TECHNICAL REPORT

Aluno: Samuel Henrique da Silva

1. Introdução

O dataset utilizado neste projeto é o **StudentsPerformance.csv**, que contém dados educacionais de estudantes, incluindo suas notas em matemática, leitura e escrita, além de atributos demográficos como gênero, nível educacional dos pais, entre outros. O objetivo principal da análise é explorar os dados, aplicar técnicas de clusterização (agrupamento não supervisionado), reduzir a dimensionalidade dos dados e realizar interpretação semântica dos grupos encontrados. Além disso, foi feita uma avaliação do impacto da redução de dimensionalidade na performance de modelos de classificação.

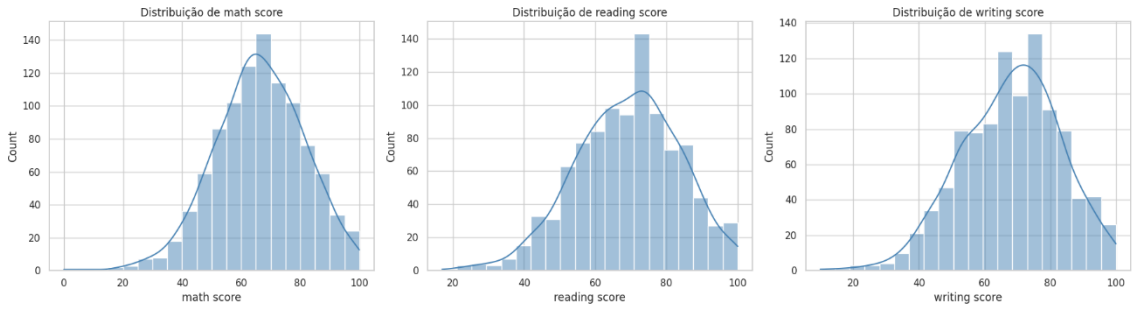
2. Observações

Em questões de observações não encontrei nenhuns imprevistos técnicos durante a execução dos scripts. Todos os pacotes necessários foram carregados corretamente, e os dados estavam limpos e prontos para uso, o que facilitou o processo de análise.

3. Resultados e discussão

1. Análise Exploratória com Foco em Redução de Complexidade

A análise exploratória foi conduzida para entender a distribuição das notas e identificar possíveis padrões. Foram gerados histogramas e boxplots para visualizar as distribuições e outliers em cada uma das três notas principais (matemática, leitura e escrita). Também foi gerado um mapa de calor de correlação entre as notas.



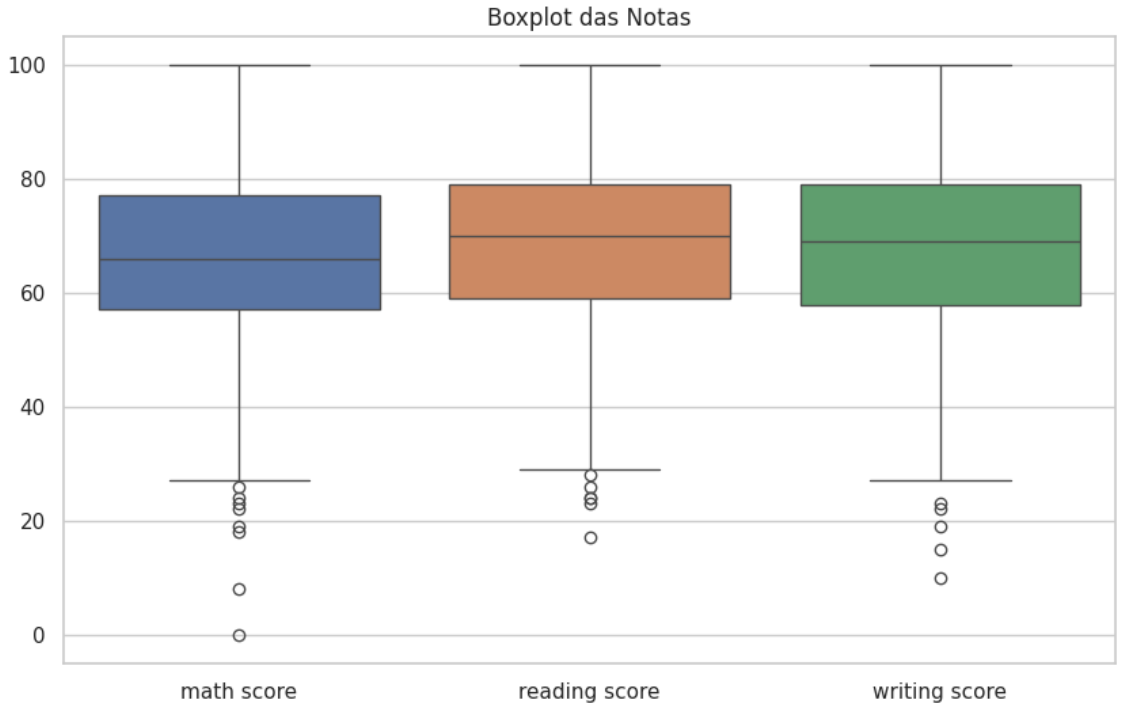
Aqui são histogramas para compreensão melhor dos dados:

