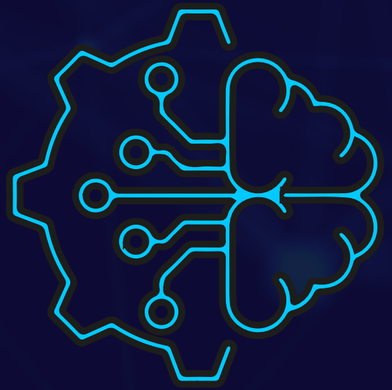


Inteligência Artificial

Classificação do Ensino Médio usando Dados do INEP

Felipe da Costa Coqueiro, NUSP: 11781361
Fernando Alee Suaiden, NUSP: 12680836
Laura Camargo, NUSP: 13692334
Adhemar Molon Neto, NUSP:14687681
Allan Garcia Silva, NUSP: 13731222

ICMC USP SÃO CARLOS

Introdução

Contexto

- Desigualdade educacional no Brasil e importância do ENEM como indicador.
- Infraestrutura escolar (laboratórios, biblioteca, internet, saneamento, energia) como possível fator de desempenho.

Problema Central

A infraestrutura física das escolas influencia o desempenho médio no ENEM?

Abordagem

Uso de **mineração de dados** e **KDD** para analisar milhares de escolas da base do INEP (Censo Escolar + ENEM 2024)

INEP enem

Objetivos

Objetivo Central

- **Classificar as escolas em grupos** (clusters) com características semelhantes e **identificar se instalações** como laboratórios, bibliotecas e saneamento **são fatores determinantes para notas mais altas**.

Objetivos Específicos

- Integrar e tratar bases de dados do INEP
- Aplicar técnicas de clusterização (K-Means) para classificar as escolas em grupos semelhantes.
- Utilizar Árvore de Decisão (XAI) para explicar as regras que distinguem os clusters.
- Analisar o perfil social dos clusters segundo o tipo de dependência administrativa (pública vs. privada).

Metodologia

22k escolas analisadas

1. **Seleção e Coleta de Dados**

Censo Escolar 2024
Microdados ENEM 2024

2. **Pré-processamento e Transformação**

Limpeza, engenharia de atributos, normalização

3. **Mineração de Dados**

Clusterização (K-Means)

Perfilamento dos clusters e análise social

4. **Pós-processamento e Interpretação**

Metodologia baseada em KDD - Knowledge Discovery in Databases

Metodologia

Dados e Pré-processamento

Integração das Bases:

- Chaves de ligação
  - Censo Escolar: CO_ENTIDADE
  - ENEM: CO_ESCOLA
- Inner join: apenas escolas presentes em ambas as bases.

Engenharia de Atributos:

- SCORE_INFRA: Soma de variáveis binárias de infraestrutura
  - Água potável, energia, esgoto, biblioteca, laboratórios, internet, quadra esportiva etc.
- NOTA_GERAL: Média das notas
  - NU_NOTA_CN, NU_NOTA_CH, NU_NOTA_LC, NU_NOTA_MT, NU_NOTA_REDACAO.

Filtro de Relevância:

- Remoção de escolas com menos de 10 alunos participantes no ENEM (QTD_ALUNOS ≥ 10)

