CONIC SEMESP

25º Congresso Nacional de Iniciação Científica

TÍTULO: ANÁLISE EXPLORATÓRIA DE DADOS EDUCACIONAIS: UM ESTUDO DE CASO DO PROGRAMA SABARÁ

CATEGORIA: CONCLUÍDO

ÁREA: CIÊNCIAS EXATAS, DA TERRA E AGRÁRIAS

SUBÁREA: Computação e Informática

INSTITUIÇÃO: INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS - IFMG

AUTOR(ES): THIAGO PEDRO FERREIRA DE MORAES

ORIENTADOR(ES): RENATO MIRANDA FILHO, CARLOS ALEXANDRE SILVA

Realização: SEMESP

IES parceiras: PUC CAMPINAS

CATEGORIA CONCLUÍDO

1. RESUMO

Este artigo apresenta uma análise exploratória de dados da edição 2023 do Programa Sabará, iniciativa de extensão do IFMG – Campus Sabará que promove a inclusão digital por meio de oficinas de programação e robótica para alunos do ensino fundamental II. Foram analisadas respostas de 283 estudantes utilizando estatística descritiva, relatórios automáticos e *k-means* ($k = 3$). Identificaram-se três perfis: o **Cluster 0** ($n = 184$) apresentou melhor desempenho (*média 15,4/20*); o **Cluster 1** concentrou majoritariamente estudantes do 6º e 7º anos; e o **Cluster 2**, com menor representatividade, reuniu alunos de uma escola distante do IFMG. Observou-se correlação fraca entre pontuação, tipo de escola e distância até a instituição. Os resultados oferecem subsídios pedagógicos e logísticos para futuras edições do Programa e evidenciam o potencial da ciência de dados aplicada à educação.

2. INTRODUÇÃO

Sabará é um dos municípios mais antigos de Minas Gerais e, segundo o IBGE, registrou 129.380 habitantes no Censo de 2022 e 134.576 habitantes na estimativa de 2025. Apesar do IDHM classificado como “alto” (0,731; 2010), o município ainda apresenta desafios socioeconômicos, refletidos, por exemplo, no PIB per capita de R$ 30.474,48 (2021). Nesse contexto, o IFMG – Campus Sabará desenvolveu o Programa Sabará, ação de extensão voltada à inclusão digital e à educação tecnológica para mitigar desigualdades educacionais por meio do ensino de programação e robótica.

Em 2023, uma nova edição do Programa Sabará foi implementada com fomento da Secretaria de Educação Profissional e Tecnológica do Ministério da Educação (SETEC/MEC). A iniciativa ofereceu oficinas de programação, pensamento computacional e robótica para estudantes do ensino fundamental de escolas públicas e privadas de Sabará e da Região Metropolitana de Belo Horizonte. Nesta edição, o projeto atendeu aproximadamente 450 alunos, configurando-se como uma das mais abrangentes iniciativas de educação tecnológica já realizadas no município.

As atividades desenvolvidas no projeto utilizaram ferramentas tecnológicas acessíveis, como a linguagem Logo, a plataforma de programação em blocos Scratch e kits de robótica LEGO e Arduino. O objetivo principal foi proporcionar experiências práticas que estimulem o pensamento lógico, a criatividade e a capacidade de resolução de problemas. O público atendido foi formado por estudantes do 6º ao 9º ano do ensino fundamental, abrangendo diferentes realidades socioeconômicas. Além de promover o desenvolvimento educacional, a iniciativa buscou favorecer a integração social nas áreas de ciência, tecnologia e engenharia, com potencial para gerar impactos positivos no contexto econômico e social do município.