- Das as distribuições seguem um **padrão aproximadamente normal**, com leve assimetria à esquerda (cauda mais longa em notas mais baixas).

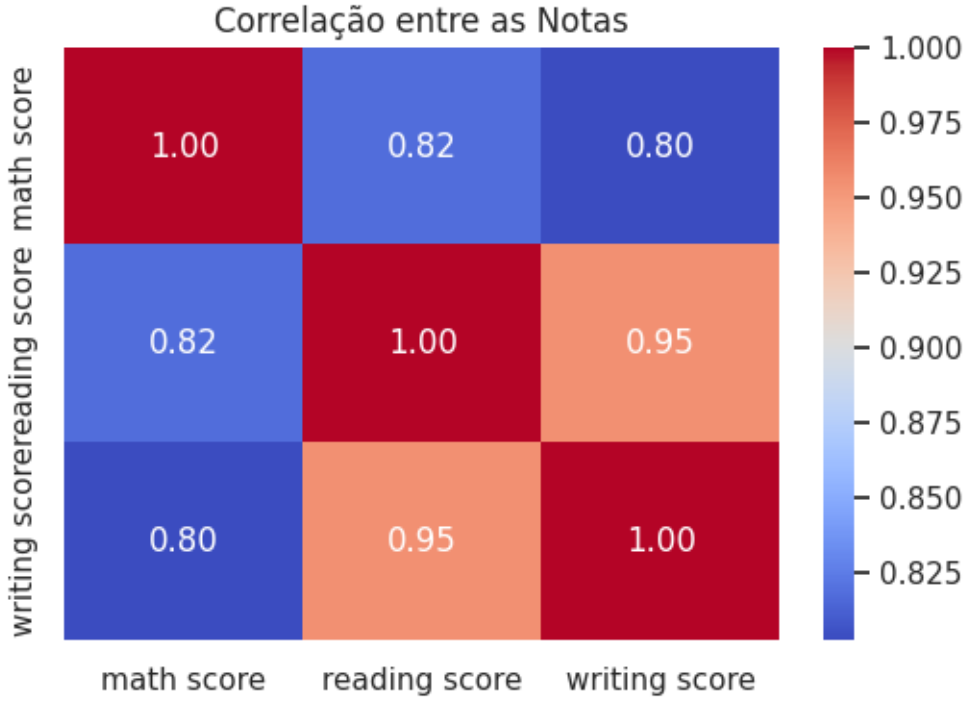
- A maior concentração de alunos apresenta notas na faixa de **60 a 80 pontos**, sugerindo uma tendência central bem definida.

- Poucos alunos têm notas extremamente baixas (menores que 30) ou extremamente altas (maiores que 95), indicando **baixo impacto de outliers**.

- A escala de todas as variáveis é semelhante (0 a 100), porém as variâncias apresentam pequenas diferenças.



- *Aqui são bloxplots das notas de matemática, leitura e escrita*
- *É possível também visualizar que existem diversos outliers nas notas, porém prevalece outliers das notas de matemática.*



O mapa de calor apresenta a correlação entre as três variáveis numéricas: **math score**, **reading score** e **writing score**. Observa-se que:

- A correlação entre **reading score** e **writing score** é **muito alta (0,95)**, indicando forte redundância entre essas duas variáveis.
- A correlação entre **math score** e as demais variáveis também é alta, sendo **0,82 com reading score** e **0,80 com writing score**.
- Não há correlação negativa entre as variáveis, sugerindo que, em geral, quando uma nota aumenta, as demais também tendem a aumentar.

2. Redução de Dimensionalidade Múltipla: PCA vs T-SNE

Antes de fazer a comparação do PCA vs T-SEM, vou listar o que foi feito antes:
 Preparação dos dados:

- Selecionar apenas as variáveis numéricas (math score, reading score, writing score).
- Aplicar **padronização** (StandardScaler).

