UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE QUIXADÁ

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

MATHEUS LEANDRO DE MELO SILVA

DESIGUALDADES EDUCACIONAIS NO BRASIL: UMA ANÁLISE POR CLUSTERIZAÇÃO DE INDICADORES EDUCACIONAIS E DESEMPENHO ESCOLAR

QUIXADÁ

2025

MATHEUS LEANDRO DE MELO SILVA

DESIGUALDADES EDUCACIONAIS NO BRASIL: UMA ANÁLISE POR CLUSTERIZAÇÃO DE INDICADORES EDUCACIONAIS E DESEMPENHO ESCOLAR

Projeto de pesquisa apresentado ao Curso de Graduação em Engenharia de Software do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Orientadora: Profa. Dra. Lívia Almada Cruz.

QUIXADÁ

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S581d Silva, Matheus Leandro de Melo.
Desigualdades educacionais no brasil : uma análise por clusterização de indicadores educacionais e desempenho escolar / Matheus Leandro de Melo Silva. – 2025.
62 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2025.
Orientação: Profa. Dra. Lívia Almada Cruz.
Coorientação: Prof. Dr. Régis Pires Magalhães.

1. clusterização. 2. k-means. 3. indicadores educacionais. 4. análise de dados. I. Título.
CDD 005.1

MATHEUS LEANDRO DE MELO SILVA

DESIGUALDADES EDUCACIONAIS NO BRASIL: UMA ANÁLISE POR CLUSTERIZAÇÃO DE INDICADORES EDUCACIONAIS E DESEMPENHO ESCOLAR

Projeto de pesquisa apresentado ao Curso de Graduação em Engenharia de Software do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Aprovada em: _____/_____/_____.

BANCA EXAMINADORA

Profa. Dra. Lívia Almada Cruz (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Régis Pires Magalhães (Coorientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Paulo de Tarso Guerra Oliveira
Universidade Federal do Ceará (UFC)

Ma. Tatieures Gomes Pires
Universidade Federal do Ceará (UFC)

RESUMO

As desigualdades educacionais no Brasil representam um desafio persistente, refletindo disparidades socioeconômicas e regionais que impactam o desempenho escolar. Este trabalho propõe uma análise dessas desigualdades por meio da clusterização de indicadores educacionais e do desempenho escolar, utilizando técnicas de aprendizado de máquina não supervisionado, com foco no algoritmo K-Means. O estudo busca identificar padrões e tendências na qualidade do ensino público brasileiro ao longo dos anos de 2015, 2019 e 2021. A metodologia adotada envolve a coleta e limpeza de dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), análise exploratória, transformação e normalização dos indicadores, além da aplicação de técnicas de clusterização para agrupar municípios com perfis educacionais semelhantes. A validação dos clusters foi realizada por meio de métricas como o coeficiente de silhueta e o método do cotovelo. Os resultados revelam uma evolução positiva no Índice de Desenvolvimento da Educação Básica (IDEB), mas também evidenciam disparidades regionais e desafios persistentes, como a taxa de distorção idade-série. A análise da evolução dos clusters ao longo do tempo mostrou migrações significativas entre grupos, sugerindo mudanças nas condições educacionais e socioeconômicas. O estudo conclui que a clusterização é uma ferramenta valiosa para a formulação de políticas públicas educacionais, contribuindo para a equidade e a melhoria da qualidade do ensino no Brasil.

Palavras-chave: clusterização; k-means; indicadores educacionais; análise de dados.

ABSTRACT

Educational inequalities in Brazil represent a persistent challenge, reflecting socio-economic and regional disparities that impact school performance. This study proposes an analysis of these inequalities through the clustering of educational indicators and school performance, using unsupervised machine learning techniques, with a focus on the K-Means algorithm. The study seeks to identify patterns and trends in the quality of Brazilian public education over the years 2015, 2019, and 2021. The methodology involves data collection and cleaning from the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), exploratory analysis, transformation and normalization of indicators, as well as the application of clustering techniques to group municipalities with similar educational profiles. The validation of the clusters was carried out through metrics such as the silhouette coefficient and the elbow method. The results reveal a positive evolution in the Índice de Desenvolvimento da Educação Básica (IDEB), but also highlight regional disparities and persistent challenges, such as the age-grade distortion rate. The analysis of the evolution of the clusters over time showed significant migrations between groups, suggesting changes in educational and socio-economic conditions. The study concludes that clustering is a valuable tool for the formulation of educational public policies, contributing to equity and the improvement of education quality in Brazil.

Keywords: clustering; k-means; educational indicators; data analysis.

SUMÁRIO

1 INTRODUÇÃO 7

1.1 Objetivos 8

1.1.1 Objetivo Geral 8

1.1.2 Objetivos Específicos 8

2 FUNDAMENTAÇÃO TEÓRICA 9

2.1 Indicadores de Desempenho 9

2.1.1 Indicadores Educacionais 9

2.1.1.1 Índice de Desenvolvimento da Educação Básica *10*

2.1.1.2 Taxa de Distorção Idade-Série por Escola *10*

2.1.1.3 Taxa de Rendimento Escolar *11*

2.1.1.4 Índice do Nível Socioeconômico (INSE) *11*

2.2 Aprendizado de Máquina 12

2.2.1 Aprendizado Não Supervisionado 12

2.2.1.1 Dados Não Rotulados *13*

2.2.2 Clusterização 13

2.2.2.1 Algoritmo K-Means *14*

2.3 Métricas de Avaliação de *Clusters* 18

2.3.1 Coeficiente de Silhueta 18

2.3.2 Índice de Davies-Bouldin 18

2.3.3 Índice de Calinski-Harabasz 19

2.3.4 Método do Cotovelo 19

2.4 Evolução dos *Clusters* 20

3 TRABALHOS RELACIONADOS 22

3.1 *Clustering Analysis for Classifying Student Academic Performance in Higher Education* 22

3.2 *Density-Based Unsupervised Learning Algorithm to Categorize College Students into Dropout Risk Levels* 23

3.3 *A decision-making framework for school infrastructure improvement programs* 24

3.4 ***Analysis of the socioeconomic impact due to COVID-19 using a deep clustering approach*** 25
3.5 **Análise Comparativa** 26
4 **METODOLOGIA** 27
4.1 **Seleção e Coleta dos Dados** 27
4.2 **Processamento e Limpeza dos Dados** 28
4.3 **Análise Exploratória** 30
4.4 **Transformação dos Dados para Clusterização** 32
4.5 **Escolha dos Parâmetros do Algoritmo** 34
4.6 **Execução e Validação da Clusterização** 35
4.7 **Análise dos Perfis dos *Clusters*** 36
4.8 **Análise da Evolução dos *Clusters*** 37
5 **RESULTADOS** 38
5.1 **Análise Exploratória** 38
5.2 **Escolha dos Parâmetros do Algoritmo** 44
5.3 **Execução e Validação da Clusterização** 47
5.4 **Análise dos Perfis dos Clusters** 49
5.5 **Análise da Evolução dos Clusters** 51
6 **CONCLUSÕES E TRABALHOS FUTUROS** 55
REFERÊNCIAS 57

1 INTRODUÇÃO

Nos últimos anos, a análise de dados emergiu como uma ferramenta crucial para compreender e abordar uma ampla gama de questões complexas em diversos campos, impulsionada pelo grande aumento na disponibilidade de dados (Dhar, 2013). Essa abordagem permite identificar padrões, tendências e correlações que podem informar decisões políticas e estratégicas em áreas como saúde, economia e educação (Jain, 2010). Dentre as técnicas de análise de dados, a clusterização se destaca por sua capacidade de agrupar conjuntos de dados distintos e determinar suas inter-relações, oferecendo percepções valiosas sobre problemas sociais e econômicos (Xu; Wunsch, 2008).

No contexto educacional, a clusterização tem sido amplamente utilizada para analisar desempenho estudantil, identificar fatores de evasão e orientar políticas públicas. Um exemplo notável é o trabalho de Nafuri *et al.* (2022), que empregou a clusterização para identificar as causas subjacentes ao elevado índice de desistência universitária na Malásia. Uma vez que, essa técnica permite categorizar conjuntos de dados em grupos distintos, agrupando os dados mais semelhantes entre si e revelando padrões que podem informar intervenções mais eficazes (MacQueen *et al.*, 1967).

No Brasil, o sistema educacional enfrenta desafios históricos e estruturais, marcados tanto por disparidades socioeconômicas, como pelo acesso desigual e baixos índices de desempenho acadêmico, especialmente em regiões vulneráveis. O Programa Internacional de Avaliação de Estudantes (Pisa) de 2022 revelou que menos da metade dos estudantes brasileiros de 15 anos atingiu o aprendizado mínimo em matemática e ciências (CNN, 2023). Essa defasagem, evidenciada por estudos anteriores (Cutler; Lleras-Muney, 2012), ressalta como as desigualdades regionais e demográficas prejudicam o acesso à educação e os resultados escolares.