Esta pesquisa tem como objetivo analisar o perfil e o desempenho dos estudantes participantes da edição de 2023, identificando padrões, relações e características relevantes por meio de uma análise exploratória multivariada de dados educacionais. O estudo fundamenta-se em abordagens da Ciência de Dados Educacionais, que busca compreender processos de aprendizagem por meio da exploração de dados (SILVA et al., 2017). Pesquisas anteriores já aplicaram técnicas semelhantes: Cesaretti et al. (2021) combinaram robótica e *k-means* para analisar estratégias de resolução de atividades, enquanto Bogdandy et al. (2020) investigaram os impactos das transformações digitais na educação durante a pandemia de COVID-19. Além disso, os resultados obtidos podem subsidiar o planejamento pedagógico, apoiar a criação de estratégias personalizadas para os diferentes perfis de estudantes e contribuir para políticas públicas educacionais no município, em alinhamento com a Base Nacional Comum Curricular (BNCC), que incentiva o uso de ferramentas tecnológicas no ensino básico.

3. OBJETIVOS

O objetivo deste estudo é caracterizar o perfil e o desempenho dos estudantes da edição de 2023 do Programa Sabará, por meio de análise exploratória multivariada combinando estatística descritiva e técnicas de agrupamento para identificar padrões e correlações entre variáveis acadêmicas, demográficas e geográficas. Adicionalmente, busca-se gerar *insights* práticos que apoiem melhorias no planejamento pedagógico e orientem trabalhos correlatos.

4. METODOLOGIA

Esta pesquisa caracteriza-se como exploratória e quantitativa, utilizando *k-means* e a biblioteca *ydata-profiling*, em Python, para Análise Exploratória de Dados (EDA). A abordagem descritiva incluiu, também, o cálculo da correlação de Pearson, com o objetivo de identificar padrões e relações entre variáveis, possibilitando a geração de hipóteses para ações futuras. O algoritmo *k-means* foi escolhido por sua eficiência na identificação de agrupamentos em conjuntos de dados educacionais com múltiplas variáveis, permitindo encontrar padrões ocultos de desempenho e perfil dos estudantes.

Participaram do estudo 283 estudantes do Programa Sabará — edição 2023 — que realizaram a avaliação diagnóstica. Os alunos que não responderam à prova foram excluídos da análise.

Os dados foram processados em Python, incluindo limpeza, padronização e integração de tabelas. Foram normalizadas variáveis, codificados atributos categóricos e removidas *stopwords* dos nomes das escolas. A Tabela 1 detalha as bibliotecas e suas finalidades.

Tabela 1. Principais bibliotecas utilizadas e suas finalidades.

Biblioteca	Descrição
pandas	Permite a manipulação de dados, incluindo consultas, modificações, exclusões e inclusões. Também permite análises.
numpy	Operações matemáticas e computacionais eficientes.
matplotlib.pyplot	Criação de gráficos para visualização dos dados.
seaborn	Visualização estatística avançada com estética aprimorada.
sklearn.cluster.KMeans	Implementação do algoritmo *k-means* para clusterização.
sklearn.preprocessing.StandardScaler	Normalização de variáveis numéricas, padronizando média e desvio padrão.
sklearn.preprocessing.OneHotEncoder	Codificação de variáveis categóricas em formato binário (0 ou 1).
ydata-profiling	Geração de relatórios automáticos para Análise Exploratória de Dados (EDA).
sklearn.metrics.silhouette_score	Avaliação da qualidade dos agrupamentos formados pelo *k-means*.
nltk / nltk.corpus.stopwords	Processamento de linguagem natural; remoção de *stopwords* dos nomes das escolas.

Fonte: Elaborado pelos autores (2025).

5. DESENVOLVIMENTO

A coleta de dados foi feita via *Google Form*s, utilizando formulários de cadastro dos participantes e aplicação da avaliação diagnóstica. As respostas foram armazenadas em planilhas no *Google Sheets*, gerando tabelas como: **Avaliação diagnóstica** (com a pontuação por aluno), **Inscrição** (dados pessoais) e **Documentação** (informações complementares).

