTADOR	Residência com computador(es)	
Q025	INTERNET	Residência com acesso à internet	

Fonte: Elaborado pelo autor(a) (2024)

https://doi.org/10.35685/revintera.v7i1.4077

Modelagem

Para a etapa de modelagem, a biblioteca scikit-learn para linguagem de programação Python foi selecionada, uma vez que conta com inúmeras ferramentas para projetos de aprendizado de máquina. Com isso, a classe KMeans do módulo sklearn.cluster foi importada. Dessa classe, foi criado um objeto definido com os seguintes parâmetros:

•n_clusters com intervalo [2, 11] - Esse parâmetro define a quantidade de clusters em que os dados do conjunto estudado serão agrupados;

•init = ‘k-means++’ - Esse parâmetro define o método de inicialização dos centroides de cada cluster. A inicialização do tipo ‘k-means++’ seleciona os centroides iniciais a partir da utilização de amostragem que se baseia em uma distribuição de probabilidade empírica da contribuição de cada registro para a inércia geral;

• n_init = 10 - Quantidade de vezes que o k-means será executado com diferentes centroides iniciais;

• max_iter = 2000 - Esse parâmetro define a quantidade máxima de iterações do k-means em uma única execução, caso a convergência para um agrupamento estável não seja alcançada.

Do objeto criado, foi executado o método fit_predict, para computar a clusterização k-means dos 221.400 registros com os 10 atributos explicitados na coluna Pós-preparação da figura 2 e levando em consideração os parâmetros descritos anteriormente.

Ao fim de cada iteração do laço de repetição e com a clusterização modelada, foi armazenada na variável labels a lista retornada do método fit_predict com os clusters nomeados de 0 a n_clusters - 1, sendo o enésimo elemento dessa lista o cluster atribuído ao enésimo registro do conjunto de dados enviado para a modelagem.

Avaliação

Da mesma biblioteca, também foi importada a função silhouette_score, do módulo sklearn.metrics, que conta com funções de avaliação, métricas de desempenho, cálculos de distância, entre outros.

Com a lista labels resultante de cada clusterização, foi calculado o coeficiente de silhueta médio do agrupamento. Assim, para cada clusterização com n_clusters variando de 2 a 11, foi armazenado o coeficiente de silhueta médio das modelagens a fim de se definir o melhor número de clusters para o conjunto de dados estudado. Com n_clusters = 4, o

coeficiente de silhueta atingiu seu melhor valor, alcançando 0,427 e, por isso, o número ideal de clusters foi fixado em n_clusters = 4.

Ainda que difiram na edição do Enem estudada, nos dados, no tratamento e processamento dos dados, na quantidade de registros e atributos considerados para modelagem do k-means, o que dificulta a comparação de métricas, o valor do coeficiente de silhueta alcançado nesse estudo é bem próximo aos valores encontrados nos estudos de Silva et al. (2020) e Silva (2021), quase alcançando o coeficiente do primeiro e superando o coeficiente do segundo.

Por fim, com a modelagem que agrupou os dados em 4 clusters distintos, foram realizados o estudo das características dos elementos de cada cluster, a comparação entre eles, a análise de atributos não aplicados à modelagem e, ainda, a verificação de existência e identificação de padrões entre os perfis socioeconômicos dos participantes e suas performances. Os resultados dessa etapa são expostos na seção de resultados e discussões.

Implantação

Seguindo o ciclo de desenvolvimento proposto pelo modelo de processo Crisp-DM a etapa de implantação para esse estudo trata-se do detalhamento, exposição, organização e estruturação de todas as etapas, estratégias e ferramentas utilizadas no projeto, além da documentação dos resultados, constatações e conclusões obtidos a fim de revelar o conhecimento adquirido através da mineração. Para isso, sumarização das informações, estatística descritiva e técnicas de visualização de dados também foram utilizadas visando a facilitação da exposição dos resultados e a compreensão das descobertas obtidas.

