Pattern identification in academic trajectories within the Brazilian Military School System using machine learning

William Moreira de Pinho[1], **Laura O. Moraes**[1]

[1]Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro – RJ – Brasil

william.pinho@edu.unirio.br, laura@uniriotec.br

Abstract. *Admission to public universities and military schools represents an opportunity to expand educational and professional perspectives. However, high competition and limited availability of positions make this process challenging for students from public education, who present diverse academic trajectories. This study aims to identify profiles of students graduating from the third year of high school in the Brazilian Military School System (SCMB), analyzing their academic trajectories and factors associated with admissions. A data analysis pipeline was proposed, emphasizing feature engineering to reduce dimensionality while preserving informational power. Unsupervised and supervised techniques were applied, including K-Means and DBSCAN, along with XGBoost for feature importance and SHAP for explainability. The results indicate three distinct student profiles, differing in academic performance, admission pathway, and family background.*

Resumo. *O ingresso em universidades públicas e escolas militares representa uma oportunidade de ampliar perspectivas educacionais e profissionais. No entanto, a alta concorrência e a disponibilidade limitada de vagas tornam esse processo desafiador para estudantes oriundos da rede pública de ensino, que apresentam trajetórias acadêmicas diversas. Este estudo tem como objetivo identificar perfis de estudantes concluintes do terceiro ano do ensino médio no Sistema Colégio Militar do Brasil, analisando suas trajetórias acadêmicas e fatores associados ao ingresso. Foi proposto um pipeline de análise de dados, com ênfase em engenharia de atributos para reduzir a dimensionalidade, preservando o poder informacional. Foram aplicadas técnicas não supervisionadas, K-Means e DBSCAN, juntamente com técnica supervisionada, XGBoost, para análise de importância das variáveis e SHAP para explicabilidade. Os resultados indicam três perfis distintos de estudantes, que diferem quanto ao desempenho acadêmico, forma de ingresso e contexto familiar.*

1. Introdução

O ingresso em universidades públicas e escolas militares amplia as perspectivas educacionais e profissionais de estudantes concluintes do ensino médio. Contudo, a elevada concorrência e a limitação de vagas tornam esses processos seletivos desafiadores, especialmente diante das diferentes trajetórias na educação básica, cenário que motivou políticas de reserva de vagas analisadas por [Lima et al. 2014].

A Mineração de Dados Educacionais (MDE) tem se consolidado como uma área relevante para a análise de dados educacionais. De acordo com [Romero and Ventura 2020], a MDE pode ser definida como a aplicação de técnicas de mineração de dados relacionados a ambientes educacionais, com o objetivo de abordar problemas relevantes para a educação. Uma das abordagens da área de MDE consiste na identificação de perfis de estudantes com características semelhantes, permitindo analisar diferentes trajetórias acadêmicas e seus possíveis desfechos, como pode ser visto no trabalho de [Najdi and Er-Raha 2016].

A identificação de perfis estudantis associados a desfechos educacionais pode contribuir para a compreensão de fatores relacionados ao ingresso em universidades públicas e escolas militares, permitindo que gestores e profissionais da educação desenvolvam estratégias mais direcionadas ao acompanhamento acadêmico dos estudantes. Com esse objetivo, este trabalho utiliza como estudo de caso o Sistema Colégio Militar do Brasil (SCMB), analisando dados acadêmicos e socioeconômicos de estudantes concluintes do Ensino Médio. Durante a exploração inicial dos dados, observou-se que a base apresentava características frequentemente encontradas em ambientes educacionais reais, como alta dimensionalidade, atributos categóricos de elevada cardinalidade, categorias com baixa frequência de ocorrência e forte concentração de registros com características semelhantes. Tais aspectos dificultam a aplicação direta de técnicas de aprendizado de máquina para descoberta e interpretação de perfis estudantis, motivando a investigação de abordagens capazes de lidar com esse tipo de complexidade.

Para enfrentar esse problema, este trabalho propõe uma abordagem baseada em engenharia de atributos orientada ao domínio, redução de dimensionalidade e técnicas de aprendizado de máquina. A abordagem foi avaliada por meio de métricas quantitativas e qualitativas de clusterização e explicabilidade das variáveis relevantes utilizando o *framework* SHAP, do inglês *SHapley Additive exPlanations*, proposto por [Lundberg and Lee 2017], tendo suas contribuições apresentadas na Tabela 1. Diante disso, este estudo busca responder às seguintes questões de pesquisa:

QP1: Como identificar perfis associados a desfechos educacionais em bases de dados caracterizadas por alta dimensionalidade, cardinalidade e forte presença de variáveis categóricas?