Diante desse cenário, compreender a evolução dos padrões educacionais ao longo do tempo é essencial para orientar políticas educacionais mais eficazes. A análise de clusters aplicada a indicadores educacionais permite identificar mudanças estruturais no sistema de ensino, revelar tendências de desenvolvimento e avaliar a efetividade de políticas implementadas. Assim como, avaliar evoluções na qualidade do ensino, tanto negativa quanto positivamente. No entanto, a complexidade e a constante modificação das classificações de dados representam um desafio significativo, tornando o uso de técnicas de clusterização não supervisionada uma abordagem promissora para analisar esses dados sem a necessidade de categorização prévia (Li *et al.*, 2021).

Portanto, este trabalho visa investigar e analisar a evolução da qualidade do ensino nas escolas públicas brasileiras, por meio dos indicadores educacionais. Ao empregar técnicas de análise de dados, como a clusterização, na interseção entre os campos da educação e a realidade das escolas de diferentes municípios, pretendemos identificar padrões e relações que forneçam uma análise clara da qualidade do ensino público no Brasil. Essa abordagem permitirá uma compreensão mais profunda das disparidades regionais e das necessidades específicas de cada área, orientando a alocação de recursos de forma mais eficiente e equitativa (Murtagh; Contreras, 2012).

1.1 Objetivos

Esta seção é referente ao objetivo geral e específicos do presente trabalho.

1.1.1 Objetivo Geral

O principal objetivo deste trabalho é analisar o panorama atual da educação pública brasileira com base na clusterização dos indicadores educacionais.

1.1.2 Objetivos Específicos

- Identificar os principais indicadores educacionais relevantes para a análise.
- Avaliar a correlação entre os diferentes indicadores educacionais.
- Caracterizar municípios brasileiros utilizando a clusterização para identificar diferentes perfis de indicadores educacionais.
- Avaliar a evolução dos diferentes *clusters* ao longo do tempo.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os principais conceitos e termos que constituem a base teórica deste trabalho. Inicialmente, na Seção 2.1 será explicado os indicadores usados neste trabalho e sua importância para a correta realização do mesmo. Em seguida, na Seção 2.2 é abordado o aprendizado de máquina, destacando sua usabilidade, técnicas e importância. Além disso, serão discutidas as métricas usadas para validar os clusters gerados na Seção 2.3. Por fim, na Seção 2.4, será abordado a análise da evolução dos clusters, onde será explorando como as mudanças nos indicadores podem influenciar a formação de novos agrupamentos ou a alteração dos já existentes.

2.1 Indicadores de Desempenho

Indicadores de desempenho são ferramentas fundamentais para a análise e planejamento de políticas públicas. Porque, eles fornecem informações cruciais sobre o desenvolvimento humano e social de uma população e por isso são amplamente utilizados por governos, organizações não governamentais e instituições de pesquisa. Além disso, esses indicadores permitem mensurar fenômenos complexos de uma forma simplificada e objetiva, facilitando a análise e a interpretação de dados em diversas áreas do conhecimento. Logo, sua importância reside na capacidade de transformar dados brutos em informações valiosas para a tomada de decisões e o desenvolvimento de estratégias eficazes.

2.1.1 Indicadores Educacionais

Os indicadores educacionais são fundamentais para mensurar o desempenho, a qualidade e o acesso à educação em uma sociedade. Eles fornecem uma visão clara sobre o estado da educação em diferentes níveis, desde a educação básica até o ensino superior, e permitem identificar desigualdades, avanços e áreas que necessitam de melhorias (Gonçalves *et al.*, 2017). Disponíveis no portal do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira[1], podem incluir taxas de alfabetização, níveis de escolaridade, índices de evasão e retenção escolar, além do desempenho em avaliações nacionais e internacionais.

[1] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais

2.1.1.1 Índice de Desenvolvimento da Educação Básica

O IDEB é um indicador educacional, criado em 2007 pelo INEP, que sintetiza a qualidade do ensino no Brasil, combinando os índices de aprovação escolar e o desempenho dos estudantes em avaliações padronizadas como o Sistema de Avaliação da Educação Básica (SAEB). Realizado a cada dois anos, seu cálculo se da por meio da multiplicação da taxa de rendimento com o desempenho em avaliações como o SAEB, podendo variar de 0 a 10 e servindo como referência para a definição de metas e o acompanhamento da evolução da educação básica em todas as esferas de gestão.

Além disso, a metodologia empregada pelo IDEB equilibra as dimensões de fluxo escolar e aprendizagem, incentivando a busca por um ensino de qualidade. Ao estabelecer metas claras e transparentes, o indicador contribui para a responsabilização dos gestores educacionais e para a mobilização da sociedade em prol da melhoria contínua da educação (INEP, 2024c). Logo, sua simplicidade e a abrangência permitem a sua utilização como uma ferramenta para a formulação e implementação de políticas públicas educacionais. Em resumo, o IDEB é fundamental para a gestão da educação no Brasil, pois ao fornecer um diagnóstico da situação do ensino, o indicador possibilita a identificação de áreas que requerem maior investimento e a adoção de medidas específicas para superar os desafios e alcançar os objetivos estabelecidos.

2.1.1.2 Taxa de Distorção Idade-Série por Escola

A Taxa de Distorção Idade-Série é um indicador educacional que mede a porcentagem de alunos com idade superior à recomendada, mais de 2 anos de atraso escolar, para o ano ou série em que estão matriculados (INEP, 2024a). Seu cálculo é feito com base na relação entre as matrículas em determinada série e a idade dos alunos, considerando como adequada a idade esperada acrescida de um ano. Por exemplo, no 6º ano do ensino fundamental, a idade adequada é de 6 a 7 anos, de modo que alunos mais velhos nessa série são classificados como distorção.

Além disso, sua discrepância geralmente ocorre devido a reprovações ou atrasos no início da vida escolar, resultando em problemas relacionados a gestão pedagógica, a retenção escolar e a qualidade de ensino. Por isso, sua importância para formulação de políticas de combate ao fracasso escolar e à evasão. Pois, reprovações têm uma série de efeitos negativos na permanência dos alunos na escola, principalmente, na motivação para estudar (QEdu, 2024).

2.1.1.3 Taxa de Rendimento Escolar

Produzidas anualmente, as taxas de rendimento escolar são fundamentais para verificar e acompanhar os dados educacionais de escolas e municípios. Além disso, elas também são incorporadas ao cálculo do Índice de Desenvolvimento da Educação Básica (IDEB). Vale ressaltar que, o cálculo das taxas de rendimento tem como base as informações de desempenho e movimentação dos alunos, coletadas na segunda etapa do Censo Escolar, denominada "Situação do Aluno". Nessa fase, são preenchidos dados sobre a situação acadêmica dos estudantes, como aprovação ou reprovação, matrícula inicial, abandono escolar, transferência, falecimento, entre outros (INEP, 2024b).

Assim, este indicador torna-se essencial para monitorar o progresso educacional e identificar escolas ou regiões que necessitam de maior atenção nas políticas públicas. Por avaliar o desempenho dos alunos ao longo do período letivo, é uma métrica central para medir a eficácia das estratégias pedagógicas e a qualidade do ensino oferecido.

2.1.1.4 Índice do Nível Socioeconômico (INSE)

O Índice do Nível Socioeconômico (INSE) é um importante indicador criado para avaliar as condições socioeconômicas dos estudantes e suas famílias. Seu objetivo principal é analisar a relação entre o contexto econômico dos alunos e seu desempenho escolar, ajudando assim a entender as desigualdades educacionais no Brasil. Para tanto, este índice é formado por dados essenciais, como renda familiar, escolaridade dos pais, acesso a bens de consumo e condições de moradia. Essas informações são cruciais para compreender, portanto, como o nível socioeconômico afeta o acesso à educação de qualidade e o sucesso acadêmico.

Além disso, o INSE permite correlacionar condições socioeconômicas com resultados acadêmicos, oferecendo uma visão mais justa em avaliações como o Enem. Pois, essa abordagem considera as desigualdades contextuais, proporcionando uma interpretação equitativa do desempenho escolar. Ademais, a análise desses dados revela padrões que evidenciam a influência do nível socioeconômico nas oportunidades educacionais e resultados acadêmicos, que podem ser usados por governos e instituições para melhorar o acesso e a qualidade da educação, especialmente para estudantes em vulnerabilidade social. Por fim, o cálculo do INSE baseia-se em dados provenientes de censos educacionais, questionários socioeconômicos aplicados em avaliações nacionais e outras fontes de dados populacionais.

2.2 Aprendizado de Máquina

Aprendizado de máquina é um subcampo da Inteligência Artificial que se concentra no desenvolvimento de algoritmos e modelos capazes de aprender padrões a partir de dados (Shinde; Shah, 2018). Ao contrário dos programas tradicionais, programados para realizar tarefas específicas, aprendizado de máquina ensina as máquinas a processarem dados de forma mais eficiente (Mahesh, 2020).