Metodologia

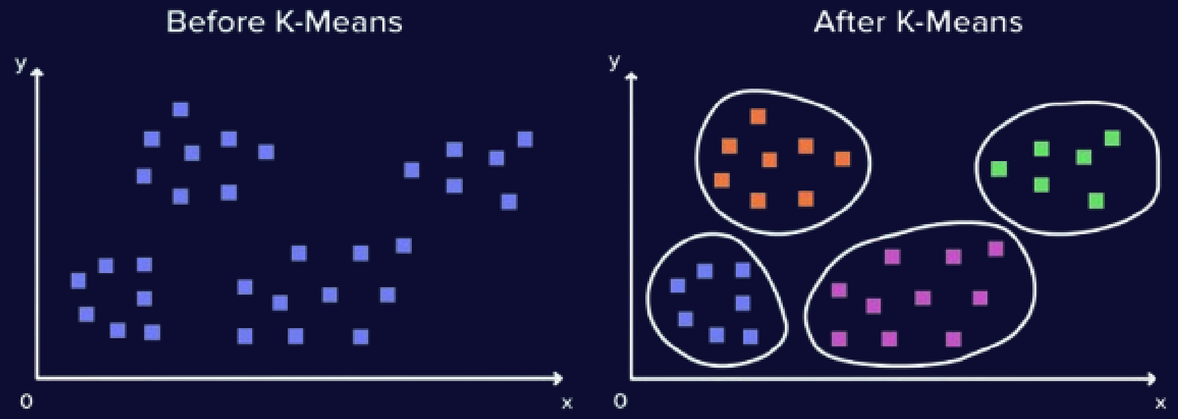
Normalização e Clusterização

Normalização

- Uso do RobustScaler:
  - Baseado em mediana e intervalo interquartil (IQR).
  - Menos sensível a outliers do que StandardScaler ou MinMaxScaler.

Definição do Número de Clusters

- Aplicação do Método do Cotovelo (Elbow Method):
  - Teste de K de 1 a 10.
  - Escolha de K = 4 como ponto de inflexão da curva de inércia.



K-Means (K = 4)

- Entradas do modelo:
  - SCORE_INFRA e NOTA_GERAL normalizadas.
- Saída:
  - Agrupamento das escolas em 4 clusters com perfis distintos.

Metodologia

Explicabilidade (XAI) e Análise Social

Árvore de Decisão (XAI):

- Treinada com:
  - Features: SCORE_INFRA e NOTA_GERAL (sem normalização)
  - Target: CLUSTER
- Gera regras do tipo:
  - Se INFRA > X e NOTA > Y → Cluster Z

- **Investigação Social:**
  - Mapeamento da variável TP_DEPENDENCIA em:
    - Federal, Estadual, Municipal e Privada.
  - Criação de tabela de contingência (crosstab) para ver:
    - Percentual de cada tipo de escola em cada cluster.
  - Gráfico de barras empilhadas mostrando a composição social dos grupos.

Resultados

Visão Geral dos Clusters

Cluster 0 - Infraestrutura Alta, Notas Medianas

- Composição:
  - Cerca de 93% de escolas Estaduais/Federais.
- Perfil:
  - Maior infraestrutura média de toda a amostra (SCORE_INFRA ≈ 7,44, superior ao cluster privado).
  - Desempenho médio inferior ao da elite privada (NOTA_GERAL ≈ 502,7).

Interpretação:

- Ter infraestrutura física boa não é suficiente para atingir excelência. Indica influência de fatores como: Contexto socioeconômico dos alunos, gestão escolar e políticas pedagógicas e condições de trabalho docente.

Cluster 1 - "Elite Acadêmica"

- Composição:
  - Aproximadamente 81% de escolas privadas.
- Perfil:
  - Alta infraestrutura (SCORE_INFRA ≈ 7,11).
  - Maior desempenho médio no ENEM (NOTA_GERAL ≈ 614,6).

Interpretação:

- Modelo típico de escolas particulares bem equipadas. Confirma o domínio da rede privada no topo do ranking do ENEM.

Cluster 3 - Vulnerabilidade Máxima

- Representa aproximadamente 10% da amostra.
- Composição:
  - Cerca de 98% de escolas públicas (Estaduais/Municipais).
- Perfil:
  - Infraestrutura crítica (SCORE_INFRA ≈ 4,0).
  - Piores notas no ENEM (NOTA_GERAL ≈ 459).

Interpretação:

- Falta de infraestrutura básica atua como barreira estrutural ao aprendizado. Grupo que mais demanda investimento urgente em estrutura e apoio pedagógico.