QP2: Quais perfis acadêmicos podem ser identificados a partir de dados acadêmicos e socioeconômicos de estudantes do SCMB?

Este artigo está organizado da seguinte forma: a Seção 2 discute os trabalhos relacionados; a Seção 3 descreve o conjunto de dados; a Seção 4 apresenta o processo SAPM (*Student Academic Profile Mining*), e suas estratégias de preparação de dados; a Seção 5 expõe os resultados obtidos face às questões de pesquisa; e a Seção 6 traz as considerações finais e limitações do estudo.

Tabela 1. Contribuições do trabalho

Eixo Dados Educacionais	Eixo Colégio Militar
1. Proposição do SAPM, um processo metodológico que sistematiza a mineração de dados educacionais para descoberta e interpretação de perfis em bases complexas. 2. Demonstração da efetividade do SAPM por meio de um estudo de caso com dados reais do Sistema Colégio Militar do Brasil.	1. Identificação e caracterização de perfis acadêmicos de estudantes concluintes do ensino médio provenientes de diferentes colégios militares brasileiros. 2. Análise das relações entre características acadêmicas, fatores socioeconômicos e desfechos reais de aprovação obtidos por esses estudantes.

Fonte: Elaborado pelos autores.

2. Trabalhos relacionados

Diferentes abordagens na literatura utilizam o aprendizado de máquina para analisar o contexto educacional. [Silva et al. 2025] aplicam o algoritmo K-Means a indicadores do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), para avaliar desigualdades educacionais, demonstrando a eficácia do algoritmo em bases públicas, embora sem focar no perfilamento individual de alunos; complementarmente, utilizam o AutoGluon[1] como modelo supervisionado. Focando na predição de perfis, [Nafuri et al. 2022] comparam algoritmos de clusterização combinados a preditores supervisionados para avaliar o desempenho de universitários. Também voltados ao comportamento estudantil, [Yang et al. 2021] empregam a clusterização hierárquica para associar padrões de aprendizagem aos resultados das avaliações. Embora os trabalhos analisados empreguem técnicas de aprendizado de máquina em contextos educacionais, diferem quanto aos tipos de dados utilizados, estratégias de engenharia de atributos, mecanismos de explicabilidade e desfechos investigados. Um comparativo entre os trabalhos pode ser visto na Tabela 2.

3. Dados

O estudo utilizou duas bases de dados com registros de 2021 a 2023. Os desfechos educacionais não são registrados pelo SINCOMIL[2], mas em registros próprios dos colégios militares, de modo que as duas bases se complementam quanto às informações acadêmicas e de desfecho educacional. O período foi delimitado de modo a evitar anos letivos afetados pela pandemia de COVID-19, uma vez que a literatura aponta impactos significativos da pandemia sobre o desempenho acadêmico e a comparabilidade entre avaliações presenciais e remotas, como visto no estudo de [Vicario et al. 2024], o que poderia introduzir viés na caracterização dos perfis estudantis. A primeira base, extraída do SINCOMIL, reúne variáveis acadêmicas e socioeconômicas de concluintes do Ensino Médio de 11 Colégios Militares (CM). A segunda mapeia as aprovações em universidades públicas e escolas militares. Por depender de registros voluntários dos estudantes ou de listas oficiais, esta última base restringe-se aos casos confirmados pelos CM. O código

[1] *Framework* que automatiza o processo de treinamento, seleção e avaliação de modelos.

[2] Sistema Integrado dos Colégios Militares, utilizado pelo Sistema Colégio Militar do Brasil para registro, gestão de informações acadêmicas e administrativas de alunos e responsáveis.

Tabela 2. Comparação metodológica dos trabalhos relacionados

Trabalho	**Dados**	**Clusterização**	**Aprendizado Supervisionado**	**Engenharia de Atributos**	**Avaliação / Explicabilidade**	**Desfecho**
Silva et al. (2025)	Indicadores educacionais (INEP)	K-Means	Sim (AutoGluon)	Agregação + normalização	Índice normalizado de desempenho dos *clusters* / Sem explicabilidade	Evolução temporal
Nafuri et al. (2022)	Dados acadêmicos e socioeconômicos	K-Means, BIRCH, DBSCAN	Sim (Decision Tree, Random Forest e ANN)	Seleção + derivação de atributos	DBI, Silhouette Score e Índice Calinski-Harabasz / Sem explicabilidade	Evasão, graduação, emprego
Yang et al. (2021)	Logs comportamentais (e-learning)	Hierárquico (Ward)	Não	Extração de indicadores	Interpretação de perfis diretamente / Sem explicabilidade	Notas em exame
Este trabalho	Trajetórias acadêmicas + aprovação	K-Means, DBSCAN	Sim (XGBoost)	Redução de dimensionalidade orientada ao domínio	DBI, Silhouette Score, t-SNE, Distribuições por *cluster* e SHAP	Aprovação real