Uma das principais características que distingue o aprendizado de máquina é sua capacidade de adaptação e melhoria contínua, à medida que o sistema é exposto a mais dados. Isso permite que os sistemas identifiquem padrões, façam previsões e tomem decisões sem intervenção humana direta. Assim, os algoritmos aprendem interativamente, utilizando dados de treinamento para resolver problemas e descobrindo padrões sem a necessidade de programação explícita (Janiesch *et al.*, 2021).

Vale ressaltar que o aprendizado de máquina tem importância crescente em diversas áreas, possibilitando avanços em campos como finanças, segurança da informação e saúde (Badillo *et al.*, 2020). Além disso, existem dois principais tipos de aprendizado, o supervisionado, onde o modelo é treinado com dados rotulados, permitindo ao algoritmo aprender a partir de exemplos já classificados para fazer previsões e o não supervisionado, onde o algoritmo trabalha com dados não rotulados, buscando padrões ocultos ou estruturas nos dados (Bi *et al.*, 2019).

Portanto, neste projeto, o foco será no aprendizado não supervisionado, uma vez que o objetivo é explorar e identificar agrupamentos naturais nos dados, por meio de um grande conjunto de dados não rotulados.

2.2.1 Aprendizado Não Supervisionado

No aprendizado não supervisionado, os algoritmos são encarregados de identificar padrões e relacionamentos nos dados sem a necessidade de rótulos de saída previamente definidos (Usama *et al.*, 2019). Esse tipo de aprendizado é amplamente utilizado em tarefas como a clusterização, que agrupa dados com base em similaridades, e a redução de dimensionalidade, cujo objetivo é diminuir o número de atributos no conjunto de dados, simplificando a complexidade da análise e reduzindo os custos computacionais associados (Yang; Sinaga, 2019).

2.2.1.1 Dados Não Rotulados

Devido à abundância de informações disponíveis eletronicamente, processá-las eficientemente torna-se mais viável quando os dados são organizados para que dados similares fiquem agrupados. Além disso, com o baixo custo de grandes conjuntos de dados disponíveis em diversas aplicações, o uso de dados não supervisionados se tornam mais vantajosos para o aprendizado de máquina (Zhang; Oles, 2000).

Nesse contexto, dados não rotulados são aqueles que não possuem categorias ou rótulos previamente definidos (Tan *et al.*, 2016). A principal tarefa com esses dados é descobrir padrões e estruturas sem informações prévias sobre a classe dos dados. A análise desses dados pode revelar padrões ocultos e percepções valiosas, além de reduzir o tempo e os custos na preparação dos dados. Contudo, a interpretação dos resultados pode ser subjetiva, especialmente sem conhecimento prévio ou objetivos claros.

2.2.2 Clusterização

Clusterização é uma técnica de aprendizado não supervisionado que agrupa dados com características semelhantes, mantendo grupos distintos para dados com diferenças significativas. Sendo uma das principais técnicas de aprendizado de máquina não supervisionado, a clusterização forma agrupamentos com base em padrões encontrados nos dados não rotulados, definidos automaticamente pelo algoritmo escolhido (Kriegel *et al.*, 2011).

Contudo, a escolha do algoritmo de clusterização é um dos pontos centrais no processo analítico, pois segundo Pitafi *et al.* (2023) diferentes métodos podem produzir resultados variados dependendo do conjunto de dados. Entre os algoritmos mais populares estão o *K-Means*, o DBSCAN e o Hierárquico (Nikita Sachdeva, 2023).

Aplicações da clusterização podem ser encontradas em várias áreas. No marketing, ela possibilita a segmentação de clientes com base em padrões de compra e comportamento, permitindo ações mais precisas e personalizadas (Liu; Ong, 2008). Na biologia, a técnica é amplamente utilizada na classificação celular e na análise de dados genômicos, oferecendo visões importantes para a pesquisa científica (Nugent; Meila, 2010). Além disso, em setores como a economia e a educação, a clusterização pode revelar padrões latentes em indicadores que de outra forma passariam despercebidos (Shaulska *et al.*, 2020).

Neste trabalho, a clusterização será empregada para analisar a evolução e correlação dos grupos formados por indicadores educacionais ao longo do tempo. O objetivo é compreender as mudanças na qualidade do ensino, identificando padrões de progresso ou retrocesso dentro dos agrupamentos. Essa abordagem possibilita uma visão mais detalhada sobre as tendências educacionais, contribuindo para avaliações mais precisas e embasando estratégias de melhoria no sistema educacional brasileiro.

2.2.2.1 Algoritmo K-Means

O algoritmo *K-Means* é uma técnica de aprendizado de máquina não supervisionado que agrupa pontos de dados não rotulados em grupos distintos, conhecidos como **k** *clusters*, onde **k** é um parâmetro definido pelo usuário. Esse algoritmo pretende minimizar a distância entre pontos dentro de *clusters* semelhantes e maximizar a distância entre *clusters* diferentes. Essa abordagem se destaca por sua eficácia na identificação de padrões em conjuntos de dados não rotulados, oferecendo uma solução simples e escalável para a análise de agrupamentos. Devido à sua simplicidade conceitual e capacidade de revelar estruturas subjacentes em dados, o *K-Means* é amplamente utilizado em diversas áreas, como análise de dados, reconhecimento de padrões e segmentação de mercado.

Mais especificamente, o *K-Means* encontra aplicações em diferentes domínios, incluindo a segmentação de clientes no setor de marketing, a categorização de documentos e a detecção de anomalias em sistemas financeiros. Na área educacional, essa técnica tem sido empregada para analisar padrões de desempenho estudantil, identificar grupos de alunos com características semelhantes e auxiliar no desenvolvimento de estratégias pedagógicas personalizadas. Sua versatilidade e escalabilidade tornam-no uma ferramenta essencial para a análise de grandes volumes de dados em diversas áreas do conhecimento.

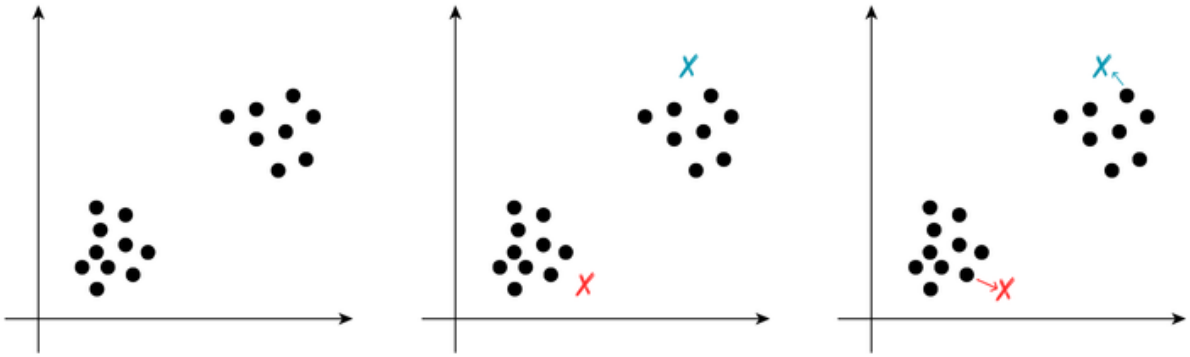
Por conseguinte, uma de suas principais vantagens é a eficiência computacional, principalmente em conjuntos de dados grandes. Porém, a escolha da quantidade de *clusters* é sua desvantagem, uma vez que geralmente requer conhecimento prévio do domínio ou o uso de técnicas de validação de *clusters*, como o método do cotovelo ou o coeficiente de silhueta. Vale ressaltar que ele também possui limitações como a inicialização dos centroide por conta do uso de números aleatórios, resultando em soluções sub ótimas. Por fim, o Algoritmo 1 representa um pseudocódigo para o algoritmo *K-Means*, e as Figuras 1 e 2 demonstram os passos de execução do mesmo, onde é utilizado um conjunto de dados fictícios de trabalhadores.

Algoritmo 1: K-Means

```
Function kmeans(dados, k, max_iter):
    centroides ← inicializar_centroides(dados, k);
    for iter = 1 até max_iter do
        clusters ← atribuir_pontos(dados, centroides);
        centroides_atualizados ← atualizar_centroides(dados, clusters, k);
        if centroides_atualizados são iguais a centroides then
            break;
        end
        centroides ← centroides_atualizados;
    end
    return clusters;
Function inicializar_centroides(dados, k):
    centroides ← selecionar_k_pontos_aleatórios(dados, k);
    return centroides;
Function atribuir_pontos(dados, centroides):
    clusters ← estrutura_vazia_com_k_listas();
    for cada ponto em dados do
        centroide_mais_proximo ← encontrar_centroide_mais_proximo(ponto,
         centroides);
        adicionar ponto ao cluster associado a centroide_mais_proximo;
    end
    return clusters;
Function atualizar_centroides(dados, clusters, k):
    centroides_atualizados ← lista vazia;
    for cada cluster do
        novo_centroide ← calcular_media(cluster);
        adicionar novo_centroide a centroides_atualizados;
    end
    return centroides_atualizados;
```

Fonte: Elaborada pelo autor.