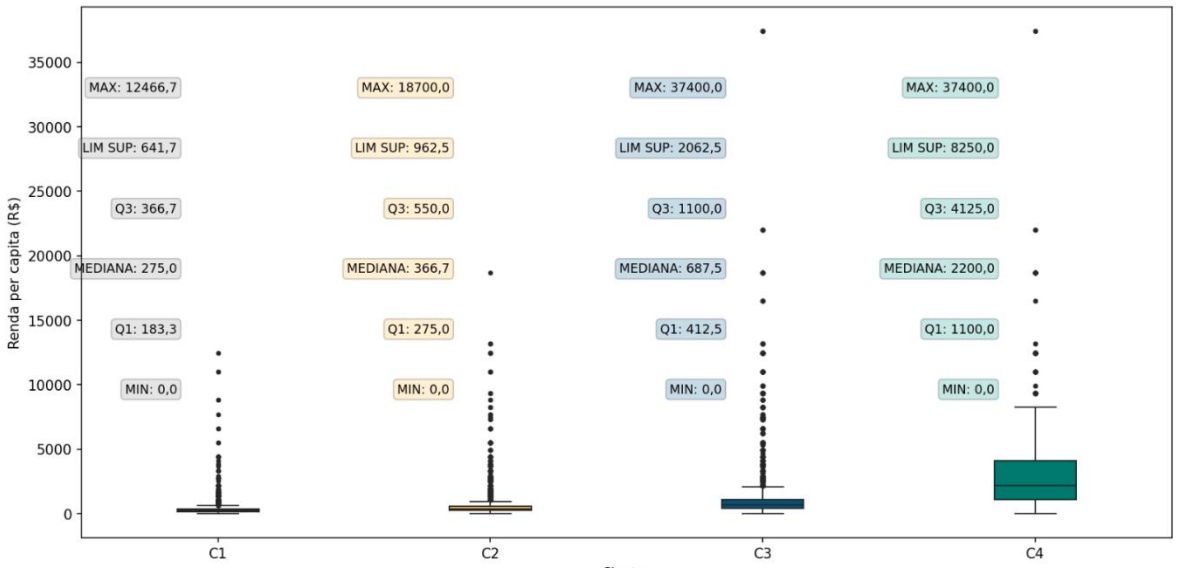
RESULTADOS E DISCUSSÕES

A primeira análise refere-se à renda familiar per capita dos participantes, da região Norte, do Enem 2021. A figura 1 exibe a renda familiar per capita dos participantes em relação a cada cluster formado. O gráfico de diagrama de caixa representa a distribuição dos dados e, devido a concentração na parte inferior do eixo vertical, as métricas principais estão em destaque à esquerda de cada diagrama.

É possível notar que, com exceção do valor mínimo observado de renda familiar per capita em cada grupo, que é igual a 0 para todos os clusters, todas as outras métricas crescem ao partir do cluster C1 em direção ao C4. Além disso, pelos valores de quartis do C3, 75% dos integrantes desse cluster possuem renda familiar per capita mais alta que 75% dos integrantes do C1 e 50% possuem renda familiar per capita mais alta que 75% dos

integrantes do C2. Já em relação ao C4, 75% dos integrantes possuem renda familiar per capita mais alta que 3 vezes a renda familiar per capita, que 2 vezes a renda familiar per capita e que 1 vez a renda familiar per capita de aproximadamente 75% dos integrantes de C1, C2 e C3, respectivamente. Não obstante, é possível ainda constatar que, para o cluster C4, 50% das rendas per capita estão concentradas entre o intervalo de R$1.100,00 e R$4.125,00, enquanto nos clusters C1, C2 e C3, apenas pontos dispersos apresentam renda familiar per capita acima de R$641,70, R$962,50 e R$2.062,50, respectivamente.

