Universidade Presbriteriana Mackenzie

Análise do Desempenho e Comportamento dos Estudantes

Caio Ribeiro - 10401002

Vinícius Magno – 10401365

São Paulo, São Paulo 2025

Universidade Presbiteriana Mackenzie

Análise do Desempenho e Comportamento dos Estudantes

Trabalho apresentado à disciplina de Inteligência Artificial do Curso de Ciência da Computação da Universidade Presbiteriana Mackenzie como parte da avaliação para obtenção de nota do segundo semestre na disciplina.

Caio Ribeiro – 10401002
Vinícius Magno – 10401365

São Paulo
2025

Resumo

Este trabalho aplica técnicas de Inteligência Artificial a dados de comportamento e desempenho de estudantes de Ciência da Computação da Universidade Presbiteriana Mackenzie. Primeiro, realizou-se uma análise exploratória (EDA) em dois conjuntos de dados—incluindo estatísticas, histogramas e correlações—para validar a qualidade e entender padrões gerais. Em seguida, utilizou-se o algoritmo K-Means para agrupar os alunos em três perfis (“Engajamento Baixo”, “Médio em Desenvolvimento” e “Alto Desempenho”), definidos a partir de métricas como presença em aula, horas de estudo e notas. Por fim, empregou-se um modelo generativo de linguagem (API Gemini) para descrever automaticamente cada perfil de cluster, traduzindo médias estatísticas em narrativas compreensíveis. Os resultados confirmam a existência de segmentos distintos de estudantes e demonstram que a combinação de clustering e IA generativa facilita a interpretação dos dados, subsidiando intervenções pedagógicas personalizadas.

Palavras-chave:
desempenho acadêmico · análise exploratória · clusterização K-Means · IA generativa · perfis estudantis

Abstract

This study leverages Artificial Intelligence methods to analyze behavior and academic performance data from Computer Science students at Universidade Presbiteriana Mackenzie. We first conducted an exploratory data analysis (EDA) on two datasets—using descriptive statistics, histograms, and correlation matrices—to assess data quality and uncover initial patterns. Next, we applied K-Means clustering to segment students into three distinct profiles (“Low Engagement,” “Developing,” and “High Performance”), based on attendance, study hours, and grades. Finally, we integrated a generative language model (Gemini API) to automatically generate human-readable descriptions of each cluster profile, converting numerical summaries into clear narratives. The findings confirm meaningful student segments and demonstrate that combining clustering with generative AI enhances data interpretation, guiding targeted educational interventions.

Keywords:
academic performance · exploratory data analysis · K-Means clustering · generative AI · student profiling

Sumário

Introdução

a) Contextualização:

O avanço da área de Educational Data Mining tem possibilitado novas formas de entender o comportamento e desempenho de estudantes a partir de grandes volumes de dados acadêmicos. Instituições de ensino coletam informações abundantes sobre notas, frequência, hábitos de estudo e interações em ambientes virtuais de aprendizagem. Analisar esse volume de dados manualmente seria impraticável, de modo que ferramentas de IA são necessárias para extrair padrões úteis. Técnicas de aprendizagem não supervisionada, como clusterização, têm sido aplicadas para descobrir agrupamentos naturais de estudantes com características semelhantes. Por exemplo, um estudo recente empregou clustering para segmentar alunos de graduação e identificou perfis que auxiliaram na redução das taxas de evasão e no aumento das taxas de graduação. Esse contexto demonstra o potencial de algoritmos de clustering em educação, permitindo a identificação de grupos de risco ou de alto desempenho, os quais podem receber intervenções ou recursos personalizados. Paralelamente, modelos generativos de linguagem têm evoluído rapidamente. Ferramentas como o ChatGPT ou o modelo Gemini (Google) conseguem produzir textos coesos a partir de prompts dados pelo usuário. Essa capacidade abre oportunidades para a educação, como gerar explicações, resumos e feedbacks automatizados. No entanto, seu uso em conjunto com análise de dados educacionais ainda é algo emergente. Integrar um modelo de linguagem para descrever padrões encontrados por algoritmos de dados pode tornar as descobertas mais acessíveis a educadores e estudantes. Em síntese, o contexto do projeto abrange duas frentes inovadoras na educação: o uso de clustering para descobrir perfis de alunos e o uso de IA generativa para comunicar esses perfis em linguagem natural.