Figura 1 – Primeiras etapas da clusterização



(a) Conjunto inicial (b) Definir centroides (c) Atribuir centroides

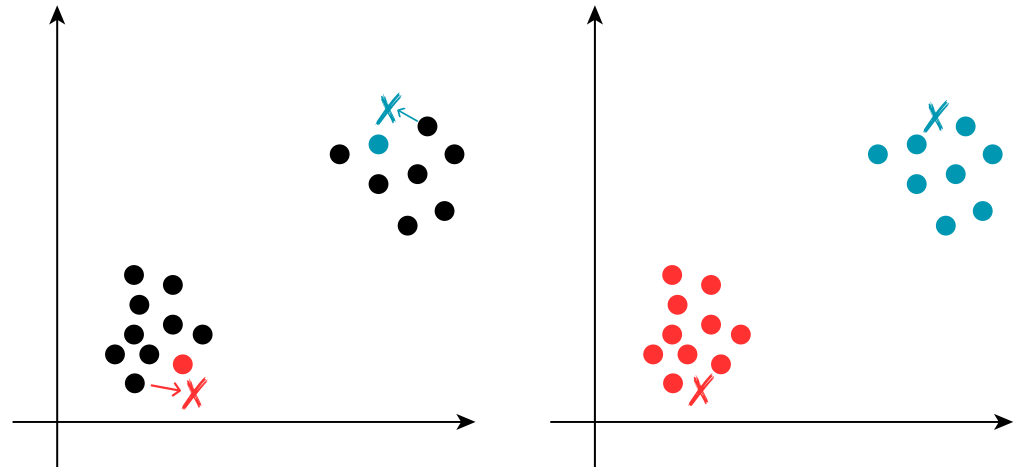
Fonte: Elaborada pelo autor.

Primeiramente, com base em conhecimentos prévios dos dados e uso de técnicas como o gráfico do cotovelo, o usuário deverá decidir e definir um número de *clusters* que melhor funcionará com os dados que serão usados. Então o usuário chamará a função `kmeans` passando o conjunto de dados, o número de *clusters* e em alguns casos o número de interações que normalmente possui um valor base. Por conta de ter apenas dois conjuntos de dados, nesse exemplo o número de *clusters* passado será de (k = 2), o valor de interações máxima será 20 e será passado um dado fictício contendo idades e salários de trabalhadores.

Em seguida, na função `kmeans`, os centroides serão inicializados utilizando a função `inicializar_centroid`. Por conseguinte, um laço é executado, onde os pontos são atribuídos aos *clusters*, os centroides são atualizados e um novo conjunto de centroides é definido. Após isso, o laço continua até que os centroides não mudem entre duas iterações consecutivas ou até que o número máximo de iterações seja atingido. Após o término do laço, os *clusters* finais são retornados como resultado da função `kmeans`. Ou seja, nessa parte do exemplo temos apenas os conjuntos de dados iniciais prontos para serem agrupados, conformo visto na Figura 1a.

Inicialmente, como não sabemos os centros exatos dos *clusters*, os centroides, selecionamos aleatoriamente alguns pontos de dados e definimos esses pontos como centroides para cada *cluster*. Conforme a Figura 1b, dois centroides são inicializados no conjunto de dados. A função `inicializar_centroides`, onde é passado os dados e o número de *clusters*, seleciona aleatoriamente *k* pontos do conjunto de dados para serem os centroides iniciais. Estes serão os centros dos *clusters* no início do algoritmo chamado no início da função `kmeans`.

Figura 2 – Etapas finais da clusterização



(a) Repetição do processo (b) Resultado final

Fonte: Elaborada pelo autor.

Na Figura 1c, cada ponto de dados é atribuído ao *cluster* cujo centroide está mais próximo. Para isso, calcula-se a distância entre cada ponto e todos os centros usando a métrica de distância euclidiana. O ponto encontrado é então associado ao *cluster* cujo centroide possui a menor distância. Como mostrado na Figura 2a, esse processo se repete para todos os pontos do conjunto. A função `atribuir_pontos` realiza essa operação percorrendo os dados e aplicando a métrica de distância para determinar a melhor atribuição.

Depois que todos os pontos de dados são atribuídos aos *clusters*, os centroides são atualizados recalculando-os como a média de todos os pontos de dados atribuídos a cada *cluster*. Este passo é essencial para reposicionar os centroides com base nos novos membros de cada *cluster*. A função `atualizar_centroides` recalcula os centroides de cada *cluster* como a média de todos os pontos atribuídos a esse *cluster* na etapa anterior. Isso move os centroides para uma posição mais central dentro de seus *clusters*.

Como mencionado na Figura 2a, o processo se repete até que a posição dos centroides varie minimamente entre duas etapas consecutivas ou até que um número máximo de iterações seja atingido. Isso indica que os pontos de dados não estão mais mudando de grupos e que o algoritmo atingiu a convergência. Como visto na função `kmeans`, esse critério pode ser a estabilidade dos centroides ou o limite de iterações. Quando os centros se estabilizam, os grupos finais são definidos, concluindo o processo de clusterização, como mostrado na Figura 2b.

2.3 Métricas de Avaliação de *Clusters*

A classificação de dados para identificar regularidades torna o conceito de similaridade cada vez mais relevante no processamento inteligente de dados, pois, nos algoritmos de clusterização, a detecção dessa similaridade é fundamental para a escolha adequada das métricas de avaliação. Assim, por meio dessas métricas, a qualidade dos *clusters* formados deve ser rigorosamente avaliada para garantir agrupamentos significativos (Grabusts, 2011). Logo, nesta seção serão apresentadas as principais métricas de avaliação, que auxiliam na escolha do número ideal de *clusters* e na validação dos resultados obtidos.

2.3.1 Coeficiente de Silhueta

O coeficiente de silhueta é uma métrica que pode otimizar o método *K-Means* na formação ou determinação do número de *clusters*. Seu principal uso, no entanto, é na avaliação e validação da qualidade da clusterização. Pois, ele testa a distância entre os *clusters* e a densidade dos mesmos, medindo a proximidade ou distância relativa entre os objetos e outros *clusters* (Hartama; Anjelita, 2022). Assim, este método será utilizado para avaliar o número de *clusters* formados pelo algoritmo *K-Means*, o qual é utilizado neste trabalho.

Seu funcionamento se dá pela combinação dos fatores de coesão e separação. A coesão é a semelhança entre o objeto e o *cluster*, enquanto a separação refere-se à comparação desse objeto com outros *clusters*. Essa comparação é realizada por meio do valor da Silhueta, que varia entre -1 e 1. Um valor de Silhueta próximo de 1 indica uma forte relação entre o objeto e o *cluster*. Portanto, se um *cluster* de dados apresenta um valor de Silhueta elevado, ele pode ser considerado adequado e aceitável (Yuan; Yang, 2019).

2.3.2 Índice de Davies-Bouldin

O Índice de Davies-Bouldin é uma métrica interna utilizada para avaliar a qualidade dos *clusters* em métodos de clusterização. Assim, a avaliação ocorre com base na compacidade, quando os elementos em um mesmo grupo são similares, e a separação entre eles, buscando maximizar a diferença entre *clusters* distintos e minimizar a variação interna de cada *cluster*. Logo, quando a distância *inter-cluster* é elevada, as características dos *clusters* são distintas, tornando as diferenças entre eles mais evidentes, enquanto uma menor distância indica que os objetos possuem uma alta similaridade de características (Mughnyanti *et al.*, 2020).

Por fim, seu cálculo é dado pela média da razão entre a soma das dispersões dentro de cada *cluster* e a distância entre os *centróides* de clusters distintos. Ou seja, quanto menor o valor do Índice de Davies-Bouldin, melhor é a qualidade do agrupamento. Além disso, um valor baixo indica que os *clusters* são compactos internamente e bem separados entre si. Por isso, essa métrica é amplamente utilizada devido à sua simplicidade e eficiência computacional. Além disso, é especialmente útil em situações onde não se dispõe de rótulos externos para comparação, oferecendo uma avaliação interna e objetiva da qualidade dos *clusters*.

2.3.3 Índice de Calinski-Harabasz

O Índice de Calinski-Harabasz, também conhecido como Razão de Variância, é uma métrica utilizada para avaliar a qualidade dos clusters formados em algoritmos de clusterização. Além disso, como outras métricas internas, ele não depende de rótulos externos, tornando-o ideal para situações em que a classe real dos dados é desconhecida, tendo como foco a maximização da separação entre clusters e a minimização da dispersão interna de cada cluster.

Formalmente, este índice é calculado com base na razão entre a dispersão total entre os clusters e a dispersão interna dentro de cada cluster. Portanto, sua fórmula considera o número de clusters e o número total de pontos no conjunto de dados. Onde, quanto maior o valor do índice, melhor é a qualidade do agrupamento, já que valores altos indicam que os clusters são bem separados entre si e que os elementos dentro de cada cluster estão mais compactos (Wang; Xu, 2019).