Figura 1: Renda familiar per capita dos participantes em relação aos clusters



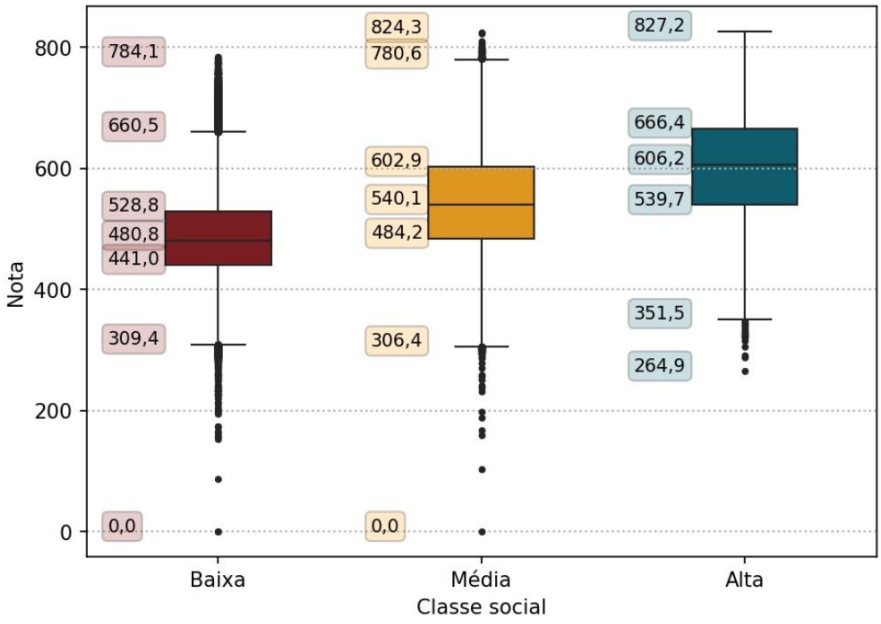
Fonte: Elaborado pelo autor(a) (2024).

Para aprofundar essa análise, a figura 2 apresenta as notas médias dos participantes em relação à categoria de classe social. É possível observar diferenças significativas entre as principais medidas dos diagramas de caixa das notas médias ao comparar diferentes classes, sendo o primeiro quartil de notas da classe média superior à mediana de notas da classe baixa, ou seja, 75% dos participantes da classe média têm nota mais alta que 50% dos participantes da classe baixa. Entre as classes média e alta, o cenário é semelhante: o primeiro quartil de notas da classe alta é bem próximo da mediana de notas da classe média e a mediana daquele grupo é superior ao terceiro quartil de notas desse, ou seja, 50% dos participantes da classe alta têm nota mais alta que 75% dos participantes da classe média.

Ainda, é possível reparar que a mediana de notas médias dos participantes de classe baixa é 59,3 pontos mais baixa que a mediana dos participantes de classe média e 125,4 pontos

mais baixa que a mediana dos participantes de classe alta. O grupo de participantes da classe média, por sua vez, tem mediana com 66,1 pontos abaixo da mediana do grupo de classe alta. Por fim, o grupo de classe baixa tem limite superior 120,1 e 166,7 pontos mais baixo que o limite superior do grupo de classe média e alta, respectivamente. Tais informações evidenciam um padrão no qual a classe social a partir da renda familiar per capita pode impactar no desempenho dos participantes.