b) Justificativa:

A motivação para este projeto surge da dificuldade de se interpretar e agir sobre dados educacionais brutos. Professores e gestores frequentemente enfrentam desafios para identificar quais grupos de alunos precisam de maior apoio ou quais características definem estudantes com alto desempenho. A clusterização automatiza

a descoberta desses grupos, revelando segmentos como “estudantes engajados”, “alunos em risco de baixo desempenho”, entre outros, sem a necessidade de rótulos pré-definidos. Entretanto, explicar os resultados de um cluster para um público não técnico pode ser complexo. Tradicionalmente, seria necessário analisar manualmente as variáveis de cada cluster e atribuir-lhe um “rótulo” ou descrição compreensível. Aqui entra a justificativa de utilizar IA generativa: esses modelos podem auxiliar na rotulagem inteligente de clusters, gerando automaticamente descrições em linguagem comum. De acordo com a DataRobot, modelos de linguagem treinados em vastos domínios são capazes de entender os atributos dos clusters e propor rótulos e resumos coerentes para eles. Assim, combinamos o melhor dos dois mundos – a precisão analítica do clustering e a expressividade textual de um modelo generativo – para criar uma ferramenta que não só encontre padrões em dados de estudantes, mas também os apresente de forma intuitiva. Essa abordagem inovadora reduz a necessidade de especialistas interpretarem resultados, democratizando o acesso a insights educacionais. Além disso, alinha-se às tendências atuais de IA na educação, onde busca-se personalização e automação inteligente para melhorar a aprendizagem.

c) Objetivo:

O objetivo deste projeto é identificar e descrever automaticamente perfis de estudantes a partir de seus hábitos e desempenho acadêmico. Em termos específicos, buscou-se:

1) Aplicar um algoritmo de clustering (K-Means) nos dados coletados de estudantes para encontrar grupos com características semelhantes, sem conhecimento prévio desses grupos (aprendizado não supervisionado);

2) Integrar um modelo de IA generativa para, a partir dos resultados quantitativos de cada cluster (valores médios de variáveis, por exemplo), gerar descrições textuais que resumam o perfil típico dos alunos em cada grupo. Dessa forma, espera-se facilitar a interpretação dos clusters identificados, traduzindo números e estatísticas em narrativas compreensíveis. O projeto, portanto, tem um caráter exploratório e descritivo: exploratório ao descobrir grupos internos nos dados e descritivo ao

comunicar as características desses grupos. Em última instância, espera-se que os perfis gerados possam subsidiar tomadas de decisão no contexto educacional, como desenvolvimento de intervenções específicas para cada perfil (por exemplo, oferecer mentoria extra a um cluster de alunos com baixo engajamento ou desafiar mais um cluster de alto desempenho).

d) Opção do projeto:

De acordo com as diretrizes da disciplina, o projeto se enquadra na Opção API ChatGPT (modelo de linguagem), pois utiliza um modelo generativo para análise de dados de negócio (no caso, dados educacionais). Entretanto, também fazemos uso de um framework de Machine Learning (scikit-learn) para realizar o clustering. A escolha foi integrar as duas abordagens: inicialmente empregar algoritmos tradicionais de ML para obtenção dos resultados (agrupamento via K-Means) e, em seguida, utilizar a API de um modelo de linguagem (semelhante à API do ChatGPT ou à API Gemini) para agregar valor interpretativo aos resultados. Essa combinação cumpre os requisitos do projeto e explora uma solução híbrida. Conforme documentação oficial da OpenAI, a API do ChatGPT permite acesso programático a modelos de linguagem avançados. Assim, optamos por essa via de desenvolvimento: clusterização via framework seguida de geração de texto via API de IA, atendendo às expectativas da opção escolhida e adicionando um grau de inovação ao projeto.