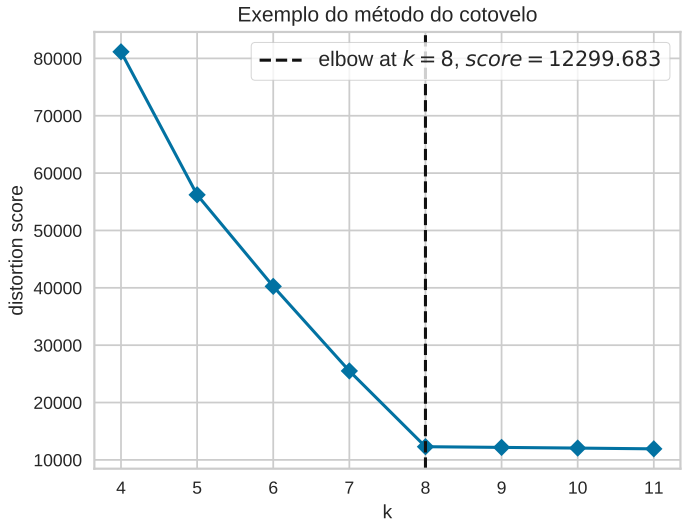
Além disso, esse índice é eficiente do ponto de vista computacional e é comumente aplicado em problemas onde a avaliação objetiva da qualidade dos agrupamentos é fundamental (Lima; Cruz, 2020). Contudo, o Índice de Calinski-Harabasz tende a favorecer soluções com mais clusters, o que pode levar a uma super segmentação se não for aplicado em conjunto com outros critérios de validação, como o método do cotovelo ou o Índice de Davies-Bouldin

2.3.4 Método do Cotovelo

O método do cotovelo é uma técnica amplamente utilizada na análise de dados para determinar o número ideal de *clusters* em algoritmos de agrupamento, como o *K-Means*. Sua função principal é definir a quantidade ideal de k no *K-Means*, uma vez que essa escolha garante uma maior qualidade nos *clusters* gerados. Isso ocorre porque o uso de um valor maior de k nem sempre resulta em uma melhoria nos resultados (Cui *et al.*, 2020).

Esse método é mais adequado para valores relativamente pequenos de k e opera por meio do cálculo da soma das diferenças quadráticas. À medida que k aumenta, o grau médio de distorção diminui, resultando em uma redução no número de amostras em cada categoria e fazendo com que essas amostras fiquem mais próximas ao centro de gravidade. O ponto em que a melhoria no grau de distorção passa a diminuir de forma menos significativa corresponde ao valor de k relacionado ao cotovelo (Bholowalia; Kumar, 2014).

Figura 3 – Método do cotovelo



Fonte: Elaborada pelo autor.

Na Figura 3, gerada a partir de dados fictícios, é possível observar que, à medida que k aumenta, a soma das diferenças quadráticas diminui de forma acentuada, indicando uma melhora significativa na qualidade dos *clusters*. Contudo, em um determinado ponto, a taxa de diminuição começa a desacelerar, formando um cotovelo. Esse ponto é, portanto, considerado o valor apropriado de k, pois representa um equilíbrio entre a complexidade do modelo e a qualidade do agrupamento.

2.4 Evolução dos *Clusters*

A análise da evolução dos *clusters* é uma etapa fundamental na segmentação de dados ao longo do tempo, visando a compreensão das mudanças ocorridas em contextos específicos. Essa análise não apenas permite identificar a formação inicial dos *clusters*, mas também acompanhar o comportamento desses grupos de dados ao longo do tempo, revelando dinâmicas de mudança significativas.

Por meio da análise temporal dos *clusters*, é possível observar o comportamento desses grupos de dados ao longo do tempo, identificando padrões como a criação, sobrevivência, desaparecimento, fusão, divisão, expansão e retração dos *clusters*. Cada uma dessas dinâmicas oferece informações sobre a natureza das variáveis em estudo e permite uma compreensão mais profunda das forças que influenciam os dados. Por exemplo, a criação de novos *clusters* pode sugerir a emergência de novas tendências ou a introdução de variáveis que antes não eram observadas. Já a fusão de *clusters* pode indicar uma convergência de características ou fatores que antes eram vistos como distintos.

A análise da evolução dos *clusters* também oferece percepções sobre a estabilidade ou volatilidade dos grupos ao longo do tempo. A sobrevivência de um *cluster* pode refletir a continuidade de determinadas condições ou a persistência de determinados fatores, enquanto a retração ou desaparecimento de um *cluster* pode sinalizar a perda de relevância de determinados fenômenos ou a obsolescência de fatores anteriormente importantes. Tais mudanças são cruciais para entender a dinâmica de contextos como o educacional, onde a variação dos índices de desempenho pode refletir a eficácia das políticas públicas ou a evolução das condições socioeconômicas nas diferentes regiões analisadas.

Ademais, a evolução dos *clusters* não se limita apenas à identificação de tendências de mudança. Através dessa análise, é possível investigar os fatores subjacentes a essas transformações. Fatores como alterações nas políticas educacionais, variações nos investimentos governamentais podem ter impactos significativos na dinâmica dos *clusters*. Essa investigação detalhada oferece uma oportunidade para direcionar ações de intervenção, como políticas públicas focadas em áreas ou grupos de maior vulnerabilidade, ou iniciativas para fortalecer regiões com tendências de melhoria.

A análise da evolução dos *clusters* também permite uma visão estratégica sobre o desenvolvimento de intervenções. No caso dos indicadores educacionais, por exemplo, a detecção de clusters em retração ou expansão pode sinalizar mudanças relevantes que exigem atenção. A retração de um *cluster* associado a bons resultados pode indicar um problema, enquanto a expansão de um *cluster* com indicadores negativos pode ser um alerta para a necessidade de novas estratégias. Por outro lado, a retração de um *cluster* problemático ou a expansão de um *cluster* positivo pode sinalizar impactos favoráveis de políticas implementadas. Por isso, a capacidade de identificar e interpretar essas mudanças ao longo do tempo é importante para os pesquisadores, formuladores de políticas e profissionais envolvidos na análise e gestão de dados.

3 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados os principais trabalhos relacionados a esta pesquisa. A Seção 3.1 aborda um estudo sobre desistências em instituições públicas de ensino superior na Malásia, focando em alunos de baixa renda por meio da clusterização de dados. A Seção 3.2 descreve um modelo de aprendizado de máquina para agrupar alunos conforme o risco de evasão. A Seção 3.3 propõe um framework para priorizar investimentos nas escolas usando técnicas de clusterização. Na Seção 3.4, é analisada a evolução de clusters para identificar o impacto da COVID-19 na economia. Por fim, a Seção 3.5 compara este trabalho com artigos relacionados.

3.1 *Clustering Analysis for Classifying Student Academic Performance in Higher Education*

Nesse estudo, Nafuri *et al.* (2022), por meio de uma abordagem de aprendizado de máquina não supervisionado, realiza uma clusterização para os estudantes categorizados no mais baixo nível de renda, baseado no desempenho em universidades, tendo como motivação o alto índice de desistência dos alunos das universidades. Esse estudo visa auxiliar o governo da Malásia a reduzir o número de desistência, e consequentemente, aumentar o número de formações.

Primeiramente, foi realizado o pré-processamento de dados para garantir a adequação aos algoritmos de clusterização e ferramentas de mineração de dados. A integração dos dados combinou seis arquivos fonte em um único conjunto. Em seguida, o conjunto de dados foi filtrado para incluir somente cidadãos malaios cuja renda familiar estivesse no grupo mais baixo. Vale ressaltar que a limpeza dos dados foi realizada para remover atributos com muitos valores ausentes, repetidos e redundantes, por isso atributos com poucos valores ausentes foram substituídos manualmente por valores específicos, resultando em um conjunto final de 16 atributos, a partir de um total de 53.

Muitos atributos podem dificultar o processo e análise dos dados, pois, o cálculo da distância por algoritmos de clusterização pode não ser eficaz para dados de alta dimensionalidade. Por isso, para resolver esse problema, foram aplicados métodos de seleção de atributos para identificar os mais relevantes para este estudo. Portanto, os atributos no conjunto de dados passaram pelo processo de seleção de atributos supervisionados usando técnicas como *random forest*. Após a execução, os atributos receberam pesos alocados conforme a sua relevância relativa e depois foram ordenados.

Por fim, o experimento abordado demostrou o desenvolvimento, com sucesso, de três modelos de clusterização usando *K-Means*, *BIRCH* e *DBSCAN*. A análise mostrou que o *K-Means* produziu o melhor desempenho quando comparado aos outros modelos. Em conclusão, baseado no estudo, a clusterização dos estudantes cuja renda familiar estivesse no grupo mais baixo, usando o *K-Means*, pode ser usado para reduzir o índice de desistência, uma vez que o artigo conseguiu identificar os níveis de desempenho durantes os estudos dos alunos.