 Aplicação do PCA:

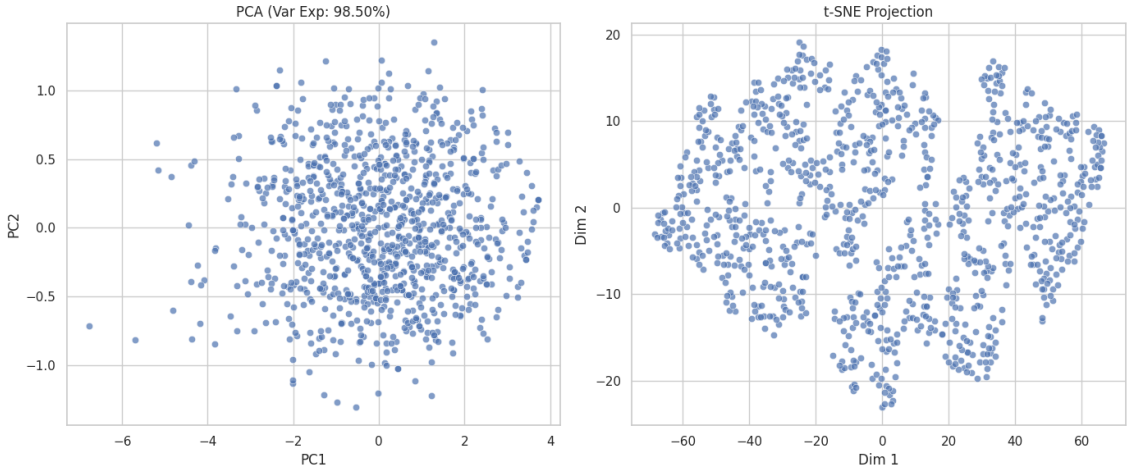
- Calcular variância explicada.
- Gerar projeção em 2D.
- Interpretar preservação da variância.

 Aplicação do T-SNE:

- Gerar projeção 2D.
- Comparar estrutura de agrupamentos com PCA.

 Comparação:

- **PCA**:
  - Preserva variância linear.
  - Baixo custo computacional.
  - Não detecta relações não lineares.
- **T-SNE**:
  - Preserva estrutura local (distâncias entre vizinhos).
  - Custo computacional maior.
  - Ótimo para visualizar clusters.



Agora a comparação do PCA VS T-SNE:

PCA:

- Variância explicada pelo 1º e 2º componente:
  - PC1: 90.6%, PC2: 7.9% → Total ≈ 98.5%
- Custo computacional: ~0.003 s (extremamente rápido)
- Visualização: Os pontos estão distribuídos em uma linha quase reta, indicando que as 3 notas são altamente correlacionadas → PCA captura isso bem.

t-SNE

- Custo computacional**: ~**57 s **(muito mais lento)**
- Visualização**: Cria uma distribuição mais espalhada, com pequenas variações, mas sem clusters evidentes (o dataset não tem grupos naturais marcantes).**
- Objetivo**: Preserva estrutura local, útil quando há clusters complexos (não é o caso aqui).**

3. Clusterização com K-Means e Hierárquico

Foi aplicado o algoritmo K-Means para agrupar os dados com base nas três notas. O número ideal de clusters foi determinado por meio dos métodos Elbow e Silhouette, que indicaram **k = 2** como a melhor opção.

Resultados:

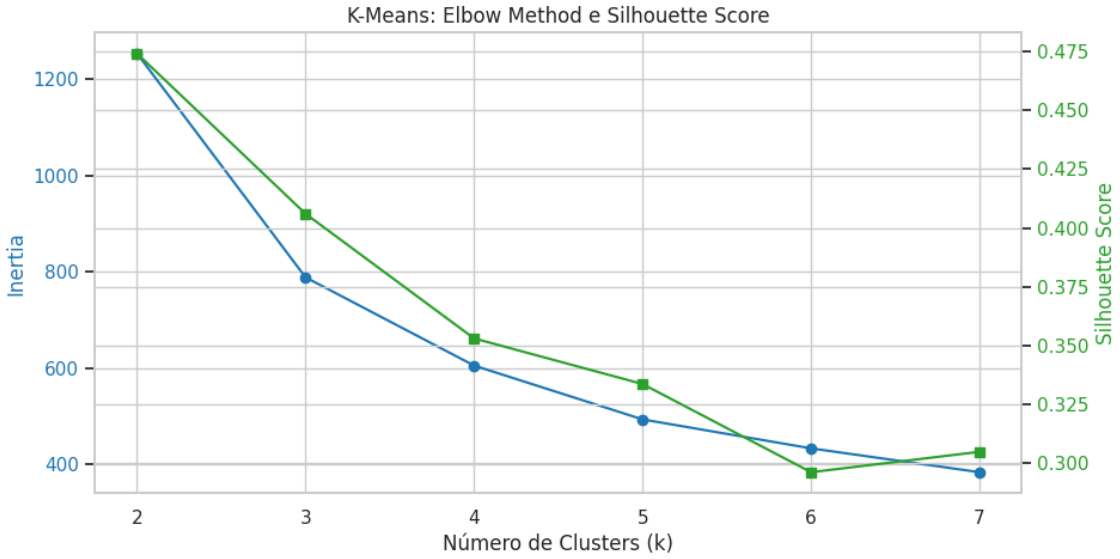
- O K-Means gerou três grupos bem definidos visualmente nas projeções PCA e t-SNE.
- A clusterização hierárquica (métodos average e complete) foi comparada com base no Silhouette Score, com o método **average** se aproximando mais do K-Means.