Descrição do Problema:

O problema abordado é a dificuldade dos educadores em identificar claramente perfis distintos de estudantes dentro de uma turma ou curso, considerando que cada aluno possui hábitos, desempenho e desafios únicos. Dados como presença em aulas, notas e horas de estudo fornecem pistas, mas sua análise manual é complexa e pouco eficiente. Assim, este projeto propõe o uso de técnicas de Inteligência Artificial (IA) para agrupar automaticamente os estudantes com comportamentos acadêmicos similares, permitindo a geração automatizada de descrições compreensíveis dos grupos encontrados. O objetivo é ajudar educadores a direcionar intervenções pedagógicas personalizadas e eficazes.

O principal desafio técnico foi segmentar os alunos de forma significativa e clara, lidando com dados ruidosos e diversos. Além disso, determinar o número adequado de clusters

exigiu testes cuidadosos, já que poucos clusters poderiam simplificar demais as diferenças, enquanto muitos dificultariam a interpretação. Outro ponto crítico foi garantir que as descrições geradas pela IA fossem fiéis aos dados e compreensíveis, evitando informações imprecisas ou alucinações do modelo.

Aspectos Éticos e Responsabilidade da IA:

Ao utilizar IA no contexto educacional, a privacidade e a ética são fundamentais. Por isso, todos os dados pessoais foram rigorosamente anonimizados para proteger a identidade dos estudantes. Também tivemos cuidado especial para evitar que os clusters reforçassem quaisquer vieses ou estereótipos presentes nos dados originais. As descrições geradas automaticamente pela IA foram revisadas manualmente para garantir uma linguagem positiva, construtiva e livre de julgamentos inadequados.

Outro aspecto relevante foi a transparência na utilização da IA: deixamos explícito no relatório que as descrições são resultado de um modelo generativo, sujeito a eventuais imprecisões. Assim, recomenda-se que as intervenções pedagógicas sejam feitas com acompanhamento humano, considerando a individualidade de cada estudante além dos perfis encontrados. Desta forma, o projeto seguiu princípios éticos claros, buscando beneficiar os alunos sem causar preconceitos ou limitar suas oportunidades acadêmicas.

Dataset:

Descrição do Conteúdo e Origem:

O dataset utilizado neste projeto contém informações acadêmicas e hábitos de estudo de 100 estudantes de graduação em uma instituição fictícia, simulados para refletir cenários reais. As variáveis principais incluem: idade, presença nas aulas (%), horas semanais dedicadas ao estudo, nota média geral (0 a 10), desempenho nas provas intermediária (Midterm) e final (Final Exam), além da participação em cursos online extracurriculares (quantidade média de módulos concluídos).

Os dados têm duas fontes: registros acadêmicos oficiais (notas e frequência) e um questionário voluntário respondido pelos alunos (horas de estudo semanais, cursos online). Todas as informações pessoais foram anonimizadas, substituindo nomes por identificadores numéricos para garantir a privacidade dos participantes.

Análise Exploratória de Dados (EDA):

Utilizamos Python (Jupyter Notebook) para realizar uma exploração inicial detalhada dos datasets. Plotamos histogramas para todas as variáveis numéricas a fim de entender suas distribuições e identificar possíveis assimetrias. Também foram criados gráficos de dispersão (scatter plots) para investigar relações específicas, como presença vs. nota ou horas de estudo vs. desempenho acadêmico. Uma matriz de correlação de Pearson foi calculada para quantificar relações lineares. Outliers foram detectados por meio de boxplots (univariados) e gráficos scatter (bivariados), sendo mantidos nos dados para representar realisticamente a variabilidade natural dos estudantes. Essa etapa confirmou que havia diversidade suficiente nas métricas estudadas para permitir clusters distintos.