O tratamento dos dados foi realizado com o uso das bibliotecas *Pandas* e *NumPy*, incluindo etapas de limpeza, padronização e organização. Foram removidos valores nulos, substituídas entradas *NaN* por *False* e convertidos os anos escolares para valores numéricos. Além disso, foram removidas *stopword*s da coluna Escola e criada a variável mudança de turma, baseada nos sufixos dos nomes dos alunos. Como os registros não possuíam identificadores únicos, os nomes foram normalizados para garantir consistência entre as diferentes tabelas.

Após o pré-processamento, foram derivadas novas variáveis com base nas tabelas originais, incluindo:

- **Idade**: calculada a partir da data de nascimento;
- **Três distâncias geográficas**: entre o local do curso, a escola e a residência de cada aluno;
- **Tipo de escola**: classificada como estadual, municipal ou privada;
- **Requisição de mudança de turma**: extraída a partir das solicitações registradas.

A classificação do tipo de escola foi realizada manualmente por meio de uma tabela criada no *Google Sheets*, na qual cada instituição foi registrada e atribuída à categoria correspondente. O cálculo das distâncias utilizou as APIs do Google (*Geocoding* e *Routes*). Primeiramente, os endereços foram convertidos em coordenadas de latitude e longitude; em seguida, essas coordenadas foram processadas pela *Routes* API, que retornou as distâncias em metros entre os pontos de origem e destino.

A Tabela 2 apresenta a listagem das variáveis utilizadas nas análises. Após a consolidação dos dados, foi aplicada a clusterização utilizando o algoritmo *k-means* com todas as variáveis derivadas, exceto a pontuação.

Tabela 2. Listagem das variáveis usadas nas análises.

Variável	Tipo de dado
ANO	Numérico
DIA_AULA	Alfanumérico
ESCOLA	Alfanumérico
IDADE	Numérico
MUDANCA_DE_TURMA	Numérico
PONTUACAO	Numérico
SEXO	Alfanumérico
TIPO_ESCOLA	Alfanumérico
TURMA	Alfanumérico
DISTANCIA_IFMG_CASA	Numérico
DISTANCIA_ESCOLA_IFMG	Numérico
DISTANCIA_ESCOLA_CASA	Numérico

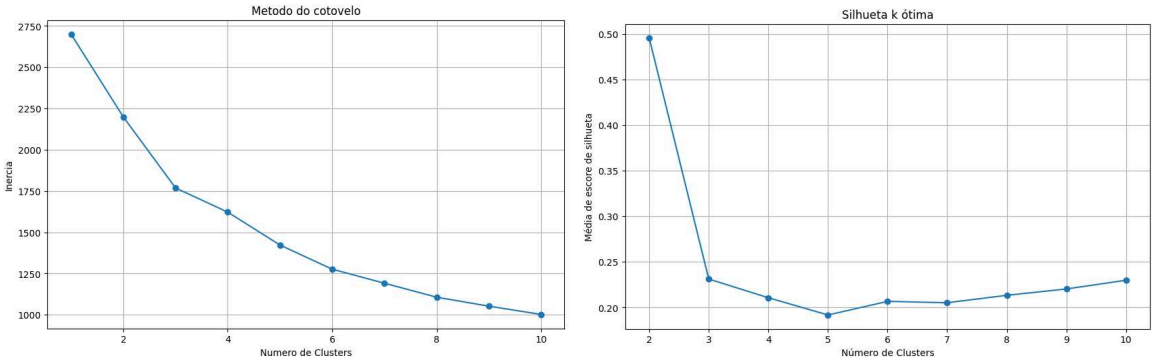
Fonte: Elaborado pelos autores (2025).

6. RESULTADOS

6.1 Determinação do número de clusters

O número ideal de agrupamentos foi definido a partir do método do cotovelo e da métrica de silhueta. A Figura 1 apresenta os dois gráficos: o cotovelo demonstra estabilização da inércia intra-cluster em três grupos (a), enquanto a silhueta sugeriu dois grupos, mas foi inadequada por criar clusters desbalanceados. (b). Assim, optou-se por três clusters para manter melhor a distribuição dos alunos.