Silhouette Scores:

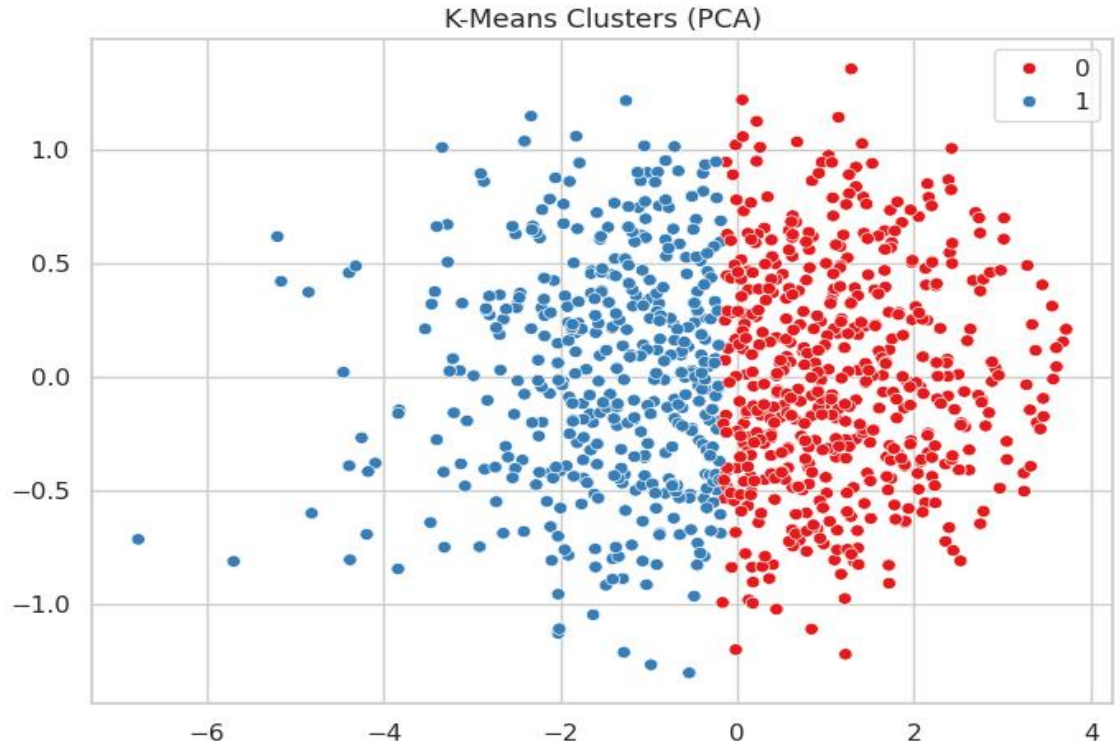
- K-Means: **0.474**
- Hierárquico (average): **0.452**
- Hierárquico (complete): **0.431**

Segue alguns do gráficos de comparação:

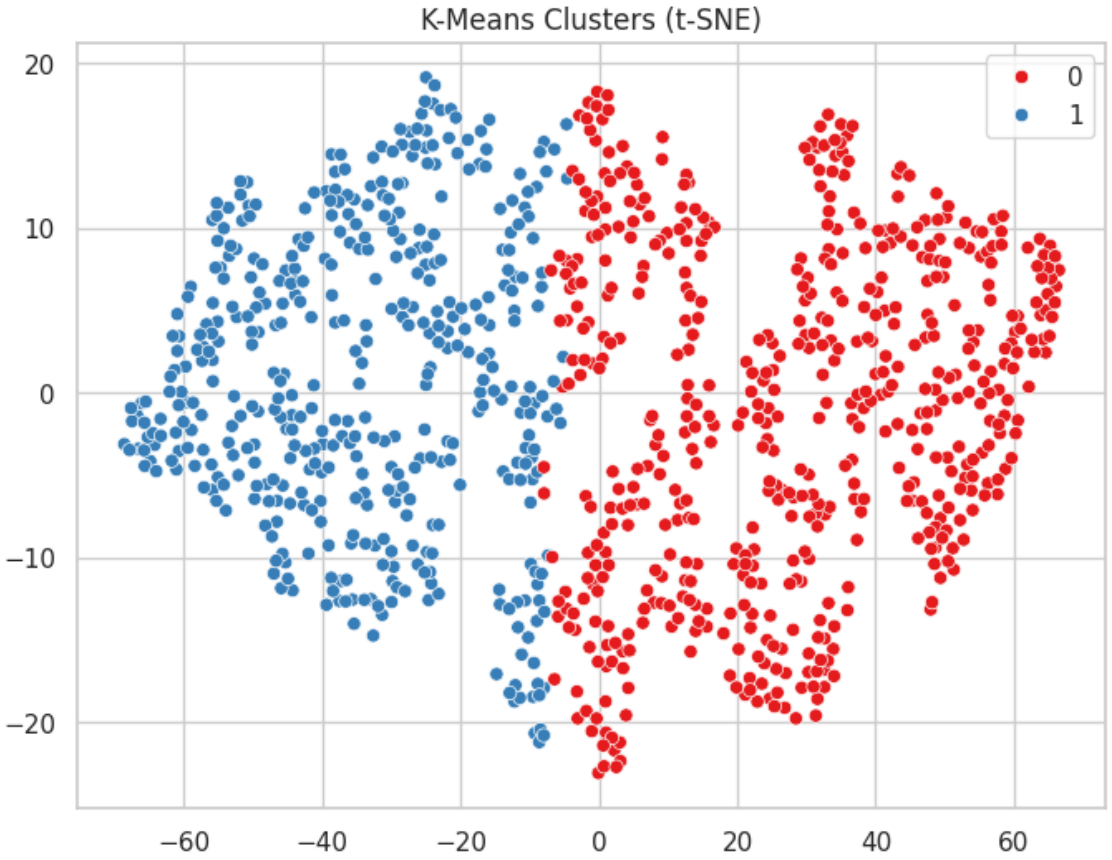
K-Means: Eibow Method e Silhouette Score



K-Means Clusters (PCA)



K-Means Clusters (t-SNE)



4. Interpretação Semântica dos Agrupamentos

Em questão, aqui temos o resultado das Médias por cluster e vou comentar seus resultados:

Cluster	Média Math	Média Reading	Média Writing
0	75,82	79,34	78,51
1	53,71	56,23	54,74

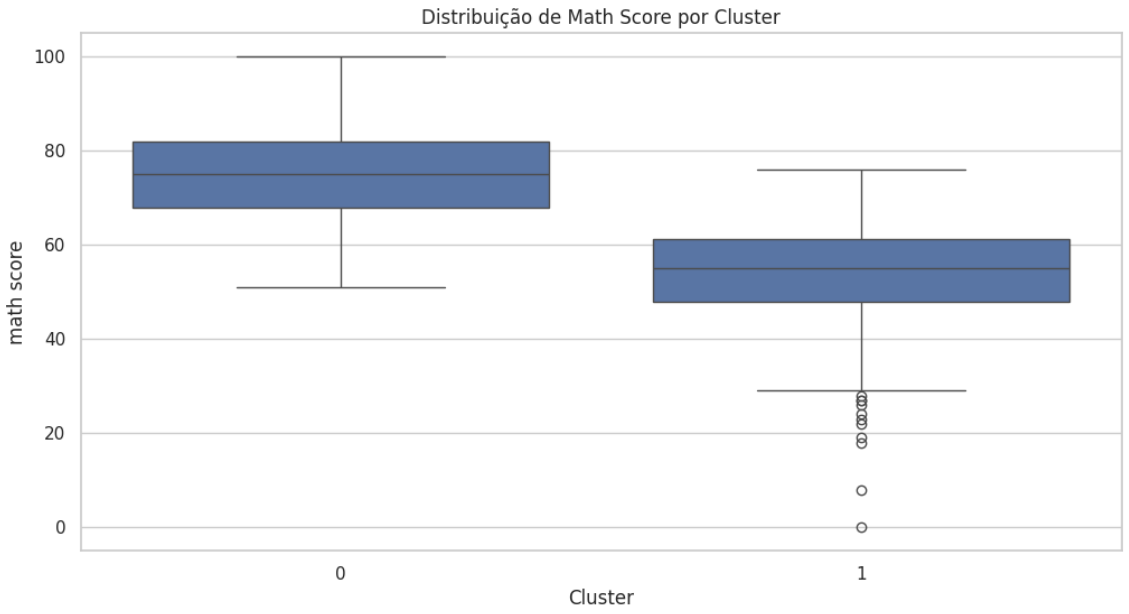
Cluster 0:

- Apresenta médias elevadas em todas as disciplinas (notas entre **75 e 80**).
- Este grupo representa os alunos com **alto desempenho geral**, tanto em matemática quanto em leitura e escrita.
- Pode ser nomeado como **“Alunos de Alto Rendimento”**

- **Cluster 1:**
    - Médias significativamente mais baixas (entre **53 e 56**).
    - Representa os alunos com **baixo desempenho geral**, apresentando dificuldades nas três áreas.
    - Pode ser nomeado como **"Alunos de Baixo Rendimento"**.

Conclusão da comparação dos clusters: Os clusters têm diferenças claras **e** consistentes nas médias, o que reforça que a segmentação faz sentido do ponto de vista pedagógico. É possível que esses grupos estejam associados a fatores socioeconômicos ou nível educacional dos pais, mas isso exigiria uma análise complementar usando as variáveis categóricas.

Distribuição de Math score por cluster:



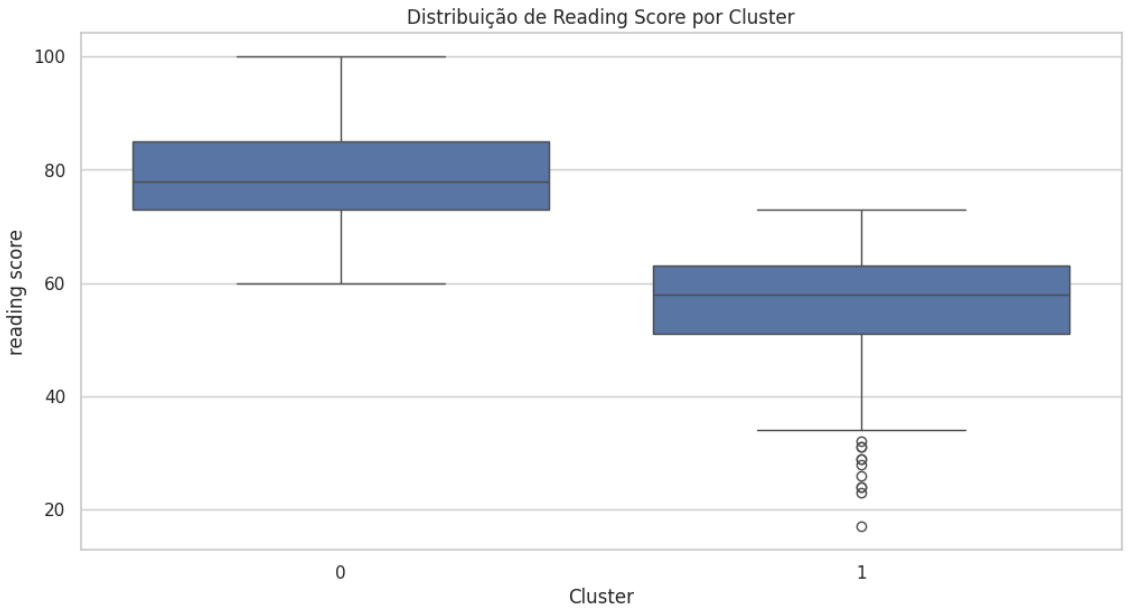
Cluster 0:

- **Média/mediana elevadas**: Os estudantes desse cluster têm, em geral, notas significativamente mais altas.
- **Distribuição mais compacta**: A maioria dos valores está entre 70 e 90.
- **Sem outliers visíveis**, indicando consistência no desempenho do grupo.
- **Interpretação**: Esse grupo representa **estudantes com desempenho acadêmico forte em matemática**.