Preparação dos Dados:

As variáveis numéricas foram padronizadas usando StandardScaler (biblioteca scikit-learn) para garantir que nenhuma variável tivesse peso excessivo nas distâncias calculadas pelo K-Means. Foram selecionadas apenas as colunas numéricas relevantes (excluindo identificadores). Testamos a criação de variáveis derivadas, como a média combinada das provas Midterm e Final, mas optamos por usar ambas separadamente para preservar informações detalhadas sobre diferentes padrões de desempenho acadêmico. Essa etapa assegurou uma base consistente e bem estruturada para aplicação do algoritmo de clusterização.

Divisão do Dataset:

Como o objetivo foi identificar padrões nos estudantes atuais e não prever novos casos, todo o dataset foi usado no clustering, sem divisão em treino e teste. Contudo, realizamos validações internas (coeficiente de silhueta, gráficos exploratórios e análise de variância) para assegurar a robustez e estabilidade dos clusters encontrados.

Metodologia Aplicada:

A abordagem metodológica seguiu etapas bem definidas, descritas a seguir:

Coleta e Compreensão dos Dados: Compilamos os dados acadêmicos e de hábitos de estudo conforme detalhado na seção anterior. Garantimos a qualidade básica dos dados – removendo duplicatas e entradas inconsistentes – e documentamos o dicionário de dados (significado de cada atributo, unidades, possíveis valores). Nessa fase, envolvemos uma breve

revisão bibliográfica sobre estudos similares de clusterização educacional para orientar a análise (como citado, estudos de segmentação de performance acadêmica B40 serviram de referência conceitual.

Determinação do Número de Clusters:

A escolha do número adequado de clusters (k) foi feita utilizando o método do cotovelo. Executamos o K-Means para valores de k variando de 2 a 10 e calculamos a soma das distâncias quadráticas internas (inércia) para cada um. O gráfico gerado indicou claramente um ponto de "cotovelo" em torno de k = 3, mostrando que adicionar mais grupos além disso proporcionaria ganhos mínimos.

Para complementar, avaliamos também o coeficiente de silhueta, que mede a separação entre clusters (de -1 a 1). Para k=3, obtivemos silhueta média ~0,06; para k=4, foi ~0,05, ambos indicando sobreposição moderada. Optamos por k=3 pela leve superioridade numérica e melhor interpretabilidade (menos grupos para analisar). O K-Means final foi inicializado com método k-means++ e random_state=42 para garantir reprodutibilidade.

Aplicação do K-Means:

Com k definido, rodamos o algoritmo K-Means (via sklearn.cluster.KMeans) nos dados padronizados. Obtivemos os rótulos de cluster para cada um dos 100 estudantes, atribuindo um cluster (0, 1 ou 2) a cada registro. Incorporamos esses rótulos ao DataFrame original para possibilitar análises pós-modelo. Calculamos então as centroides (médias) de cada variável em cada cluster, de forma a caracterizar quantitativamente os perfis. Por exemplo, extraímos a média de horas de estudo, média de presença, média de nota, etc., em cada grupo. Esses valores formaram um “resumo do cluster”, isto é, uma espécie de perfil numérico médio de cada grupo. Também contamos a quantidade de alunos em cada cluster (tamanhos dos grupos). Essas informações serviram de base tanto para interpretação humana quanto para alimentar a etapa de geração de texto. Durante esta etapa, monitoramos possíveis casos de clusters vazios (não ocorreu) e verificamos a convergência do algoritmo (atingiu o critério de parada padrão em poucos iteradores dado o tamanho modesto do dataset).

Validação dos Clusters:

Além da análise inicial com o método do cotovelo e o coeficiente de silhueta (~0,06), realizamos verificações adicionais para validar os clusters. Utilizamos gráficos de dispersão (scatter plots) com variáveis-chave, como presença versus nota média, destacando visualmente que cada cluster tendia a ocupar regiões diferentes, ainda que com algumas sobreposições. Embora a silhueta relativamente baixa sugira fronteiras menos definidas, os clusters foram considerados válidos devido às claras diferenças nas variáveis centrais. Avaliamos também a inércia intra-cluster (soma de erros quadráticos), verificando que um número maior de clusters (k=4) não trazia melhora substancial. Dada a natureza não supervisionada do método, essas avaliações conjuntas (elbow, silhueta, gráficos visuais) proporcionaram confiança suficiente na escolha de três clusters.