Figura 2: Notas médias agrupadas por categoria de classe social

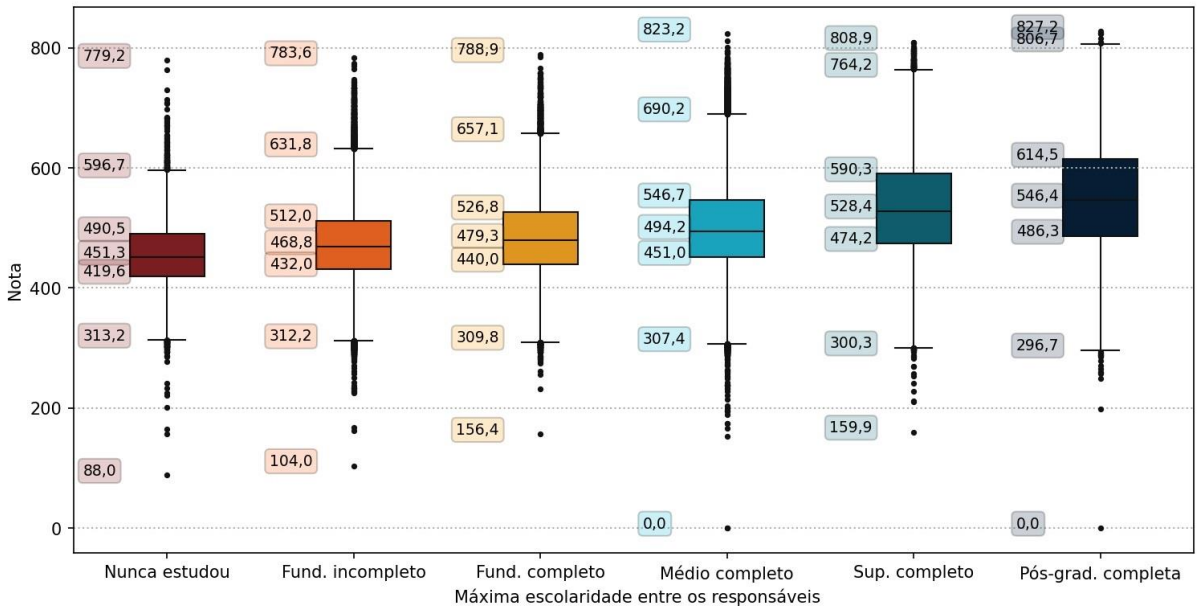


Fonte: Elaborado pelo autor(a) (2024)

Quando consideramos as notas médias dos participantes agrupadas por nível de escolaridade e ocupação profissional dos responsáveis, figura 3 (a) e (b), é possível verificar diferenças significativas das notas médias nas principais medidas dos diagramas. Todas as três medidas que formam o intervalo interquartil (primeiro quartil, mediana e terceiro quartil) e limite superior aumentam com o aumento da escolaridade máxima dos responsáveis, assim como aumentam também ao partir da ocupação de responsáveis no Grupo 1 em direção ao Grupo 5. Ainda, o grupo de participantes com responsável/responsáveis que nunca estudou/estudaram tem mediana de notas com 95,1 pontos abaixo do grupo de participantes com, pelo menos, um responsável com pós-graduação completa. De maneira semelhante, o grupo de participantes com responsável/responsáveis com ocupação

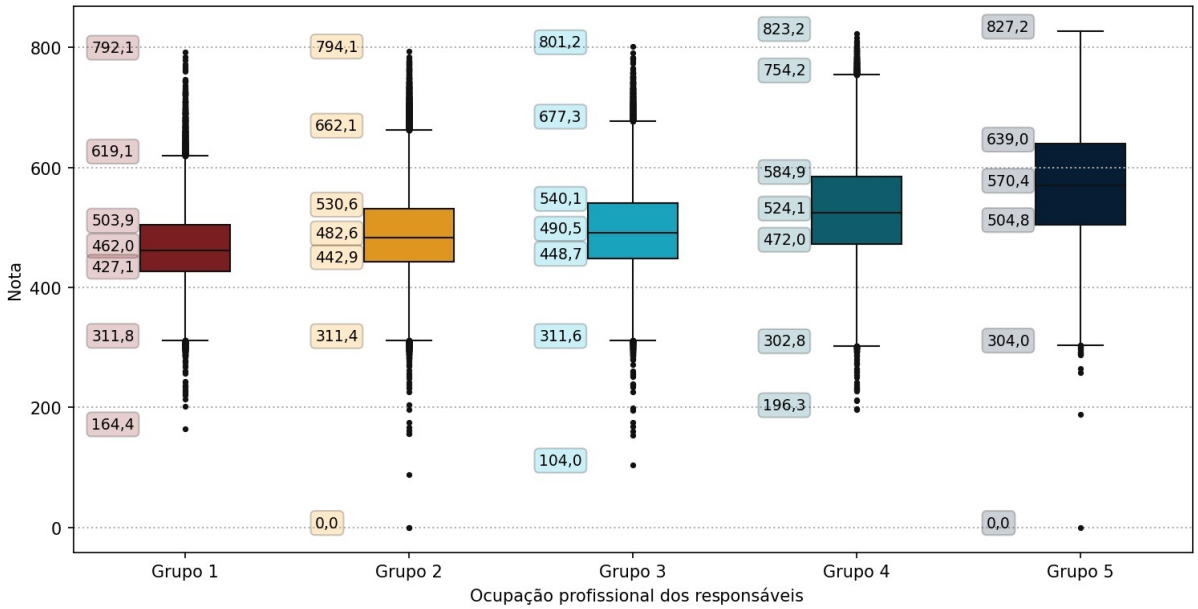
profissional do Grupo 1 tem mediana de notas com 108,4 pontos abaixo da mediana do grupo de participantes com, pelo menos, um responsável com ocupação profissional do Grupo 5. Tal figura revela tendências nas quais a escolaridade e ocupações profissionais dos responsáveis podem gerar discrepâncias no desempenho dos participantes.