Figura 1. Determinação do número ideal de clusters: (a) método do cotovelo; (b) métrica de silhueta.



(a) **(b)**

Fonte: Elaborado pelos autores (2025).

Os três grupos foram determinados: **Cluster 0** (184 alunos), **Cluster 1** (87 alunos) e **Cluster 2** (12 alunos). A Tabela 3 apresenta as médias, desvios padrão e número de alunos por cluster.

Tabela 3. Médias, desvios padrão e número de alunos por cluster.

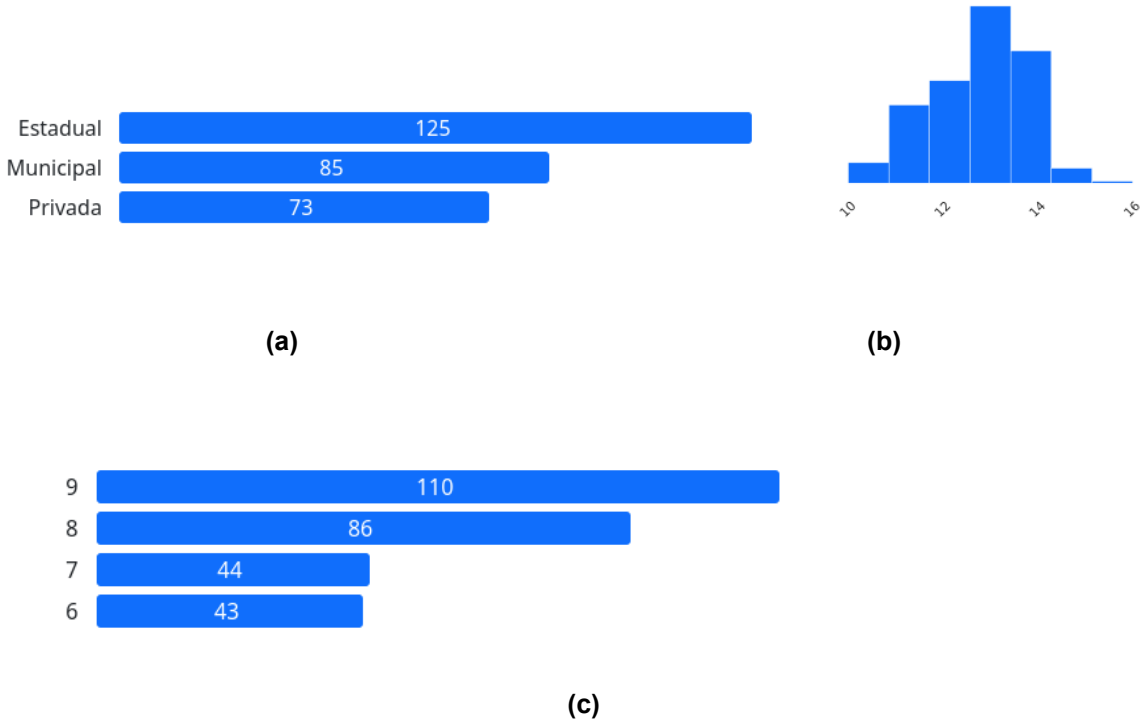
Cluster	Média	Desvio Padrão	Tamanho
0	15,44	3,23	184
1	14,85	3,57	87
2	13,91	2,94	12

Fonte: Elaborado pelos autores (2025).

6.2 Análise do perfil geral

O perfil geral dos alunos é mostrado na Figura 2, que reúne três visualizações: (a) distribuição por tipo de escola, (b) distribuição de idades e (c) distribuição por ano escolar. A maioria dos estudantes frequentam escolas estaduais, tem 13 anos e está matriculada, predominantemente, no 9º ano.

Figura 2. Perfil geral dos alunos: (a) tipo de escola, (b) idade e (c) ano escolar.