Fonte: Elaborado pelos autores.

utilizado no treinamento dos modelos, assim como informações complementares sobre o conjunto de dados empregado, está disponível em um repositório anônimo no seguinte endereço: `https://doi.org/10.5281/zenodo.20786834`.

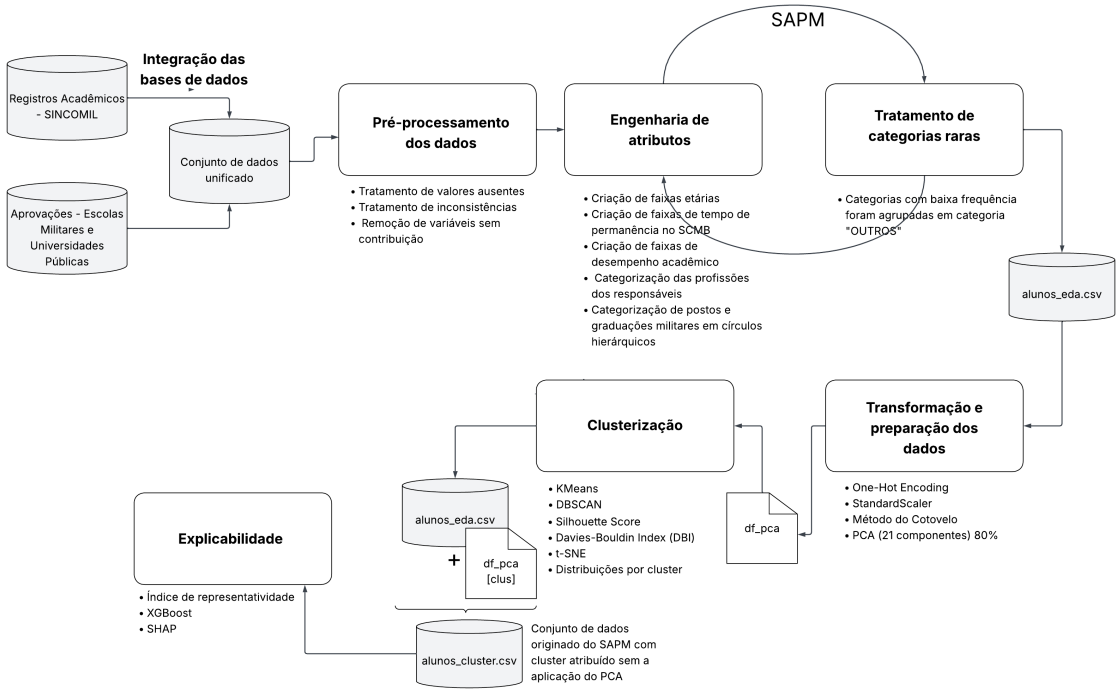
4. Metodologia

A metodologia adotada neste trabalho utilizou o SAPM, apresentado na Figura 1. É importante destacar que o SAPM não constitui uma nova técnica de MDE, mas organiza e integra técnicas consolidadas em um processo para análise de bases educacionais complexas. As descrições detalhadas do pipeline são descritas nesta seção.

Na fase de **integração das bases de dados** as bases descritas na Seção 3 foram unificadas utilizando como critério identificadores sequenciais dos estudantes e dos CM.

Durante a fase de **pré-processamento dos dados**, foi realizado um processo de limpeza e preparação. Inicialmente, foram tratados valores ausentes e inconsistências nos registros, por meio de procedimentos de padronização e validação das informações. Além disso, nesta fase, o conjunto de dados possuía originalmente 67 variáveis. As variáveis identificadoras foram removidas por não apresentarem relevância para o domínio do problema investigado. Também foi definido um *threshold* de 95% para identificar variáveis com predominância de um mesmo valor, as quais foram excluídas do conjunto de dados devido à sua baixa contribuição analítica, resultando no conjunto final utilizado nas análises contendo 21 variáveis.

Figura 1. SAPM, aplicado ao pipeline de MDE



Fonte: Elaborado pelos autores.