Figura 3a: Notas médias agrupadas por escolaridade e ocupação profissional dos responsáveis - Por escolaridade



Fonte: Elaborado pelo autor(a) (2024)

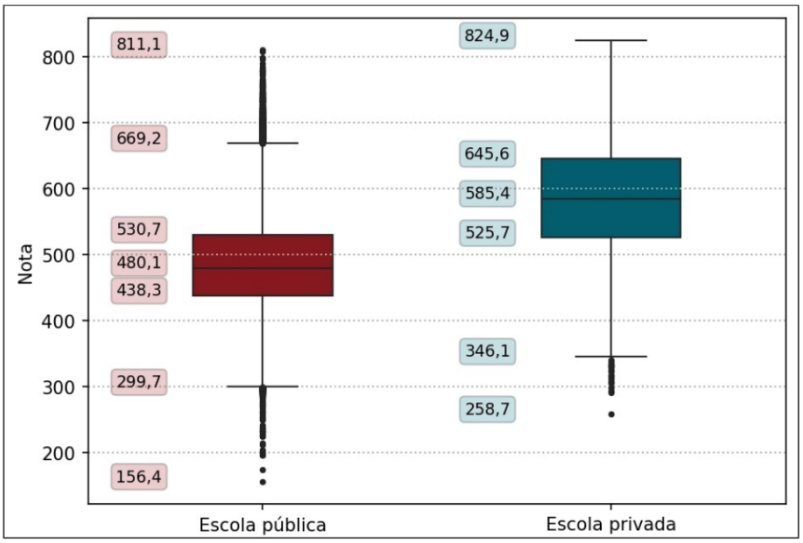
Figura 3b: Notas médias agrupadas por escolaridade e ocupação profissional dos responsáveis - Por ocupação



Fonte: Elaborado pelo autor(a) (2024)

Quanto ao tipo de escola frequentada no ensino médio, a figura 4 exibe as notas médias dos participantes. É possível observar diferenças significativas de notas médias, ficando o grupo que frequentou escola pública com as medidas principais do diagrama de caixa consideravelmente mais baixas que o grupo que frequentou escola privada: o primeiro quartil das médias de notas desse é bem próximo do terceiro quartil das médias de notas daquele, assim, 75% dos participantes que responderam ter frequentado escola pública durante o ensino médio têm nota inferior a 530,7, enquanto 75% dos participantes que responderam ter frequentado escola privada durante o ensino médio têm nota superior a 525,7. Além disso, as medianas de notas dos dois grupos distanciam-se em mais de 105 pontos de diferença e o limite superior de notas em mais de 155 pontos.