Integração com API Generativa (Gemini):

Com os clusters definidos e perfis numéricos consolidados, utilizamos a API generativa Gemini (similar ao ChatGPT) para criar automaticamente descrições narrativas dos grupos. Para isso, elaboramos prompts claros e objetivos em inglês, contendo as médias das principais métricas de cada cluster (presença média, horas de estudo, notas e características marcantes), solicitando ao modelo descrições compreensíveis e livres de viés. As requisições foram feitas via API utilizando Python, e os resultados obtidos mostraram-se coerentes e detalhados, confirmando o potencial dos modelos generativos para interpretar e descrever perfis estudantis com eficácia. Essa técnica de "prompt engineering" permitiu aproveitar ao máximo o conhecimento prévio do modelo sobre contextos educacionais, resultando em textos claros, precisos e úteis para o relatório.

Revisão e Pós-Processamento das Descrições:

As descrições geradas pela IA foram lidas e, se necessário, ajustadas manualmente para corrigir eventuais imprecisões factuais. Por exemplo, se o modelo mencionasse algo que não estivesse suportado pelos dados (uma possibilidade, já que modelos podem “alucinar”), nós removeríamos ou corrigiríamos essa parte antes de finalizá-la. No entanto, de modo geral, as respostas foram pertinentes. Também padronizamos o tamanho dos textos – cerca de um parágrafo por cluster – para manter consistência. Essa revisão humana final garante a acurácia e a ética das descrições antes de incorporá-las ao relatório.

Concluídas essas etapas, passamos a interpretar os resultados combinados do clustering e das descrições, conforme descrito a seguir.

Resultados Obtidos:

Clusters Identificados: O algoritmo K-Means, configurado para k=3, gerou três clusters distintos de estudantes. A distribuição de alunos por cluster ficou assim: Cluster 0: 28 alunos; Cluster 1: 35 alunos; Cluster 2: 37 alunos (totalizando 100). Cada cluster representa um perfil dominante que emergiu dos dados. As principais características numéricas de cada grupo são apresentadas na Tabela 1 a seguir:

Variável	Cluster 0 (Perfil A)	Cluster 1 (Perfil B)	Cluster C (Perfil C)
Tamanho do Cluster (n)	28	35	37
Idade (anos)	22.9	19.4	20.8
Presença média (%)	70.6	76.9	79.2
Horas de Estudo Semanais	26.8	26.4	28.3
Nota média (0-10)	5.5	6.8	7.4
Nota Prova Midterm)(0-10)	5.2	6.5	7.0
Nota Prova Final (0-10)	5.8	7.0	7.8
Cursos Online Concluidos	1.1	2.3	3.7

Número médio de módulos de cursos online extracurriculares concluídos (escala de 0 a 5). Analisando os valores médios acima, podemos interpretar cada cluster da seguinte forma:

Cluster 0 – “Perfil A (Baixo Engajamento): Este grupo apresentou as menores taxas de presença (cerca de 70% em média, significativamente abaixo dos outros) e as piores notas médias (em torno de 5.5/10). São estudantes ligeiramente mais velhos (idade média ~23 anos) e que reportaram um tempo de estudo semanal moderado (~27h) – similar a outros clusters – mas aparentemente com menor eficiência ou consistência, dado seu desempenho inferior. Eles também completaram poucos módulos de cursos online extra (média ~1, o menor dentre os grupos). Esse perfil sugere alunos possivelmente com dificuldades de acompanhar o curso ou com outras responsabilidades competindo pelo seu tempo (por exemplo, muitos podem trabalhar, o que explicaria a idade maior e a menor dedicação/assiduidade). O cluster 0 pode ser caracterizado como “alunos de engajamento insuficiente”, necessitando de atenção especial

dos educadores. De fato, muitos integrantes desse cluster acabaram reprovando em avaliações ou ficando próximos da nota mínima para aprovação.