A fase de **engenharia de atributos** baseou-se em características do domínio educacional para definir categorias de notas e profissões, um tratamento não observado nos trabalhos analisados (Seção 2). Foram derivados novos atributos das variáveis originais para melhorar a interpretabilidade e mitigar a maldição da dimensionalidade, fenômeno amplamente discutido na literatura e associado ao aumento exponencial do volume do espaço de atributos à medida que novas dimensões são adicionadas [Hastie et al. 2009]. As principais transformações e reduções de complexidade resultantes estão descritas na Tabela 3.

Tabela 3. Redução da complexidade dos atributos após a aplicação do SAPM

Transformação	Antes	Depois
Representação da idade	8 valores distintos	3 faixas etárias
Disciplinas	12 variáveis acadêmicas	4 áreas do conhecimento
Notas	55–56 valores distintos por disciplina	4 faixas de desempenho
Tempo colégio (faixas)	12 valores distintos	3 faixas de tempo

Fonte: Elaborado pelos autores.

Para contornar a elevada cardinalidade das variáveis categóricas originais, como as ocupações dos responsáveis, que alcançavam até 361 categorias e inviabilizavam abordagens diretas como one-hot encoding, realizou-se uma agregação em categorias. As profissões foram agrupadas com base no nível de escolaridade. Esse processo reduziu significativamente a dimensionalidade sem perda do poder informacional, beneficiando a qualidade dos agrupamentos. Essas transformações específicas estão resumidas na Tabela 4.

Por fim, para o tratamento das variáveis de desfecho, registros com aprovações simultâneas em universidades públicas e escolas militares foram decompostos em duas variáveis

Tabela 4. Transformações aplicadas às variáveis de ocupação dos responsáveis

Variável	Categorias		Descrição (Agrupamentos)
	Antes	Depois	
Posto do pai	32	3	oficiais, praças e não declarado
Profissão do pai	130	10	escolaridade e categorias específicas
Profissão da mãe	361	10	Agrupamento semelhante ao do pai
Círculo + Profissão pai	Separadas	Unificadas	Integração em uma única variável de ocupação

Fonte: Elaborado pelos autores.

binárias independentes, permitindo capturar adequadamente cada tipo de aprovação de forma isolada.

O **tratamento de categorias raras** foi necessário, uma vez que, na formação de perfis de alunos, categorias com baixas frequências de ocorrência apresentam pouco poder de separação para os algoritmos de clusterização. Assim, optou-se por agrupar essas variáveis, tratadas aqui como "categorias raras", em uma classe unificada chamada "outros".

Na fase de **transformação e preparação dos dados**, as variáveis de aprovação foram temporariamente removidas e reintroduzidas apenas após a formação dos agrupamentos. Essa estratégia garantiu que os desfechos não influenciassem a formação dos *clusters*, definindo os perfis exclusivamente por características acadêmicas e contextuais, além de mitigar o viés de seleção. Adicionalmente, incluíram-se todos os alunos disponíveis na base, mesmo sem confirmação em processos seletivos, evitando a redução da amostra e potenciais vieses. Em seguida, aplicou-se a técnica de one-hot encoding para converter as variáveis categóricas em representações numéricas e, por fim, utilizou-se o Standard Scaler para a padronização das escalas dos dados. Embora com a redução prévia de atributos, o conjunto de dados resultou em 58 variáveis após o one-hot encoding. Para contornar esse cenário, aplicou-se a Análise de Componentes Principais (*Principal Component Analysis* – PCA), uma das abordagens discutidas em [Sorzano et al. 2014]. O PCA atuou de forma complementar, com 21 componentes e 80% da variância, permitindo representar os dados em um espaço de menor dimensão e preservando a maior parte da variabilidade original, descritas no trabalho de [Jolliffe and Cadima 2016].

Na etapa de **clusterização**, foram utilizadas abordagens baseadas em centróides utilizando o K-Means, cujo número de grupos k foi avaliado pelo método do cotovelo, proposto no trabalho de [Thorndike 1953], com base na variação da dispersão *intra-cluster*, e na abordagem baseada em densidade com o DBSCAN, configurado por meio de uma busca sistemática via Grid Search, buscando verificar qual paradigma produzia agrupamentos mais adequados para análise dos perfis estudantis.