3.2 *Density-Based Unsupervised Learning Algorithm to Categorize College Students into Dropout Risk Levels*

Este estudo aborda a necessidade de reduzir a evasão de estudantes no ensino superior, para assim, melhorar a qualidade educacional. Valles-Coral *et al.* (2022) propõe um modelo de acompanhamento acadêmico e emocional que utiliza técnicas de clusterização e aprendizado de máquina para agrupar estudantes universitários de acordo com seu nível de risco de evasão.

Nele foi tomado uma abordagem descritiva aplicada, focando na análise de dados existentes para desenvolver uma solução. Dados estes, adquiridos por meio de questionários psicológicos, validados e aprovados em estudos passados, que foram classificados usando algoritmos de aprendizado não supervisionado, onde foram aplicados na Universidade Nacional de San Martín, Peru, no semestre acadêmico de 2021.2 e possuíam uma amostra de 670 participantes, com uma população alvo de 5575 estudantes.

Em seguida, os dados coletados por meio do questionário no *chatbot*, teve todos os dados que não contribuíam com valores quantitativos para o modelo removidos, em seguida, foram escalados por meio de métodos de normalização, para providenciar os algoritmos com dados no mesmo formato e escala, e assim, visualiza-lós utilizando os algoritmos *DBSCAN*, *K-Means* e *HDBSCAN*. Ressaltando que, a escolha dos parâmetros foi baseada em uma análise criteriosa dos coeficientes de validação interna, com base em métricas de agrupamento como a *Silhouette*, *Calinski–Harabasz*, e *Davies–Bouldin* utilizados para comparar os algoritmos.

Por fim, os resultados demonstraram que o algoritmo *HDBSCAN* foi o mais robusto, categorizando eficazmente os alunos em cinco níveis de risco de evasão, validados por profissionais de saúde mental. Logo, a categorização dos estudantes providência uma melhor visualização do risco de desistência dos alunos que, caso combinado com um diagnóstico prematuro, permite a tomada de medidas corretivas. Além disso, o modelo resultante poderia ser aplicado em diferentes contextos, como também, ajustados para outros tipos de testes.

3.3 *A decision-making framework for school infrastructure improvement programs*

Fernández *et al.* (2023) propõe um framework focado na otimização da alocação de recursos limitados as escolas, com o intuito de maximar a qualidade da infraestrutura escolar em termos de funcionalidade e segurança. Assim, nele é abordada a crescente necessidade de melhorias na infraestrutura escolar, assim como a complexidade de alocar eficientemente recursos limitados para maximizar os benefícios. Por isso, este framework tem como base, para sua tomada de decisões, dados quantitativos e qualitativos, assim, visando fornecer uma metodologia sistemática para apoiar os gestores na tomada de decisões informadas sobre investimentos em infraestrutura das escolas.

Neste framework, composto por cinco módulos principais, primeiramente temos o módulo onde são coletados os dados detalhados sobre as condições das escolas, onde são inclusos aspectos estruturais, funcionais e de segurança. Ainda nesta etapa, são utilizadas ferramentas de levantamento de dados e sistemas de gestão de informações escolares. Por conseguinte, o módulo seguinte consiste na criação de um *Building Quality Index (BQI)*, gerado por meio de algoritmos de análise de dados de sistemas de pontuação, com o intuito de quantificar a qualidade da infraestrutura de cada escola, tendo como base os dados coletados.

No segundo módulo, após a finalização do cálculo do *(BQI)* e codificação de cada estrutura conforme a taxonomia de cada escola, elas são agrupadas em clusters com necessidades semelhantes onde as intervenções necessárias para cada grupo são definidas, por meio de algoritmos de clusterização e técnicas de análise estáticas para formar os grupos. Por fim, o último módulo envolve a otimização dos recursos disponíveis, com o desenvolvimento de modelos que ajudam a alocar os recursos de forma otimizada, visando maximizar a melhoria da qualidade no orçamento disponível.

Em conclusão, a flexibilidade do método e sua adaptabilidade a diferentes perfis de tomadas de decisão foram demonstradas por meio de uma análise de sensibilidade, que destacou a robustez do framework em diferentes cenários. Como também, os resultados do estudo demonstraram a eficácia do framework — por meio de um estudo de caso realizado na República Dominicana. Onde teve como resultado uma melhoria significativa da qualidade da infraestrutura escolar, tendo uma melhora de qualidade de até 3,54 vezes, quando comparada com programas de melhorias tradicionais. Além disso, o framework mostrou-se capaz, também, de economizar potencialmente até 65% do orçamento com um determinado limiar de melhoria de qualidade.

3.4 *Analysis of the socioeconomic impact due to COVID-19 using a deep clustering approach*

Durante o ano de 2022, muitos países enfrentaram dificuldades em medir o impacto da pandemia do COVID-19 em diversas áreas da sociedade. Por isso, Quintero *et al.* (2022) propôs uma abordagem de aprendizado de máquina híbrida para analisar o impacto socioeconômico da COVID-19 em diferentes distritos colombianos — também conhecidos como setores da Colômbia. A abordagem combinou um modelo de previsão da série temporal de infectados com um modelo de agrupamento não supervisionado, como os algoritmos de clusterização, *K-Means* e *k-medoids*, além das métricas para a validação dos resultados encontrados, sendo elas *Silhouette Coefficient* e *Davies-Bouldin Index*.

Este estudo teve como primeiro passo a criação de um modelo de previsão dos infectados. Para isso, foi utilizada uma ampla gama de dados, incluindo variáveis relativas ao comportamento da COVID-19, clima, economia, saúde, geografia e demografia. Que ao fim da coleta, foram submetidos a um processamento para garantir a qualidade e consistência. Após isso, esses dados foram usados para o treinamento do modelo de rede neural para prever, nos 7 dias seguintes, o número de casos de infecção por COVID-19 em cada setor da Colômbia.

Em seguida, para realizar o agrupamento dos dados, os valores previstos no passo anterior de cada setor foram combinados com outras variáveis socioeconômicas para assim formar um vetor de características. Então, esse vetor foi analisado utilizando os algoritmos de agrupamento — *K-Means* e *k-medoids*, para agrupar os setores conforme a similaridade socioeconômica. Por fim, os clusters gerados pelo modelo de agrupamentos foram analisados para identificar padrões e características comuns entre os departamentos pertencentes a cada grupo. Permitindo, assim, entender como diferentes fatores socioeconômicos foram influenciados pelo impacto da COVID-19 em cada região.

Em suma, o estudo demonstrou a eficácia da clusterização para acompanhar a evolução dos impactos socioeconômicos da COVID-19 em diferentes regiões, além de oferecer uma ferramenta valiosa para gestores públicos na formulação de políticas mais assertivas. Ao identificar padrões e características comuns entre diferentes grupos de departamentos, o modelo contribui para a definição de estratégias mais direcionadas e eficazes. Além disso, também exemplifica como a análise da evolução dos clusters pode auxiliar na compreensão das transformações de diferentes variáveis ao longo do tempo e seus impactos.

3.5 Análise Comparativa

Conforme visto no Quadro 1, este trabalho propõe uma análise aprofundada da evolução da qualidade do ensino nas escolas públicas brasileiras, utilizando técnicas de clusterização para identificar padrões e relações entre indicadores educacionais. Enquanto estudos como os de Nafuri *et al.* (2022) e Valles-Coral *et al.* (2022) focaram na aplicação de algoritmos como K-Means e HDBSCAN para compreender as causas da evasão universitária, este trabalho busca compreender como os indicadores educacionais variam entre municípios ao longo do tempo, oferecendo uma visão mais ampla e dinâmica da qualidade do ensino.

A metodologia adotada também se distingue pela aplicação de métricas de qualidade da clusterização em um contexto diferente. Enquanto Nafuri *et al.* (2022) utilizou o coeficiente de silhueta para avaliar a formação de grupos de estudantes e Valles-Coral *et al.* (2022) empregou o índice Calinski-Harabasz para prever padrões de evasão, este estudo aplicará o coeficiente de silhueta junto com o método do cotovelo para analisar indicadores educacionais, permitindo uma avaliação detalhada das suas disparidades e da evolução da qualidade do ensino.

Além disso, este trabalho compartilha semelhanças metodológicas com a pesquisa de Quintero *et al.* (2022), que analisou os efeitos da COVID-19 na economia por meio do algoritmo K-medoids e do índice Davies-Bouldin. Da mesma forma, este estudo utilizará técnicas de clusterização, incluindo o *K-Means*, para acompanhar a evolução das escolas ao longo do tempo e identificar padrões que possam orientar políticas públicas. Contudo, diferencia-se da abordagem proposta por Fernández *et al.* (2023), que desenvolveu um framework para detectar escolas com deficiência de recursos financeiros com base em dados estruturais, ao focar na análise de indicadores educacionais e sua evolução temporal.