Cluster 1:

- **Média e mediana mais baixas**: As notas estão mais concentradas na faixa entre **50 e 60**.
- **Presença de muitos outliers inferiores**: Notas próximas de zero, indicando estudantes com dificuldades significativas.
- **Amplitude maior**: O grupo é mais heterogêneo, com notas variando de 0 a 75.
- **Interpretação**: Esse cluster representa um grupo **mais vulnerável academicamente**, com grande variabilidade e desempenho mais fraco.

Distribuição de Reading score por cluster:



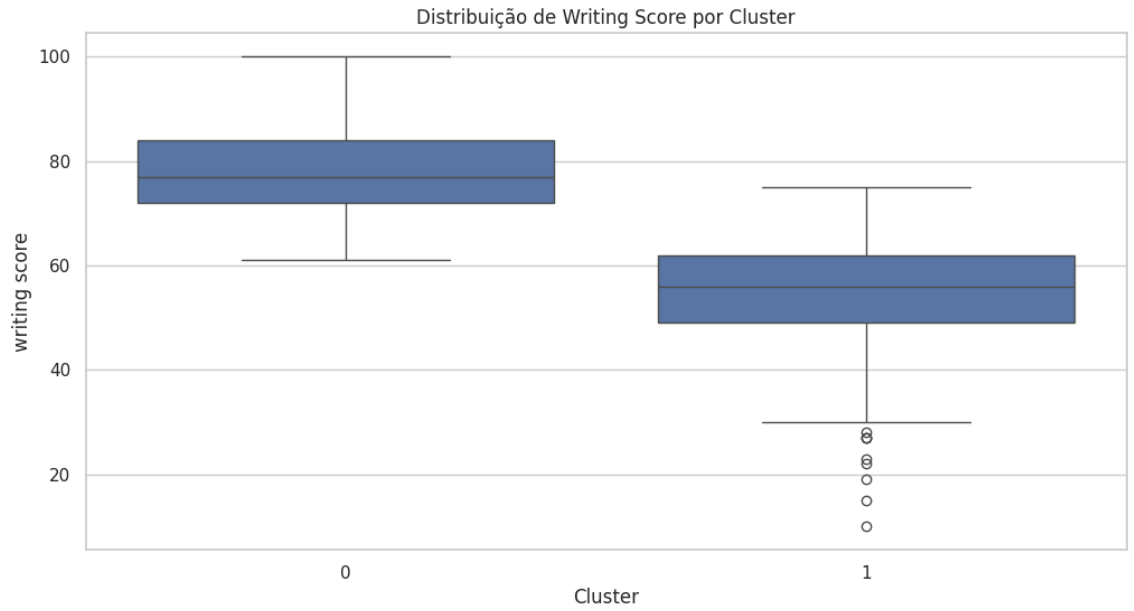
Cluster 0:

- Notas mais altas e consistentes:
  - A mediana está em torno de 80.
  - A maioria dos dados está entre 70 e 90, com alguns estudantes chegando a 100.
- Sem outliers visíveis, indicando estabilidade e homogeneidade.
- Interpretação: Este grupo corresponde a estudantes com alto desempenho em leitura e pouca dispersão nos resultados.

Cluster 1:

- Desempenho inferior:
  - A mediana está por volta de 58, indicando um desempenho médio a baixo.
  - A distribuição se estende desde 35 até 73, com muitos outliers abaixo de 35.
- Alta variabilidade:
  - Muitos valores extremos sugerem um grupo mais heterogêneo, com casos de desempenho muito fraco em leitura.
- Interpretação: Este grupo reúne estudantes com dificuldades em leitura, e maior desigualdade nos resultados.

Distribuição de Writing score por cluster:



Cluster 0:

- Notas mais altas:
  - A mediana está entre 75 e 80, indicando bom desempenho.
  - A maioria dos estudantes obteve notas entre 70 e 90, com alguns alcançando o máximo (100).
- Distribuição consistente e sem outliers visíveis.
- Interpretação: O grupo representa estudantes com forte desempenho em escrita, possivelmente com boas habilidades de comunicação e argumentação.

Cluster 1:

- Média e mediana mais baixas:
  - A mediana está por volta de 57.
  - A distribuição vai de aproximadamente 30 a 75, com vários outliers abaixo de 30.
- Alta variabilidade e presença de notas muito baixas.
- Interpretação: Este cluster agrupa estudantes com dificuldades na escrita, o que pode refletir limitações linguísticas, estruturais ou de vocabulário.

Agora irei comentar um pouco os centroides dos clusters e dos crosstab:

Centroides dos Clusters (Escala Padronizada):

Cluster	Math Score	Reading Score	Writing Score
0	0.6419	0.6969	0.6886
1	-0.8170	-0.8870	-0.8764

- Como os dados foram padronizados (Z-score), valores **positivos** indicam desempenho **acima da média**, e valores **negativos**, desempenho **abaixo da média**.