A escolha do algoritmo definitivo baseou-se na análise combinada de métricas quantitativas e aspectos qualitativos. Avaliaram-se os índices de Davies-Bouldin (DBI), proposto originalmente no trabalho de [Davies and Bouldin 1979], índice da silhueta, proposto no trabalho de [Rousseeuw 1987], da quantidade de agrupamentos formados e da distribuição dos dados entre os grupos e as projeções geradas pelo t-SNE, presente no trabalho de [van der Maaten and Hinton 2008], buscando uma configuração com boa separação, coesão interna e equilíbrio representativo. Para identificar as características mais marcantes de cada grupo, adotou-se o índice de razão de proporções, definido como $I_{c,k} = \frac{P(c|k)}{P(c)}$, onde $P(c|k)$ é a proporção da categoria c no grupo k e $P(c)$ é a proporção da mesma categoria na população total. Valores de

$I_{c,k} > 1$ indicam sobre-representação da característica no grupo.

O *pipeline* é finalizado na etapa de **explicabilidade**. O objetivo foi validar os perfis identificados treinando um modelo supervisionado utilizando o algoritmo XGBoost. Embora o PCA tenha sido utilizado na etapa de clusterização para redução da dimensionalidade, o modelo XGBoost empregado nesta etapa foi treinado utilizando as variáveis originais processadas pelo SAPM após a aplicação do one-hot encoding. Essa estratégia permitiu preservar a interpretabilidade das características originais dos estudantes durante a análise por SHAP, técnica baseada nos valores de Shapley, da teoria dos jogos cooperativos [Lundberg and Lee 2017]. O SHAP atribui a cada variável um valor de contribuição para a predição do modelo em cada instância. Valores positivos aumentam a probabilidade de a instância pertencer ao *cluster* analisado; valores negativos indicam o efeito oposto, e a magnitude reflete a intensidade dessa contribuição.

5. Resultados

Nesta seção, são apresentados os resultados da aplicação dos algoritmos de clusterização, classificação e explicabilidade sobre o conjunto de dados processado por meio do pipeline proposto.

5.1. Algoritmos de aprendizado não supervisionado

Destaca-se que foram realizados testes considerando um intervalo de $3 \leq k \leq 8$; contudo, optou-se por apresentar na Tabela 5 apenas as configurações até $k = 4$ e a análise sem o SAPM, por questões de espaço e relevância analítica, visto que soluções com maior granularidade não agregaram valor descritivo aos perfis. Embora o K-Means tenha apresentado um índice Davies-Bouldin superior ao produzido pelo DBSCAN (quanto menor o índice, melhor o *cluster*), produziu três *clusters* bem distribuídos. Por outro lado, o DBSCAN gerou agrupamentos mais concentrados, o que contribuiu para diminuir o valor do índice, mas limita sua aplicação na análise de perfis, uma vez que se tornam muito específicos. Dessa forma, o K-Means mostrou-se mais adequado para a análise proposta. Para fins de comparação, também se aplicou o K-Means diretamente sobre os dados sem as etapas de engenharia de atributos do SAPM (discretização de faixas, categorização de alta cardinalidade e agrupamento de categorias raras), resultado identificado como "Sem SAPM" na Tabela 5 e na Figura 2(b).

Tabela 5. Resultados comparativos dos algoritmos K-Means e DBSCAN.

Critérios	**K-Means** $k = 3$	**K-Means** $k = 4$	**K-Means** $k = 3$ *Sem SAPM*	**DBSCAN** $eps = 7{,}0$ $MinPts = 10$
Número de *Clusters*	3	4	3	5
Amostras em *Clusters*	5493	5493	5493	5143
Ruídos	-	-	-	350
Índice Davies-Bouldin	1,80	1,63	2,09	0,99
Índice da Silhueta	0,17	0,12	0,13	0,30
Interpretação (*Clusters*)	Alta	Alta	Muito Baixa	Baixa
Distribuição (*Clusters*)	Equilibrado	Equilibrado	Esparsos	Concentrado
Análise de perfis	Alta	Alta	Muito Baixa	Baixa

Fonte: Elaborado pelos autores.

A distribuição dos perfis pode ser observada na Tabela 6 e na Figura 2, onde é apresentada a projeção t-SNE dos melhores agrupamentos gerados pelos algoritmos K-Means e DBSCAN.