Resultados

Normalização

--- Estatísticas Originais ---

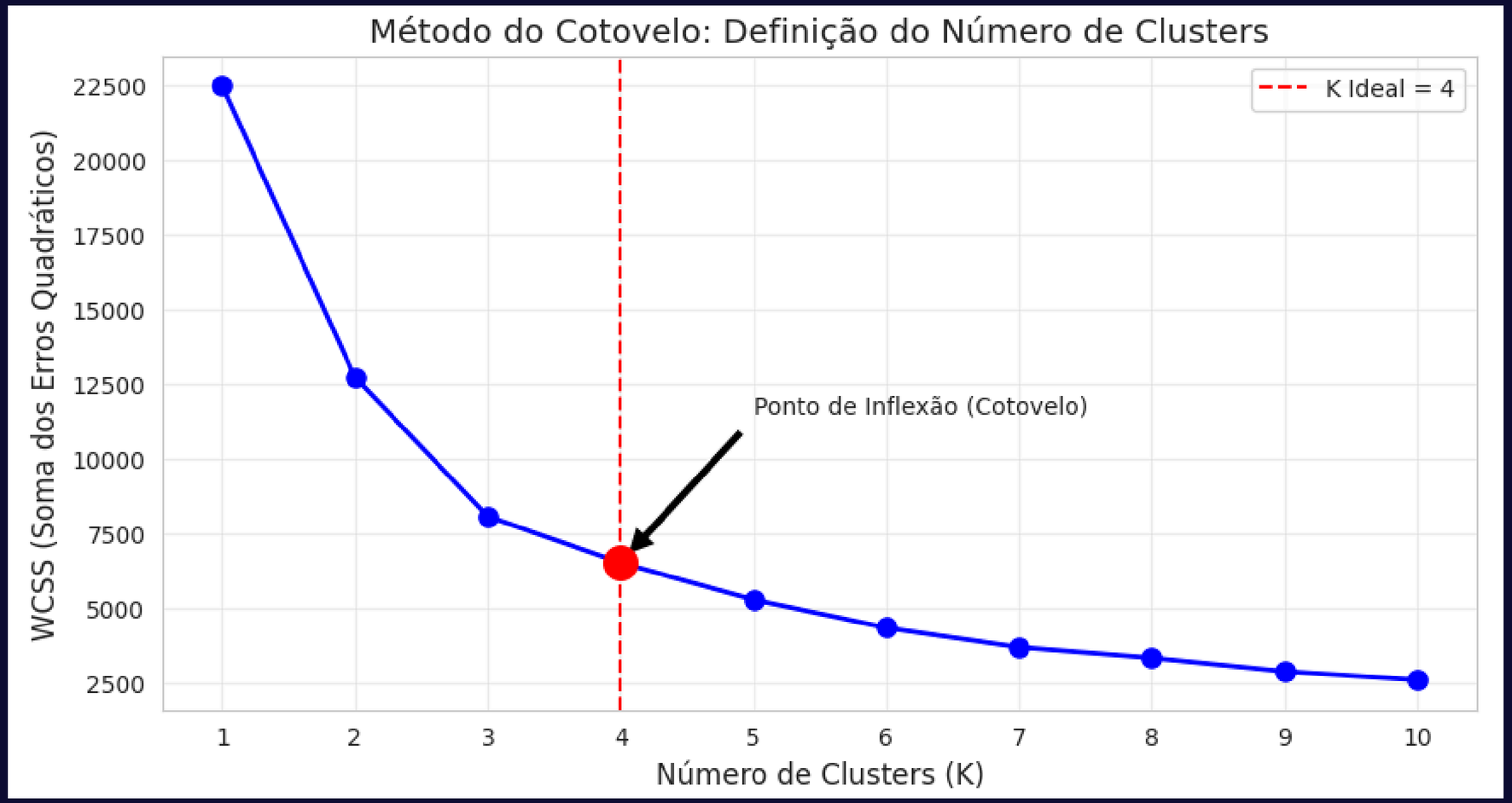
	SCORE_INFRA	NOTA_GERAL
count	22210.00	22195.00
mean	6.54	525.85
std	1.29	61.56
min	0.00	301.09
25%	6.00	482.53
50%	7.00	513.29
75%	8.00	562.23
max	8.00	787.12