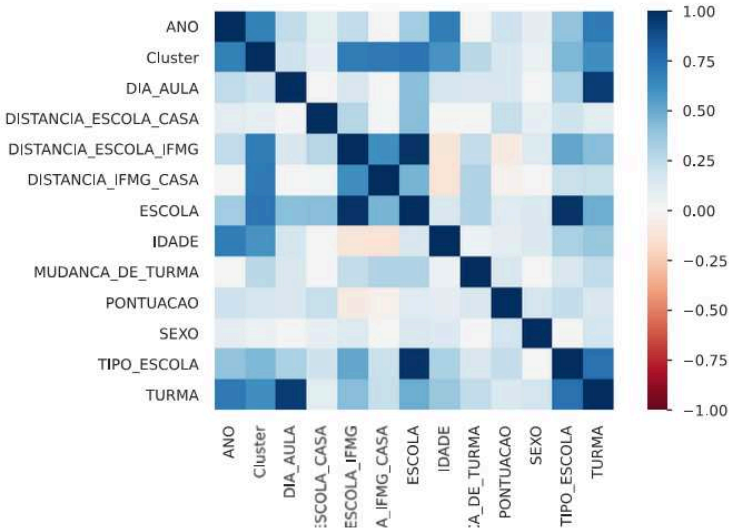
(a) **(b)**

(c)

Fonte: Elaborado pelos autores (2025).

A Figura 3 apresenta o *heatmap* das correlações de *Pearson* para todas as variáveis. De modo geral, não foram observadas correlações fortes entre a pontuação e as demais variáveis. Identificaram-se apenas relações positivas fracas com a distância entre residência e escola (*r = 0,226*) e com o tipo de escola (*r = 0,231*).

Figura 3. *Heatmap* de correlação de Pearson entre as variáveis do conjunto de dados.



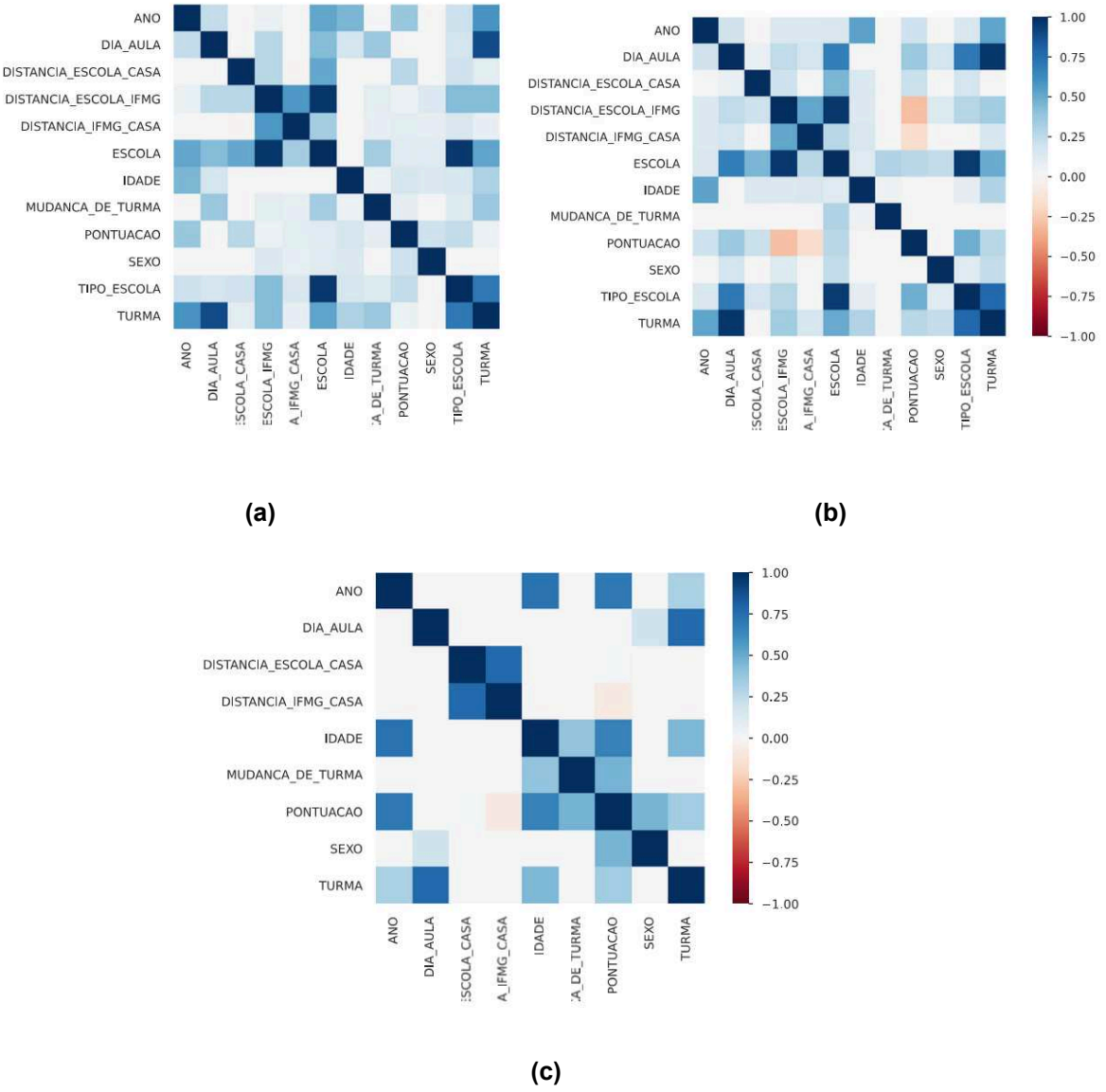
Fonte: Elaborado pelos autores (2025).