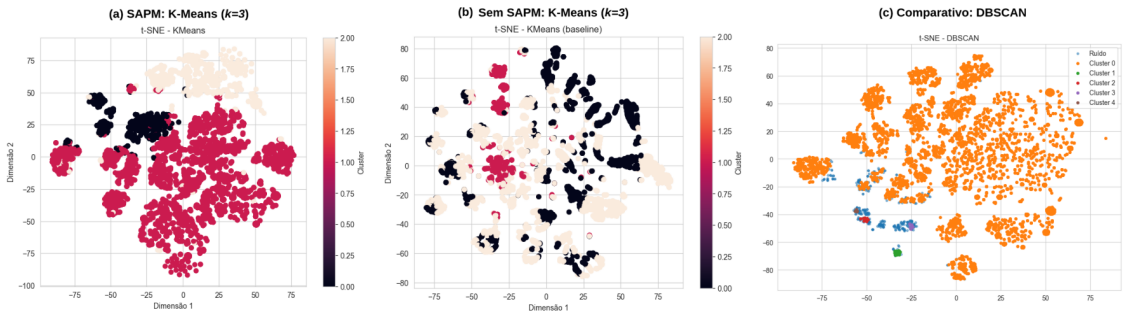
Em (a) O modelo K-Means proposto apresenta regiões bem delimitadas e volumétricas, em (b) é apresentado o algoritmo K-Means aplicado a um pipeline de MDE convencional, omitindo os processos de discretização de faixas, categorização de alta cardinalidade e agrupamento de categorias raras, o que resulta em alta fragmentação e sobreposição dos grupos, já em (c) O algoritmo DBSCAN evidencia falha por densidade, agrupando a maioria das instâncias no *Cluster* 0. O K-Means apresenta uma divisão mais equilibrada entre os *clusters*, enquanto o DBSCAN concentra a maior parte dos dados em poucos grupos e mantém uma parcela como ruído. Embora a configuração com $k = 4$ tenha apresentado um índice Davies-Bouldin ligeiramente inferior ao obtido com $k = 3$, a análise qualitativa dos agrupamentos indicou que o quarto *cluster* correspondia a uma subdivisão de um perfil já identificado. Esse agrupamento era composto predominantemente por estudantes com características muito semelhantes às observadas no *Cluster* 1, diferenciando-se principalmente pela presença de mães com vínculo militar. As demais características acadêmicas, familiares e de trajetória escolar permaneceram praticamente inalteradas.

Tabela 6. Distribuição dos *clusters* considerando os melhores resultados obtidos entre o K-Means e DBSCAN, com a utilização do SAPM.

Cluster	K-Means	DBSCAN
-1	-	350
0	605	5063
1	4001	37
2	887	21
3	-	12
4	-	10

Fonte: Elaborado pelos autores.

Figura 2. Projeção t-SNE dos agrupamentos gerados pelos algoritmos K-Means (com e sem SAPM) e DBSCAN.



Fonte: Elaborado pelos autores.

5.2. Clusterização

O número de *clusters* foi definido com base no método do cotovelo, embora não tenha sido possível observar um ponto de curva acentuada. Dessa forma foram testados valores entre $k = 3$ e $k = 8$, sendo a configuração com $k = 3$ apresentando melhor equilíbrio e discriminação entre os grupos.

5.3. Caracterização dos *clusters*

A caracterização dos *clusters* foi realizada em três etapas: (1) as taxas de aprovação em universidades públicas e escolas militares, apresentadas na Tabela 7, foram utilizadas para avaliar o tipo

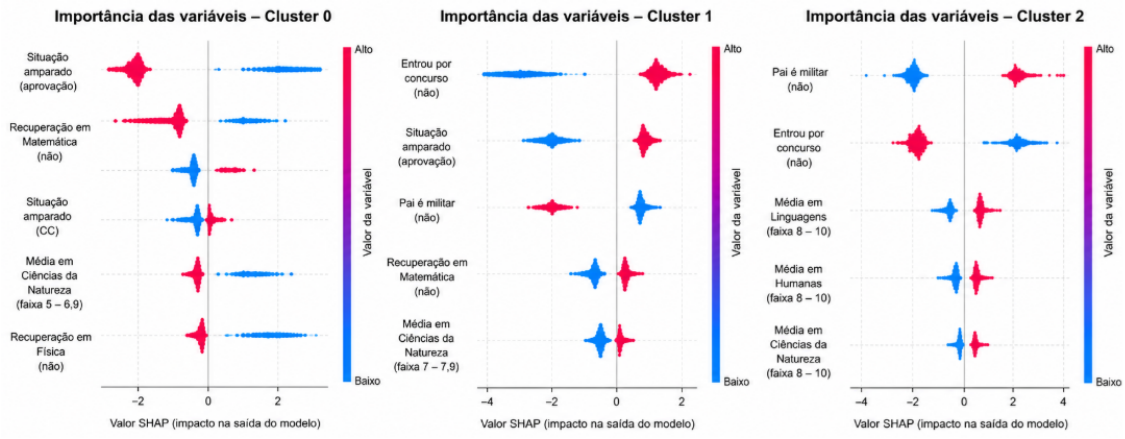
de aprovação dos perfis identificados; (2) os índices apresentados na Tabela 8 indicam as cinco variáveis com maior incidência proporcional no *cluster* em relação a população total; e (3) os fatores mais relevantes para a diferenciação dos grupos foram analisados por meio dos valores de SHAP (Figura 3), onde os pontos representam instâncias (estudantes), a cor indica o valor do atributo (rosa/alto para presença da categoria binarizada, azul/baixo para ausência). O deslocamento à direita (SHAP value > 0) indica aumento na probabilidade de pertencer ao respectivo *cluster*.