--- Dados Normalizados (RobustScaler) ---

	SCORE_INFRA	NOTA_GERAL
0	-0.5	-0.394748
1	0.0	0.447264
2	0.0	-0.135994
3	-0.5	-0.715118
4	0.0	-0.386174

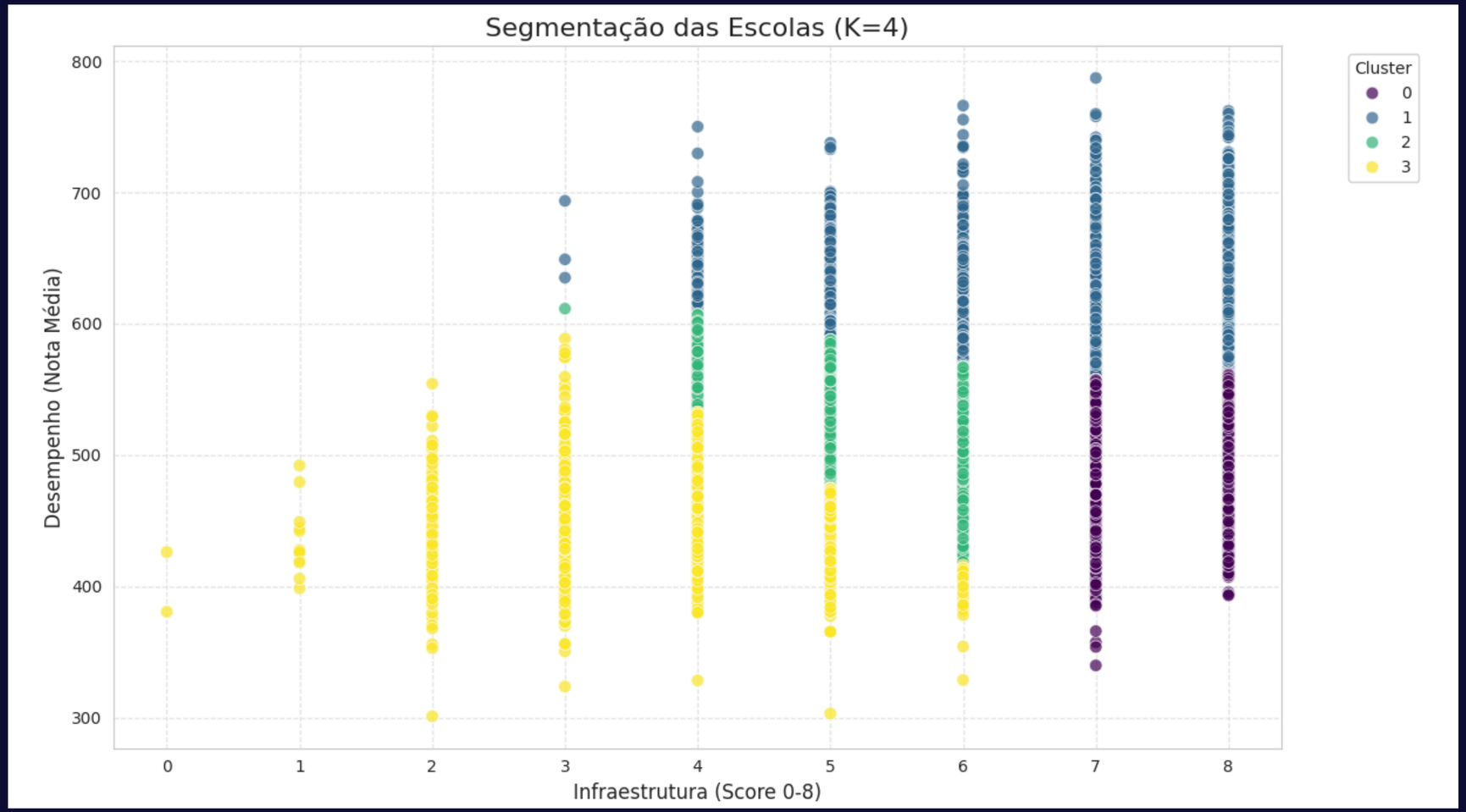
Resultados

Clusterização



Resultados

Classificação



Resultados

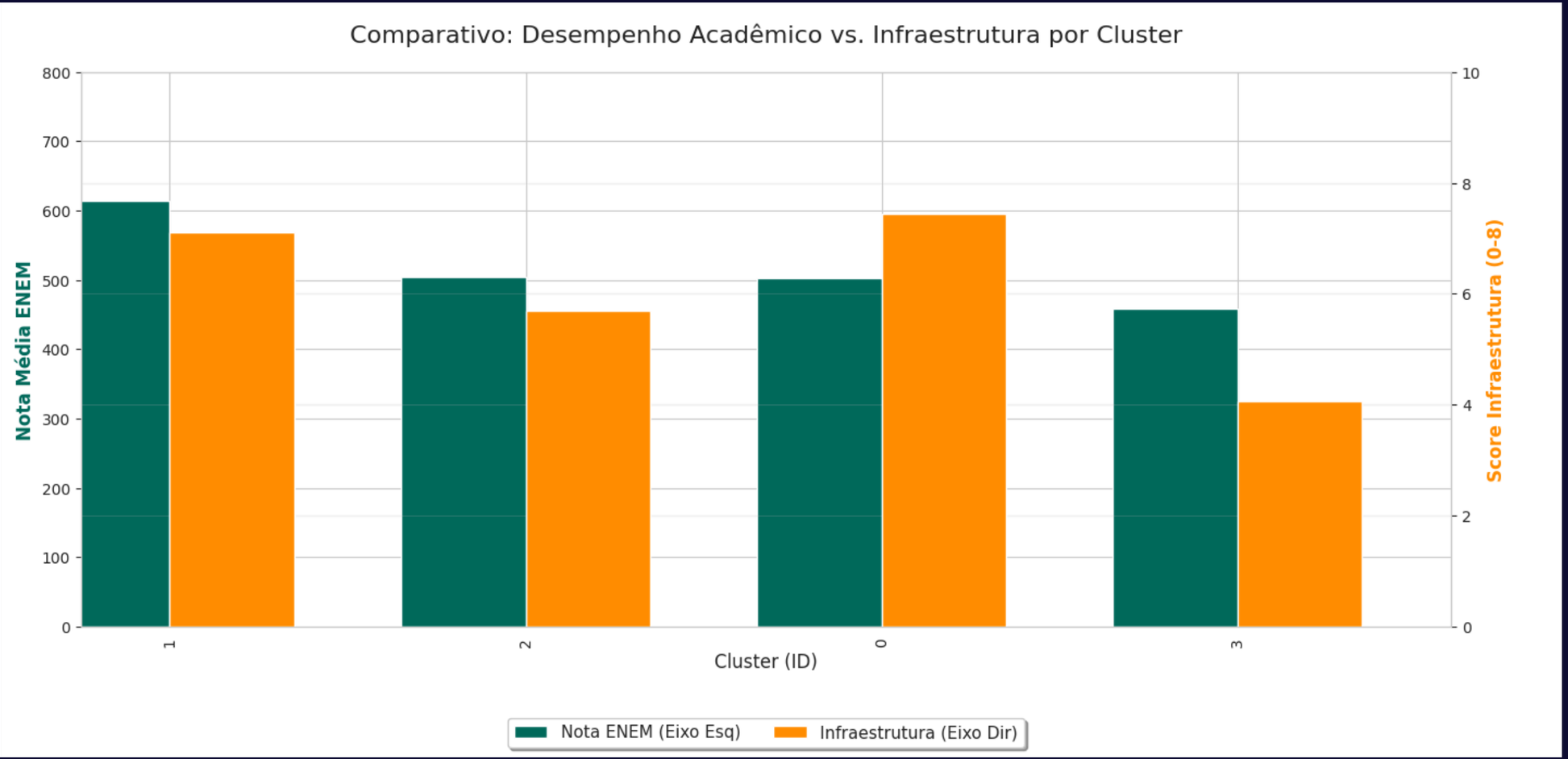
Perfilamento

```
--- INICIANDO PERFILAMENTO E VISUALIZAÇÃO ---
Tabela de Perfil dos Clusters:
```

CLUSTER	SCORE_INFRA	NOTA_GERAL	QTD_ALUNOS	QTD_ESCOLAS	%_TOTAL
1	7.11	614.61	37.86	5379	24.24
2	5.70	503.85	39.05	6187	27.88
0	7.44	502.71	51.41	8441	38.03
3	4.06	459.15	40.41	2188	9.86