Quadro 1 – Comparação dos trabalhos relacionados

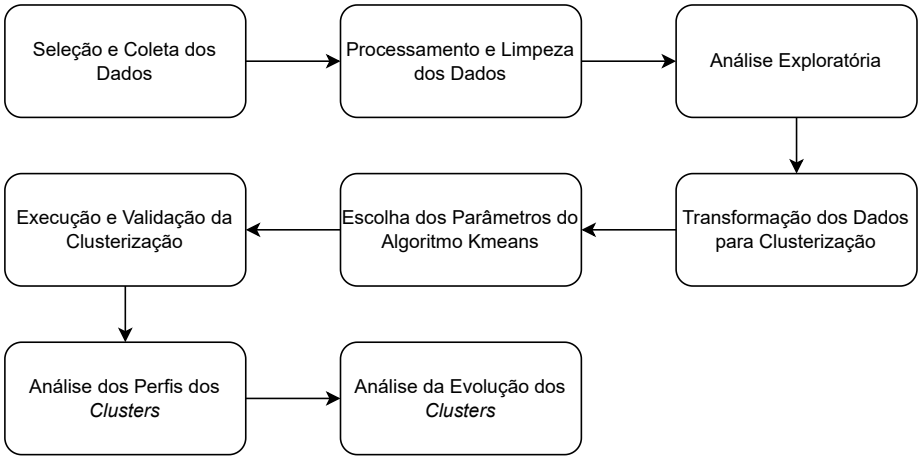
Artigos	Algoritmo	Métricas Principais	Fonte dos Dados	Domínio	Objetivo
Nafuri *et al.* (2022)	*K-Means*	Coeficiente de Silhueta	Estudantes universitários	Universitário	Ajudar o governo reduzir a evasão universitária.
Valles-Coral *et al.* (2022)	*HDBSCAN*	*Calinski–Harabasz*	*Online chatbot* com estudantes universitários	Universitário	Criar modelo de previsão da evasão dos estudantes universitários.
Fernández *et al.* (2023)	Customizado	*Building quality Index*	Dados estruturais e financeiros de escolas	Escolar e Governamental	Criar framework para identificar escolas carentes de investimento.
Quintero *et al.* (2022)	*K-medoids*	*Davies–Boulding*	Indicadores regionais	Saúde e Governamental	Identificar o impacto economico da COVID-19.
Este trabalho	*K-Means*	Coeficiente de Silhueta	Indicadores educacionais e demográficos	Escolar e Governamental	Analisar e entender as relações entre indicadores educacionais no Brasil.

Fonte: Elaborado pelo autor.

4 METODOLOGIA

Neste capítulo são apresentados os detalhes sobre a seleção dos dados utilizados, os métodos de pré-processamento aplicados para garantir a qualidade e a consistência dos dados, a escolha e aplicação dos algoritmos de clusterização, bem como a avaliação dos clusters resultantes.

Figura 4 – Metodologia da pesquisa



Fonte: Elaborada pelo autor.

4.1 Seleção e Coleta dos Dados

A análise concentrou-se no ensino fundamental inicial, dado que os indicadores do INEP são segmentados entre ensino fundamental inicial, ensino fundamental final e ensino médio. Além disso, os dados foram selecionados com base no agrupamento por escola, visto que nem todas as informações estavam disponíveis em nível municipal. Essa abordagem não apenas assegurou maior precisão, como também viabilizou, futuramente, a agregação dos dados em municípios para análises territoriais mais aprofundadas.

Para garantir a qualidade e relevância das informações utilizadas, os dados foram obtidos do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). No entanto, a heterogeneidade temporal e a diferença nos formatos de agrupamento inviabilizavam a utilização conjunta de todos os dados disponíveis para o processo de clusterização. Assim, optou-se pela seleção de dados referentes aos de 2015, 2019 e 2021, ressaltando que alguns dos indicadores são atualizados bienalmente.

Primeiramente, o ano de 2015 foi incluído por coincidir com a implementação do Plano de Ações Articuladas – PAR3, um programa do Ministério da Educação voltado para a melhoria da qualidade da educação básica. Pois, a inclusão desse ano possibilita analisar eventuais impactos iniciais dessas medidas nos indicadores educacionais. Enquanto o ano de 2019 foi selecionado por ser um dos últimos anos antes da pandemia de COVID-19, permitindo a observação dos indicadores em um cenário de normalidade. Por outro lado, 2021 representa o período afetado pela pandemia além de ser o mais recente disponivel no momento da pesquisa.

Quadro 2 – Indicadores coletados

Nome	**Linhas : Colunas**		
	2015	**2019**	**2021**
Adequação da Formação Docente (AFD)	182.606 : 44	177.714 : 44	175.593 : 44
Complexidade de Gestão da Escola (ICG)	186.441 : 10	180.610 : 10	178.370 : 10
Esforço Docente (IED)	142.695 : 33	134.152 : 33	131.804 : 33
Índice de Desenvolvimento da Educação Básica (IDEB)	64.905 : 124	64.905 : 124	64.905 : 124
Média de Alunos por Turma (ATU)	179.308 : 31	176.386 : 31	174.179 : 31
Média de Horas-aula diária (HAD)	179.305 : 31	152.509 : 30	151.777 : 30
Indicador de Nível Socioeconômico (INSE)	63.258 : 12	68.868 : 20	69.820 : 22
Percentual de Docentes com Curso Superior (DSU)	185.845 : 20	180.331 : 19	178.126 : 19
Regularidade do Corpo Docente (IRD)	169.825 : 10	166.177 : 10	165.853 : 10
Taxas de Distorção Idade-série (TDI)	179.380 : 27	132.428 : 26	130.112 : 26
Taxas de Não-resposta (TNR)	140.865 : 27	132.458 : 27	130.129 : 27
Taxas de Rendimento Escolar (TX)	140.865 : 63	132.458 : 63	130.129 : 63

Fonte: Elaborado pelo autor.

Por fim, os dados coletados, conforme apresentado no Quadro 2, abrangem aspectos pedagógicos, socioeconômicos e de desempenho escolar. Indicadores como o AFD avaliam a qualificação dos professores, enquanto o IDEB, da Seção 2.1.1.1, mede a qualidade do ensino. Além disso, variáveis como ATU e HAD analisam fatores que podem impactar o desempenho dos alunos, enquanto o INSE e o TDI, na Seção 2.1.1.2, fornecem informações sobre o contexto educacional e os desafios enfrentados pelos estudantes.

4.2 Processamento e Limpeza dos Dados

A qualidade e eficácia da análise de dados dependem significativamente da aplicação de técnicas adequadas de pré-processamento. Por isso, nesta etapa, foram realizadas a limpeza e tratamento dos dados, incluindo a identificação e tratamento de valores ausentes e inconsistências, como a presença de diferentes tipos de dados em uma mesma coluna. Uma vez que, tais inconsistências, se não tratadas, poderiam comprometer a validade da análise.

Após a seleção dos dados, diversos ajustes foram realizados para garantir a consistência, integridade e precisão das informações antes do processo de análise. Algumas colunas apresentavam inconsistências nos dados, como a mistura de valores numéricos e textuais, causada por valores ausentes representados por "ND"e outras strings. Para corrigir esse problema, os valores faltantes foram identificados e substituídos pela representação adequada de NaN. Além disso, era necessário garantir que os dados estivessem uniformemente formatados para que o processo de junção entre os indicadores fosse realizado de forma eficiente. Também foi preciso remover informações duplicadas.

Primeiramente, para garantir a consistência e a integridade dos dados, foi essencial manter o tipo das colunas que seriam utilizadas para o agrupamento dos dados, assim como garantir o padrão dos nomes de cada uma, a fim de evitar problemas de tipagem. Para isso, foi criado um dicionário para cada indicador de todos os anos utilizados, onde foram armazenados tanto os nomes das colunas como os tipos de dados correspondentes. Esse dicionário foi fundamental para assegurar que as colunas relevantes fossem corretamente identificadas e tratadas de acordo com seu tipo específico, prevenindo erros e inconsistências no processo de junção dos dados.

Código-fonte 1 – Função de limpeza dos dados

```
def clean_dataframe(dataframe, specific_columns, prefix,
    columns_type):
    dataframe = replace_invalid_values(dataframe)
    dataframe = convert_type(dataframe, specific_columns)
    dataframe = dataframe.astype(columns_type)

    dataframe = filter_columns(dataframe, specific_columns)
    dataframe = dataframe.dropna(
        how="all",
        subset=[col for col in dataframe.columns if col not
            in specific_columns],
    )

    dataframe = add_prefix(dataframe, specific_columns,
        prefix)

    return dataframe
```

Fonte: Elaborado pelo autor.