Tabela 7. Resumo comparativo dos *clusters*

Característica	***Cluster* 0**	***Cluster* 1**	***Cluster* 2**
Desempenho acadêmico	Baixo (5-6,9)	Médio (7-7,9)	Alto (8-10)
Recuperação	Alta	Baixa	Muito baixa
Forma de ingresso	Não concursado	Não concursado	Concursado
Principal fator (SHAP)	Baixo desempenho + recuperações	Contexto familiar + estabilidade acadêmica	Alto desempenho + concurso
Aprovações em Universidades	0,097	0,120	0,331
Aprovações Militares	0,00	0,013	0,060

Fonte: Elaborado pelos autores.

Figura 3. Agrupamentos via SHAP *beeswarm plots* para o classificador XGBoost.



Fonte: Elaborado pelos autores.

Cabe destacar que, por decorrência do processo de one-hot encoding, algumas variáveis binarizadas representam a ausência da categoria original — como "Recuperação em Matemática (não)", que indica que o estudante não passou por recuperação. Nesses casos, a interpretação do SHAP deve considerar essa inversão: no *Cluster 0*, por exemplo, valores altos (rosa) dessa variável, ou seja, estudantes que não fizeram recuperação, deslocam-se para a esquerda (SHAP negativo), reduzindo a probabilidade de pertencimento ao grupo; inversamente, valores baixos (azul), que indicam que o estudante de fato precisou de recuperação, deslocam-se para a direita (SHAP positivo), aumentando essa probabilidade. Essa leitura é consistente com o perfil do *Cluster 0*, caracterizado por alta incidência de recuperação (Tabela 7).

Tabela 8. Principais características identificadas em cada *cluster*

(a) ***Cluster* 0**

Variável	Valor	Índice
FISICA_REC	yes	8,98
LE_REC[1]	yes	8,88
BIOLOGIA_REC	yes	8,82
QUIMICA_REC	yes	8,63
STATUS	aprovado_cc[2]	8,50

(b) ***Cluster* 1**

Variável	Valor	Índice
MILITAR_PAI	yes	1,19
CONCURSADO	no	1,18
PROFISSAO_PAI	Oficiais	1,18
M_EXATAS_FX[3]	7_7.9	1,17
M_NATURAIS_FX[4]	7_7.9	1,13

(c) ***Cluster* 2**

Variável	Valor	Índice
PROFISSAO_PAI	outros	5,87
CONCURSADO	yes	5,75
MILITAR_PAI	no	5,38
IDADE_FAIXA	maiorigual19	2,32
M_EXATAS_FX[3]	8_10	2,27

[1] LE_REC: Recuperação em língua estrangeira.
[2] Aprovado em conselho de classe.
[3] M_EXATAS_FX: MEDIA_EXATAS_FAIXA;
[4] M_NATURAIS_FX: MEDIA_CIENCIAS_NATURAIS_FAIXA.

Fonte: Elaborado pelos autores.

O ***Cluster 0*** apresenta um perfil acadêmico de baixo desempenho, com médias concentradas entre 5 e 6,9 e alta incidência de recuperação, indicando uma trajetória irregular. A análise de SHAP aponta que variáveis associadas à dificuldade acadêmica são os principais fatores de pertencimento ao grupo. Em termos de desfecho, o *cluster* apresenta a menor proporção de aprovação em universidades públicas e desempenho nulo em escolas militares.

O ***Cluster 1*** concentra a maior parte dos estudantes e apresenta perfil intermediário e estável, com médias entre 7 e 7,9, baixa incidência de recuperação e predominância de alunos não concursados de famílias militares. O SHAP indica forte influência do contexto familiar militar e da estabilidade acadêmica. Este grupo apresenta aprovação moderada em universidades públicas e desempenho intermediário em escolas militares.