- **Cluster 0** possui centroides **positivos e elevados** em todas as notas, confirmando que esse grupo reúne alunos com **alto rendimento geral**.

- **Cluster 1** apresenta centroides **negativos e consistentes**, confirmando que são alunos com **baixo rendimento geral**.

- Essa diferença reforça que a clusterização captou uma segmentação significativa entre dois perfis distintos.

Distribuição por Gênero:

Cluster	Feminino	Masculino	Total
0	320	240	560
1	198	242	440

Cluster 0 (alto desempenho): 320 mulheres e 240 homens.

Cluster 1 (baixo desempenho): 198 mulheres e 242 homens.

Interpretação: Há uma **maior proporção de mulheres no cluster de alto rendimento** (57%) e uma proporção **maior de homens no cluster de baixo rendimento**.
Isso sugere uma possível influência do gênero no desempenho, com mulheres se destacando mais em leitura e escrita (o que é consistente com estudos educacionais).

5.T-SNE ou PCA como Pré-processamento para Classificação

Na última parte, tenho o O objetivo é verificar se a *redução de dimensionalidade* ou *normalização* melhora a performance de um classificador. Com isso, irei seguir essas etapas:

 Escolha do atributo alvo

- Variável alvo escolhida:
  - **gender** (binária → classificador simples)

 Modelos de classificação

- **Logistic Regression** (modelo linear)
- **Random Forest** (para não-linearidade)

 Cenários avaliados:

- **Cenário 1**: Dados normalizados (StandardScaler), sem redução.
- **Cenário 2**: PCA (2 ou 3 componentes).
- **Cenário 3**: t-SNE (2 componentes).

 Métrica de avaliação:

- **Accuracy** e **F1-Score** com validação cruzada.

Resultados:

Modelo	Pré-processamento	Accuracy	F1-score
Logistic Regression	Normalização	**0.878**	**0.872**
	PCA	0.869	0.863
	t-SNE	0.833	0.827
Random Forest	Normalização	0.846	0.838
	PCA	**0.849**	**0.842**
	t-SNE	0.848	0.840

Logistic Regression:

- **Melhor desempenho com normalização (Accuracy=0.878, F1=0.872)**.
  Isso faz sentido porque Logistic Regression é sensível à escala das variáveis, e a normalização garante pesos equilibrados.
- Com **PCA**, houve uma **pequena perda de desempenho**, mas ainda manteve boa performance (0.869).
- Com **t-SNE**, a queda foi maior (Accuracy=0.833), pois t-SNE não é ideal para redução de dimensionalidade em modelos preditivos (não preserva variância global).

Conclusão: Normalização é a melhor escolha para modelos lineares.

Random Forest

- Com **normalização**, o desempenho foi bom (Accuracy=0.846), mas não teve grande impacto, pois Random Forest é menos sensível à escala.
- Curiosamente, **PCA (0.849)** e **t-SNE (0.848)** tiveram performance ligeiramente superior, provavelmente por reduzir redundâncias sem prejudicar a separabilidade dos dados.
- A diferença entre PCA e t-SNE é mínima neste caso, mas **PCA é preferível** pela interpretabilidade e menor custo computacional.

Conclusão: Para modelos baseados em árvores, PCA pode trazer leve benefício, mas não é essencial.

Conclusões:

- **Normalização + Logistic Regression** é a melhor combinação para este dataset.
- **PCA** é útil quando se deseja reduzir dimensionalidade sem perda relevante de desempenho, sendo mais estável que t-SNE.
- **t-SNE** é excelente para visualização, mas **não é indicado para classificação**, pois perde relações globais importantes.

4. Conclusões

Para dar conclusão desse projeto, *queria apenas comentar que o objetivo principal desse projeto foi analisar o desempenho acadêmico de estudantes* ***por meio de técnicas de*** *análise exploratória, redução de dimensionalidade, clusterização e classificação supervisionada, utilizando o dataset StudentsPerformance.csv.*

5. Próximos passos

Em questão dos próximos passos, se é possível utilizar outros métodos de clusterização como DBSCAN ou Gaussian Mixture Models, aplicar técnicas de explicabilidade como SHAP para entender decisões dos modelos de classificação, expandir o modelo para outras variáveis-alvo (por exemplo, notas categorizadas como alto/médio/baixo) ou realizar coletar mais dados para validar os clusters em diferentes contextos escolares.