6.3 Análise dos clusters

A Figura 4 reúne os *heatmaps* dos três clusters: (a) **Cluster 0**, (b) **Cluster 1** e (c) **Cluster 2**. A seguir, apresentam-se as descrições de cada cluster:

- O **Cluster 0** é composto por 184 alunos, apresenta predominância de estudantes do 8º e 9º anos. Observa-se uma correlação moderada entre pontuação e ano escolar (*r = 0,37*), embora não confirmada no perfil geral.
- O **Cluster 1** é formado por 87 alunos, é caracterizado por 53% de escolas privadas, 40% municipais e 7% estaduais. Aqui, a correlação entre pontuação e tipo de escola é moderada (*r = 0,482*).
- O **Cluster 2**, com 12 alunos de uma única instituição distante mais de 20 km do IFMG, exige interpretação cautelosa devido à baixa representatividade.

Figura 4. *Heatmaps* de correlação de Pearson por cluster: (a) Cluster 0, (b) Cluster 1, (c) Cluster 2.



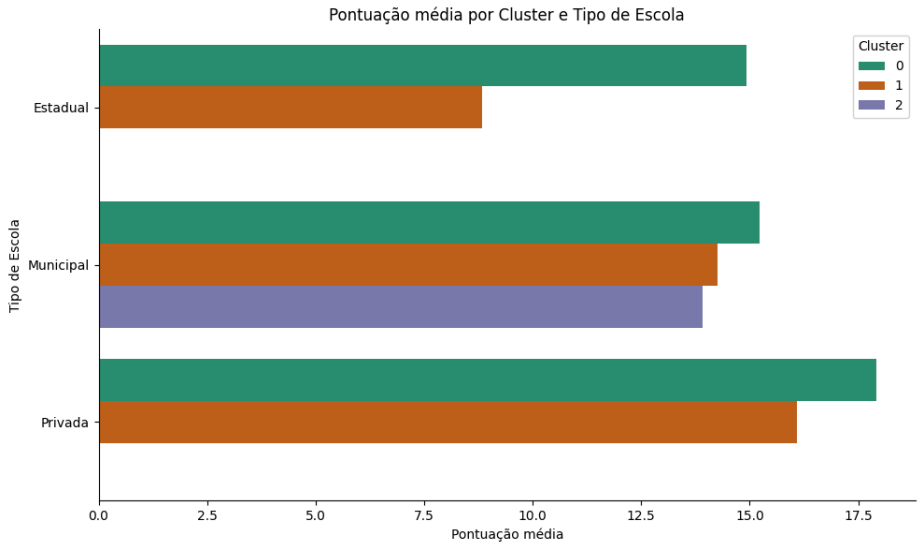
Fonte: Elaborado pelos autores (2025).