O ***Cluster 2*** é caracterizado por estudantes ingressantes por concurso público, predominantemente sem vínculo militar e com alto desempenho (médias entre 8 e 10). O SHAP confirma o impacto do sucesso escolar e do ingresso via concurso, refletindo-se nos desfechos com as maiores proporções de aprovação tanto em universidades públicas quanto em escolas militares.

5.4. Resposta às Questões de Pesquisa

A **QP1** foi respondida mediante a proposição e avaliação do processo SAPM. Em comparação ao pipeline convencional sem SAPM, a abordagem proposta reduziu o índice Davies-Bouldin de 2,09 para 1,80, correspondendo a uma melhoria de aproximadamente 13,9% na qualidade dos agrupamentos. A **QP2** foi atendida com a identificação de três perfis estudantis distintos: o *Cluster* 0, de maior vulnerabilidade e trajetória irregular; o *Cluster* 1, com desempenho intermediário e forte vínculo familiar militar; e o *Cluster* 2, composto por alunos concursados de elevado desempenho e com as maiores taxas de aprovação em concursos, sintetizados na Tabela 7.

6. Conclusão

Os resultados sugerem a existência de três perfis estudantis identificados. O primeiro grupo apresenta maior vulnerabilidade acadêmica, com as menores taxas de aprovação em escolas militares e universidades públicas. O segundo grupo, embora apresente desempenho acadêmico intermediário, é caracterizado por forte vínculo com o contexto militar, o que sugere que fatores familiares podem contribuir para os resultados obtidos, ainda que de forma complementar ao desempenho acadêmico. O terceiro é composto por estudantes com elevado desempenho acadêmico, que, mesmo sem vínculo familiar militar, apresentam maior proporção de aprovação, indicando que a excelência acadêmica constitui o principal fator associado à aprovação nesse tipo de seleção.

A engenharia de atributos mostrou-se central para a identificação dos agrupamentos, ao incorporar conhecimento de domínio na transformação de variáveis de alta cardinalidade, reduzindo a dimensionalidade sem perda de poder informacional. Os resultados oferecem ao SCMB uma visão estruturada das trajetórias estudantis, fornecendo subsídios para estratégias educacionais voltadas ao aumento de aprovações. Constatou-se que o desempenho acadêmico é o principal fator associado ao êxito, enquanto aspectos contextuais atuam de forma complementar. A análise não supervisionada e a modelagem supervisionada foram validadas pelo modelo XGBoost combinado à técnica SHAP, que ratificou a relevância das variáveis em cada perfil. Como limitações, destaca-se a disponibilidade restrita de dados de aprovação, uma vez que os desfechos foram observados apenas para os resultados de aprovações informados. Como trabalhos futuros, propõe-se avaliar quantitativamente a perda de granularidade decorrente da discretização de variáveis contínuas (Tabela 3), bem como investigar abordagens alternativas de agrupamento capazes de identificar o grau de pertencimento dos estudantes aos *clusters*, dada a presença de ruídos relevantes apontada pelo DBSCAN.

Referências

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Lima, M. E. O., Neves, P. S. d. C., and Silva, P. B. e. (2014). A implantação de cotas na universidade: paternalismo e ameaça à posição dos grupos dominantes. *Revista Brasileira de Educação*, 19:141–163.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*.

Nafuri, A. F. M., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., and Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences*, 12(19):9467.

Najdi, L. and Er-Raha, B. (2016). Implementing cluster analysis tool for the identification of students typologies. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 575–580.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wires Data Mining and knowledge discovery*.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Silva, M. L. d. M., Cruz, L. A., Magalhães, R. P., Pires, T. G., Macedo, J. A., and Andrade, R. M. d. C. (2025). Desigualdades educacionais no brasil: Uma análise por clusterização de indicadores educacionais e de desempenho escolar. In *Proceedings of the 40th Brazilian Symposium on Databases (SBBD)*, Fortaleza, CE, Brazil.

Sorzano, C. O. S., Vargas, J., and Montano, A. P. (2014). A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*.

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

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Vicario, C. M., Mucciardi, M., Perconti, P., Lucifora, C., Nitsche, M. A., and Avenanti, A. (2024). The impact of the COVID-19 pandemic on academic performance: a comparative analysis of face-to-face and online assessment. *Frontiers in Psychology*, 14:1299136.

Yang, C. C. Y., Chen, I. Y. L., and Ogata, H. (2021). Toward precision education: Educational data mining and learning analytics for identifying students' learning patterns with ebook systems. *Educational Technology & Society*.