Figura 4: Notas médias agrupadas por tipo de escola

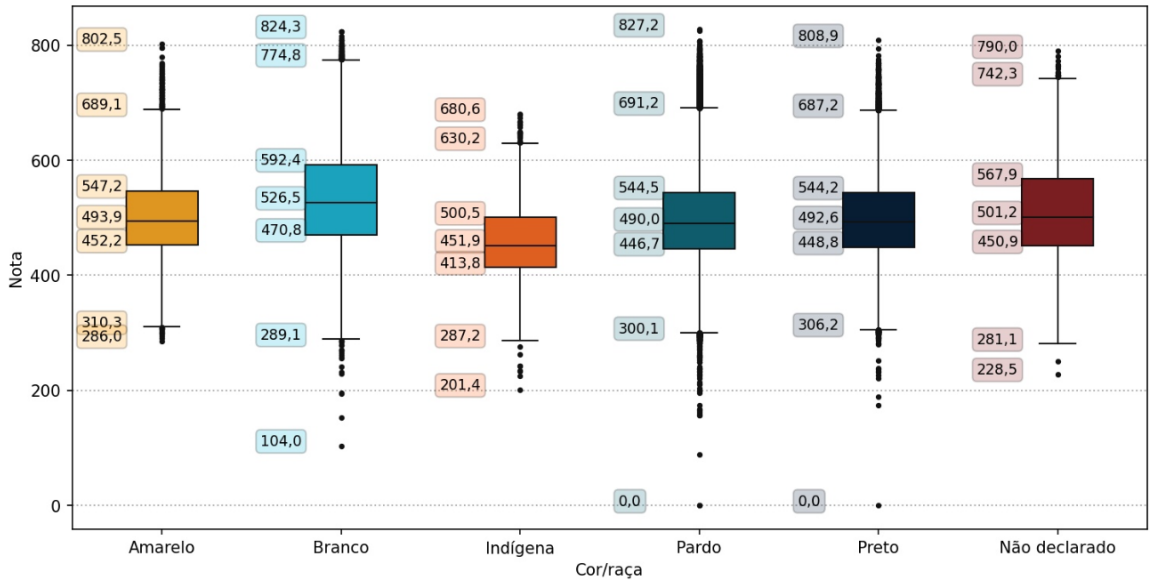


Fonte: Elaborado pelo autor(a) (2024)

É possível notar diferenças no padrão de desempenho, também, ao analisar as notas médias dos participantes de cada cluster em relação à cor/raça autodeclarada. A figura 5 apresenta esses dados e, nela, o padrão de disparidade que se destaca é o do grupo com integrantes autodeclarados indígenas ao compará-lo com os outros agrupamentos, um vez que o mesmo possui os valores de primeiro quartil, mediana, terceiro quartil e limite superior consideravelmente mais baixos que todos os outros grupos.

https://doi.org/10.35685/revintera.v7i1.4077

Figura 5: Notas médias agrupadas por raça/cor declarada



Fonte: Elaborado pelo autor(a) (2024)

Ao comparar os resultados deste projeto com os trabalhos relacionados, podemos notar algumas semelhanças e diferenças importantes. Assim como nos estudos de Silva et al. (2020) e Banni, Oliveira e Bernardini (2021), este trabalho confirmou que fatores como renda familiar e escolaridade dos pais influenciam significativamente o desempenho no Enem, reforçando a validade dessas conclusões em diferentes contextos.

Embora todos os estudos tenham usado técnicas de mineração de dados, este se destacou por seguir o modelo Crisp-DM, estruturando cada etapa da pesquisa. Além disso, focou na edição de 2021 do Enem, analisando especificamente os participantes da região Norte, enquanto estudos anteriores abordaram edições e regiões distintas, como Minas Gerais.

Em resumo, este estudo não apenas reafirma descobertas anteriores, mas também oferece novas perspectivas ao considerar o contexto único da edição de 2021 do Enem, especialmente sob a influência da pandemia.