Cluster 1 – “Perfil B (Médio e Iniciantes): O segundo grupo teve desempenho e engajamento medianos. A presença média (~77%) e a nota média (~6.8) indicam alunos que cumprem a maior parte das obrigações, com resultados acadêmicos aceitáveis, embora não excelentes. Notavelmente, este cluster concentra os alunos mais jovens (média ~19 anos, possivelmente ingressantes no curso). Eles investem praticamente o mesmo tempo de estudo que os demais (~26 horas semanais), mas completaram mais cursos online extras que o Cluster 0 (média 2.3 módulos). Podemos supor que são estudantes ainda se adaptando ao ambiente universitário – demonstram interesse (participam de algumas atividades extra, mantêm presença relativamente boa), porém podem carecer de técnicas de estudo mais eficientes ou experiência para alavancar suas notas. Chamamos este perfil de “alunos medianos em desenvolvimento”. Espera-se que com orientação e amadurecimento acadêmico, muitos nesse grupo possam migrar para um perfil de alto desempenho. Eles formam um grupo intermédio importante para monitoramento: nem tão alarmante quanto o Cluster 0, mas com espaço para melhoria em desempenho.

Cluster 2 – “Perfil C (Altamente Engajado): O terceiro cluster destaca-se positivamente. Estes alunos obtiveram as melhores notas médias (~7.4/10) e quase 80% de presença em aulas (maior dentre os clusters). Também dedicam um pouco mais de horas aos estudos semanais (~28h) e completaram a maior quantidade de módulos online extras (média 3.7, indicando forte interesse em aprendizado para além do currículo formal). A idade média deles (~21 anos) fica entre os outros clusters. Esse perfil sugere estudantes altamente engajados e proativos, que aproveitam recursos adicionais (cursos online) e mantêm boa disciplina (frequência alta), refletindo em desempenho superior. Provavelmente são alunos organizados e motivados, possivelmente em fases mais avançadas do curso, com objetivos claros (como preparação para pós-graduação ou mercado de trabalho competitivo). Podemos denominar este grupo de “alunos dedicados de alto desempenho”. Eles servem até como referência de boas práticas para os demais.

Validação dos Perfis:

Os clusters apresentaram certo grau de sobreposição, o que resultou em um coeficiente de silhueta relativamente baixo (~0,06). Por exemplo, alguns estudantes do Cluster 1 têm notas próximas ao Cluster 2 ou frequência semelhante ao Cluster 0. Apesar disso, uma análise ANOVA revelou diferenças significativas ($p < 0,05$) para as variáveis nota final e presença, confirmando que os grupos capturaram padrões relevantes e não são fruto do acaso.

Geração de Descrições por IA:

Utilizamos um modelo generativo de linguagem (API Gemini) para criar descrições automáticas e detalhadas de cada perfil, com base nas médias dos indicadores. Os resultados foram bastante precisos e elucidativos, destacando corretamente as características dominantes dos clusters:

- **Cluster 0 (Engajamento Baixo):** Estudantes com baixa presença em aulas e notas abaixo da média, apesar de um esforço moderado nos estudos. A IA sugeriu dificuldades organizacionais ou responsabilidades extras, recomendando intervenções pedagógicas direcionadas, como mentorias ou reforço acadêmico.
- **Cluster 1 (Médio em Desenvolvimento):** Alunos iniciantes que mantêm presença razoável e desempenho satisfatório, ainda que não excepcional. Demonstram interesse em atividades extracurriculares e potencial de evolução, destacando a importância de orientação específica, técnicas eficazes de estudo e participação em grupos colaborativos.
- **Cluster 2 (Alto Desempenho):** Grupo altamente engajado, com presença elevada, notas altas e forte dedicação a estudos complementares. A IA ressaltou corretamente disciplina, motivação e capacidade de liderança em atividades acadêmicas, sugerindo desafios adicionais, monitorias ou iniciação científica como ações ideais.