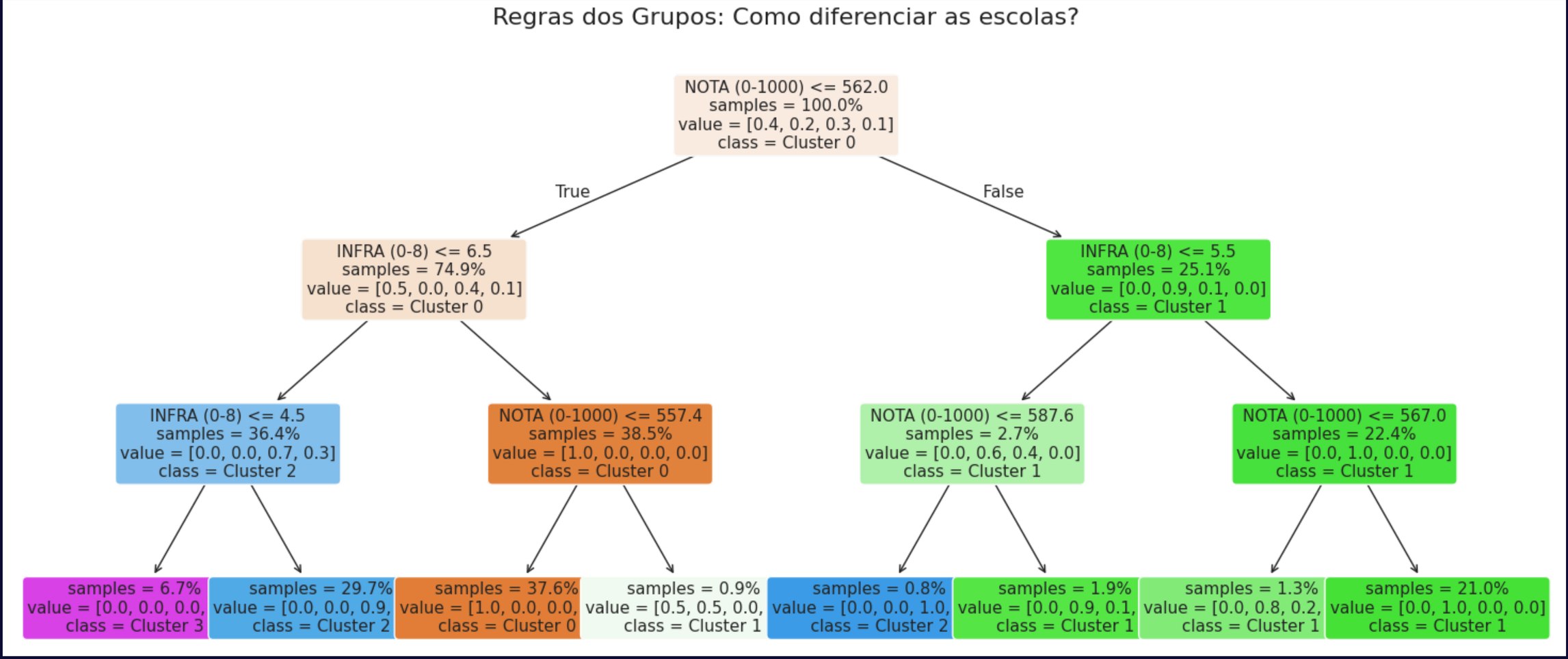
Resultados

Visualização Comparativa



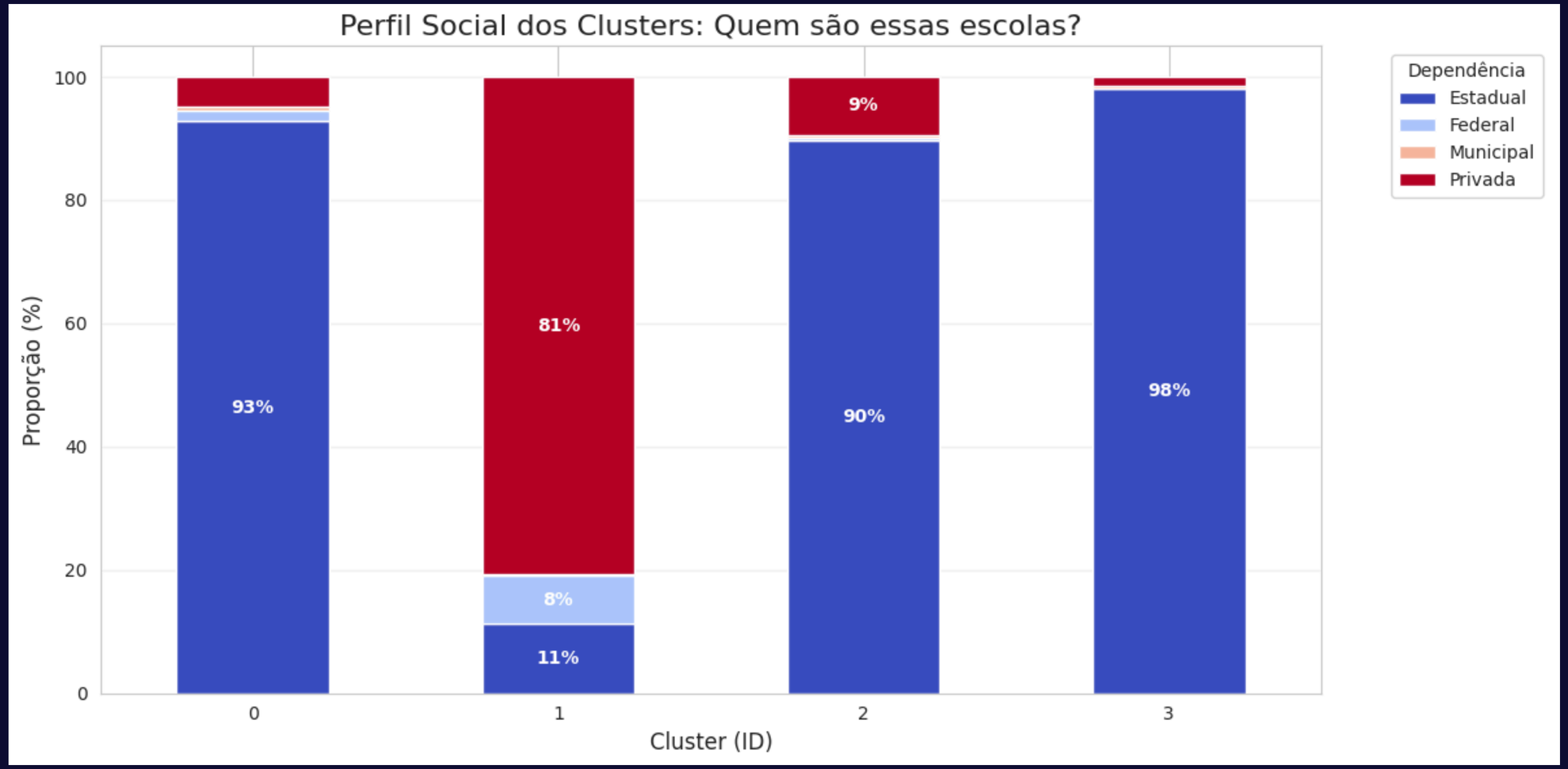
Resultados

Visualização das Regras



Resultados

Visualização dos Perfis



Conclusão

Principais Conclusões

1. Correlação positiva entre infraestrutura e desempenho:
   - Escolas com infraestrutura muito precária tendem a ter notas mais baixas (caso do Cluster 3).
2. Correlação não é linear nem suficiente:
   - Há escolas públicas bem equipadas (Cluster 0) que não atingem o desempenho da elite privada (Cluster 1).
3. Elite Acadêmica concentrada na rede privada:
   - Infraestrutura forte + condições pedagógicas e de gestão favorecem resultados altos (Cluster 1).

Veredito Geral

1. Infraestrutura é **condição necessária**, mas **não suficiente** para excelência.
2. Políticas públicas devem combinar:
   - Investimento físico (laboratórios, bibliotecas, internet).
   - Ações pedagógicas, formação docente e gestão escolar.

Limitações

Insights e Trabalhos Futuros

1. Implicações para Políticas Públicas

- Focar não apenas em obras, mas também em:
  - Gestão pedagógica.
  - Formação continuada.
  - Apoio às escolas em contexto de vulnerabilidade.

2. Possíveis Extensões do Estudo

- Incluir variáveis socioeconômicas de contexto (IDH municipal, renda, etc.).
- Analisar evolução temporal (séries históricas do ENEM e Censo).
- Ampliar o modelo:
  - Outros algoritmos de clusterização.
  - Modelos preditivos de nota por escola.

Muito obrigado!
Alguma pergunta?