CONCLUSÃO

Com base nos resultados apresentados é possível concluir que o objetivo central do presente trabalho, investigar a existência de correlações entre características socioeconômicas e o desempenho dos participantes do Enem 2021 na região Norte do Brasil, foi alcançado. Através da aplicação do

algoritmo k-means e do modelo Crisp-DM, foi possível identificar padrões significativos que confirmam a influência de fatores como renda familiar e escolaridade dos pais no desempenho dos estudantes.

Desse estudo, foram observadas contribuições na área tecnológica e social. Para o âmbito tecnológico, há a apresentação de uma metodologia de tratamento dos microdados oficiais do Enem disponibilizados pelo Inep e resultados alcançados a partir da execução do algoritmo de clusterização k-means em um conjunto de dados educacionais com 221.400 registros e 17 atributos considerados. Para o âmbito social, a principal contribuição constitui-se na apresentação de perfis de participantes no Norte brasileiro do Enem 2021 mais correlacionados a desempenhos menos satisfatórios, com base nos dados estudados, possibilitando a criação de programas sociais, estratégias, direcionamento de vagas, campanhas, etc, para tal público, a fim de contribuir com a equidade e distribuição mais justa de oportunidades para cidadãos brasileiros.

Ademais, para as pretensões profissionais e acadêmicas, este trabalho representa um passo significativo. Profissionalmente, ele fornece uma base sólida de análise de dados e compreensão das dinâmicas sociais que podem ser aplicadas em diversas áreas. Academicamente, o estudo enriquece o debate sobre a educação no Brasil, oferecendo novas perspectivas e metodologias que podem ser exploradas em pesquisas futuras. Assim, este projeto não apenas atinge seus objetivos, mas também abre portas para novas oportunidades de investigação e atuação no campo educacional.

Trabalhos futuros relacionados podem tratar do estudo de registros de participantes com desempenhos discrepantes dos demais membros de seus clusters, seja registros com notas muito abaixo ou muito acima do intervalo de notas mais prováveis/frequentes para o grupo observado, com o objetivo de investigar possíveis correlações ou mesmo causas para tais afastamentos. Além disso, pesquisas explicativas também podem ser desenvolvidas a partir dos resultados desse projeto, objetivando o aprofundamento das descobertas a partir de estudos que apresentem motivos e explicações para os padrões e tendências encontrados.

REFERÊNCIAS

BANNI, M.; OLIVEIRA, M.; BERNARDINI, F. Uma Análise Experimental Usando Mineração de Dados Educacionais sobre os Dados do ENEM para Identificação de Causas do Desempenho dos Estudantes. Anais do II Workshop sobre as Implicações da Computação na Sociedade. Porto Alegre, RS, Brasil: SBC. p. 57–66. ISSN 2763- 8707, 2021. Disponível em:

<https://sol.sbc.org.br/index.php/wics/article/view/15964>. Acesso em: 12 de fev. 2024.

HARRIS, C. R. et al. Array programming with NumPy. Nature, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>. Acesso em: 12 de fev. 2024.

INEP. Enem. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP, 2022. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>. Acesso em: 01 jun. 2022.

INEP. Microdados do Enem 2021. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP, 2022. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/ enem>. Acesso em: 18 ago. 2022

PROVOST, F.; FAWCETT, T. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. 1st. ed. Sebastopol, CA: O'Reilly Media, 2013.

SHEARER, C. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, The Data Warehousing Institute, v. 5, n. 4, p. 13–22, 2000.

SILVA, V. et al. Identificação de Desigualdades sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados. Anais do XXXI Simpósio Brasileiro de Informática na Educação. Porto Alegre, RS, Brasil: SBC. p. 72–81. ISSN 0000-0000, 2020. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/12763>. Acesso em: 18 ago. 2022.

SILVA, V. A. A. da. Uso de aprendizado de máquina para identificar desigualdades sociais na base de dados do ENEM. Universidade Federal de Juiz de Fora, 2021. Disponível em: <http://monografias.ice.ufjf.br/tcc-web/tcc?id=549>. Acesso em: 18 ago. 2022.