Em conjunto, os resultados sugerem que fatores institucionais e geográficos (tipo de escola e distância percorrida) se associam ao desempenho, indicando a pertinência de estratégias pedagógicas personalizadas por perfil (como turmas de nivelamento ou apoio focalizado) e de ajustes logísticos (como horários e deslocamento) para ampliar a participação e o aproveitamento.

6.4 Média de notas por cluster e tipo de escola

A Figura 5 apresenta a pontuação média por tipo de escola e cluster. Observa-se que, em todos os agrupamentos que incluem alunos de escolas privadas, esses estudantes obtiveram melhores resultados.

Figura 5. Pontuação média por tipo de escola. Cada cor representa um cluster.



Fonte: Elaborado pelos autores (2025).

7. CONSIDERAÇÕES FINAIS

Este estudo apresentou uma análise exploratória dos dados da edição 2023 do Programa Sabará, combinando estatística descritiva, visualizações e agrupamento por *k-means*. Identificaram-se três perfis de estudantes e diferenças de desempenho associadas a tipo de escola, ano escolar e distância até o local do curso, oferecendo evidências acionáveis para decisões pedagógicas e logísticas.

Apesar dos resultados relevantes, algumas limitações devem ser consideradas. A análise utilizou uma base de dados restrita, com destaque para o **Cluster 2**, composto por um número reduzido de alunos, o que impede conclusões robustas para esse grupo. Além disso, a coleta de dados foi feita por meio de formulários manuais e digitais, o que pode introduzir inconsistências e lacunas nas informações.

Para pesquisas futuras, recomenda-se: (i) ampliar o número de participantes e instituições para aumentar a robustez estatística; (ii) incorporar variáveis socioeconômicas e de engajamento (ex.: presença, participação em atividades, uso

de plataformas); (iii) avaliar significância estatística das correlações reportadas; e (iv) explorar modelos preditivos (ex.: regressões, árvores de decisão) para estimar desempenho e identificar perfis de risco visando intervenções precoces. Dessa forma, este estudo oferece insights valiosos para o aprimoramento do Programa Sabará e contribui para a formulação de políticas educacionais mais eficazes.

8. FONTES CONSULTADAS

BOGDANDY, B.; TAMAS, J.; TOTH, Z. Digital Transformation in Education during COVID-19: a Case Study. In: 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Mariehamn, Finland, 2020. p. 000173-000178. DOI: 10.1109/CogInfoCom50765.2020.9237840.

CESARETTI, L.; SCREPANTI, L.; SCARADOZZI, D.; MANGINA, E. *Analysis of Educational Robotics Activities Using a Machine Learning Approach*. In: MAKERS AT SCHOOL, EDUCATIONAL ROBOTICS AND INNOVATIVE LEARNING ENVIRONMENTS. Cham: Springer, 2021. p. 203–211.

IBGE. Cidades e Estados: Sabará (MG). Rio de Janeiro: IBGE, 2025. Disponível em: https://www.ibge.gov.br/cidades-e-estados/mg/sabara.html. Acesso em: 3 set. 2025.

SILVA, L. A. et al. Ciência de dados educacionais: definições e convergências entre as áreas de pesquisa. In: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO - WCBIE 2017, 6, Workshops, 2017. Anais... Recife: SBC, 2017.

GOOGLE DEVELOPERS. Disponível em: https://developers.google.com/apis-explorer. Acesso em: 23 ago. 2025.