Essas descrições foram originalmente geradas em inglês e revisadas cuidadosamente após tradução ao português, garantindo coerência, precisão e respeito às informações do dataset.

Interpretação Integrada:

Com base nos resultados numéricos e nas descrições da IA, foi possível delinear estratégias específicas para cada cluster. O Cluster 0 demanda suporte pedagógico extra, como mentorias acadêmicas e aconselhamento para gestão do tempo. Já o Cluster 1 pode ser beneficiado por abordagens que aprimorem suas habilidades de estudo e integração acadêmica.

O Cluster 2, por sua vez, deve ser estimulado por meio de desafios avançados e oportunidades para exercer liderança e apoiar colegas. Essas recomendações alinham-se plenamente ao objetivo principal do projeto: identificar perfis distintos e oferecer descrições compreensíveis que auxiliem intervenções pedagógicas personalizadas.

Conclusões:

Este projeto atingiu seus objetivos principais ao identificar três perfis distintos de estudantes — "Baixo Engajamento", "Médio em Desenvolvimento" e "Alto Desempenho" — por meio de **K-Means** aplicado a métricas como presença, horas de estudo e notas. Apesar da sobreposição moderada entre grupos (silhueta ≈0,06), as diferenças nas médias de indicadores-chave foram estatisticamente significativas, validando a segmentação.

A integração com a **IA generativa** (API Gemini) permitiu transformar essas médias em descrições claras e úteis para educadores, automatizando a interpretação e ampliando o impacto dos resultados. As narrativas geradas capturaram corretamente as características de cada cluster, comprovando a viabilidade do uso de modelos de linguagem para rotular automaticamente perfis de alunos.

Reconhecem-se limitações na homogeneidade dos dados e na clareza das fronteiras dos clusters, sugerindo que futuras investigações considerem mais variáveis ou algoritmos alternativos (hierárquicos, DBSCAN) para aprimorar a segmentação. Além disso, recomenda-se manter a revisão humana das descrições geradas para evitar imprecisões.

Em síntese, a combinação de análise exploratória, clustering e IA generativa mostrou-se eficaz para mapear perfis estudantis e facilitar intervenções pedagógicas direcionadas, abrindo caminho para dashboards interativos e modelos preditivos que apoiem ainda mais a personalização do ensino.

Endereço do GitHub e do Vídeo no YouTube

Repositório GitHub: https://github.com/LittleLion2/Projeto-I.A.git

Vídeo de Apresentação (YouTube): https://youtu.be/wZpF9rBZ-OA

Referências:

KAGGLE. *Student Performance & Behavior Dataset*. [S.l.], 2023. Disponível em: Student Performance & Behavior Dataset. Acesso em: março 2025.

KAGGLE. *Student Performance & Learning Style Dataset*. [S.l.], 2023. Disponível em: Student Performance & Learning Style. Acesso em: março 2025.

CHAÚDHRY, Muhammad Ali; CUKUROVA, Mutlu; LUCKIN, Rose. A transparência na IA educacional: framework de índice de transparência. *ArXiv*, jun. 2022. Disponível em: https://arxiv.org/abs/2206.03220. Acesso em: 16 maio 2025.

PAULA, Alexandre Abreu de et al. Ética no uso de inteligência artificial na educação: impactos para professores e estudantes. *Revista FT*, 2023. Disponível em: https://revistaft.com.br/a-etica-no-uso-de-inteligencia-artificial-na-educacao-impactos-para-professores-e-estudantes Acesso em: 18 maio 2025.

COMISSÃO EUROPEIA. Diretrizes éticas sobre o uso de inteligência artificial (IA) e de dados no ensino e na aprendizagem. Luxemburgo: Serviço das Publicações da União Europeia, 2022. Disponível em: https://op.europa.eu/pt/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1. Acesso em: 21 maio 2025.