Como pode ser visto na função presente no Código Fonte 1, o *dataframe*, uma estrutura que organiza os dados em uma tabela bidimensional de linhas e colunas, contendo os dados dos indicadores é passado, com as colunas de identificação que devem ser mantidas inalteradas, o prefixo de cada indicador e os tipo das colunas. Na primeira etapa, é chamada a função *replace_invalid_values* para substituir os valores de texto que não deveriam estar nas colunas numéricas. Em seguida, é chamada a função *convert_type* para substituir as vírgulas por pontos e transformar os valores em formato numérico de ponto flutuante.

Logo após, o dicionário criado para cada indicador é utilizado para garantir a conversão correta do tipo de dados de cada coluna e as colunas relacionadas exclusivamente ao Ensino Fundamental Inicial (1° ao 5° ano) são filtradas. Em seguida, as linhas compostas apenas por valores ausentes são removidas, ignorando as colunas de identificação. Por fim, um prefixo correspondente ao nome de cada indicador é adicionado às colunas, o que facilita a identificação de cada indicador após sua junção. Vale ressaltar que, devido a uma diferença no cálculo da média do INSE ao longo dos anos, a coluna de agrupamento foi convertida para um valor numérico, facilitando a comparação dos diferentes INSE ao longo do tempo.

4.3 Análise Exploratória

A análise exploratória dos dados é uma etapa fundamental para a compreensão do conjunto de dados antes da aplicação de algoritmos de clusterização. Seu principal objetivo é identificar padrões, detectar anomalias, testar hipóteses e verificar suposições que possam impactar a modelagem subsequente. Além de facilitar a interpretação dos dados, essa etapa permite a detecção de inconsistências e minimiza vieses, garantindo a qualidade e confiabilidade dos resultados.

Para isso, foram empregadas técnicas de visualização e estatísticas descritivas que auxiliaram na compreensão da estrutura dos dados e forneceram percepções iniciais para as próximas etapas da análise. Como os algoritmos de clusterização são altamente sensíveis à qualidade dos dados de entrada, a presença de ruídos pode comprometer a formação dos grupos e impactar a interpretação dos resultados. Além disso, a distribuição das variáveis foi analisada para identificar possíveis necessidades de transformação ou normalização. Inicialmente, os dados foram carregados utilizando a biblioteca pandas e, em seguida, empregados na geração de um relatório exploratório com a ferramenta *YData Profiling* YData (2024). Esse relatório fornece uma visão geral das variáveis presentes no conjunto de dados.

Além disso, o relatório oferece uma análise detalhada da distribuição de cada variável, tanto numérica quanto categórica. Foram apresentados histogramas, gráficos de densidade e tabelas de frequência, permitindo identificar assimetrias, valores extremos e padrões inesperados. Além disso, foi fornecida uma análise de correlação entre as variáveis numéricas, o que é essencial para detectar relações fortes que podem influenciar na etapa de clusterização, principalmente com o indicador principal de análise que será o IDEB.

Outro ponto relevante do relatório foi a detecção de valores ausentes e extremos, que auxiliou na identificação de possíveis problemas nos dados que precisavam ser tratados antes da aplicação do modelo, pois o relatório apresentava a quantidade e a proporção deles em cada variável, possibilitando sua remoção ou ajuste, caso necessário. Por isso, a análise exploratória usando *YData Profiling* do YData (2024) garantiu que o conjunto de dados estava limpo e pronto para ser utilizado no algoritmo de clusterização, além de fornecer percepções valiosas sobre a estrutura dos dados. O relatório gerado forneceu uma visão clara e abrangente, permitindo tomar decisões informadas sobre como proceder nas próximas etapas da análise.

Código-fonte 2 – Configuração do Ydata

```
custom_config = Settings(
    title="Ensino Fundamental Inicial Especifico 2021",
    interactions={
        "continuous": False,
    },
    missing_diagrams={
        "bar": False,
        "matrix": False,
        "heatmap": False,
    },
)
```

Fonte: Elaborado pelo autor.

Para evitar a geração de informações redundantes ou irrelevantes e otimizar o tempo de processamento devido ao tamanho do conjunto de dados, foi criada uma configuração personalizada do YData Profiling, conforme detalhado no Código-Fonte 2, permitindo ajustar a análise conforme as necessidades específicas do conjunto de dados. Além disso, a configuração personalizada facilitou a replicação do relatório para os outros anos, garantindo consistência e agilidade na análise dos novos conjuntos de dados.

No Código-Fonte 2 foi definido um título para o relatório gerado, o que facilitou a identificação do conjunto de dados analisado para diferentes anos. O parâmetro que controla a análise de interações entre variáveis foi ajustado para falso, desativando a análise de interações para variáveis contínuas. Os diagramas de valores ausentes foram desativados, pois a ferramenta oferece métodos alternativos para identificar dados faltantes que exigem menos processamento computacional. Dessa forma, o tempo de processamento foi drasticamente reduzido, além de evitar a geração de gráficos desnecessários.

4.4 Transformação dos Dados para Clusterização

A etapa de transformação dos dados é fundamental para garantir a eficácia da clusterização, pois assegura que os indicadores estejam devidamente preparados para a análise. Neste trabalho, este processo compreende o agrupamento dos dados, a seleção das variáveis relevantes, a normalização e a inversão da monotonicidade de determinados indicadores.

Código-fonte 3 – Função de agrupamento

```
def agrupamento_media_ponderada(data, colunas_agrupar):
    soma_pesos = grupo['QT_MAT_FUND_AI'].sum()
    indicadores_municipais = {}

    for col in colunas_agrupar:
        indicadores_municipais[col] = (data[col] * grupo['
           QT_MAT_FUND_AI']).sum() / soma_pesos

    return indicadores_municipais
```

Fonte: Elaborado pelo autor.

Inicialmente, os dados, que estavam organizados ao nível de escola, foram reagrupados para o nível municipal, permitindo uma análise mais abrangente de cada localidade. Esse agrupamento foi realizado por meio da média ponderada, utilizando como peso o número de matrículas total dos anos iniciais, em cada escola. Conforme o Código-fonte 3, a função *agrupamento_media_ponderada* recebe os dados e as colunas a serem agrupadas, calcula o total de matrículas e, para cada indicador selecionado, multiplica os valores pelo número de matrículas antes de dividir pelo total de alunos do município. Assim, assegurando que escolas com maior número de alunos tivessem uma representatividade proporcionalmente adequada.

Código-fonte 4 – Remoção de colunas não utilizadas

```
padroes_remocao = ['TX_', 'TNR_', 'INSE_']
removed_columns = [col for col in agrupado_tratado.columns
   if col.startswith(tuple(padroes_remocao))]

save_removed = agrupado_tratado[['CO_MUNICIPIO'] +
   removed_columns]

agrupado_tratado.drop(columns=removed_columns,inplace=True)
```

Fonte: Elaborado pelo autor.

Após o agrupamento, foi conduzida uma triagem dos indicadores disponíveis. Conforme o Código-fonte 4 algumas variáveis foram excluídas por não contribuírem diretamente para a análise ou por introduzirem informações já contempladas em outros indicadores. O INSE foi descartado por refletir a renda familiar dos alunos, sem estar diretamente relacionado ao desempenho escolar. Da mesma forma, a taxa de não resposta foi removida, pois sua inclusão não agregaria valor à clusterização. Além disso, a taxa de rendimento foi excluída por estar incorporada no cálculo do IDEB, evitando redundância e possível sobreposição de informações. No entanto, esses dados foram armazenados para uma análise futura após a clusterização.

Com as variáveis, do Quadro 3, devidamente selecionadas, procedeu-se à normalização dos dados, uma etapa essencial para garantir que todas as variáveis possuíssem a mesma escala e, assim, contribuíssem de maneira equitativa para a clusterização. Para isso, utilizou-se a técnica de normalização Min-Max, que transforma os valores das variáveis para um intervalo entre 0 e 1, preservando sua distribuição relativa. Essa padronização evitou que variáveis com magnitudes distintas tivessem impacto desproporcional sobre a formação dos clusters.

Por fim, foi realizada a análise da monotonicidade dos indicadores, considerando que algumas variáveis, listadas no Quadro 3, apresentavam uma relação inversa com a qualidade educacional. Enquanto a maioria das variáveis possuía uma interpretação direta — em que valores mais altos correspondiam a melhores condições educacionais —, outras, como a taxa de distorção idade-série, seguiam uma lógica inversa. Para garantir uma interpretação uniforme e coerente, foi aplicada a inversão da monotonicidade. Esse ajuste foi realizado após a normalização Min-Max, assegurando que a transformação preservasse a proporcionalidade dos dados e mantivesse a consistência da escala. Logo, todos os indicadores passaram a seguir a mesma direção interpretativa, facilitando a análise comparativa e evitando distorções nos resultados.

Quadro 3 – Lista de uso dos indicadores coletados

Indicador	**Monotonicidade**	**Clusterizado**
Adequação da Formação Docente (AFD)	Mantido	Usado
Complexidade de Gestão da Escola (ICG)	Mantido	Usado
Esforço Docente (IED)	Invertido	Usado
Índice de Desenvolvimento da Educação Básica (IDEB)	Mantido	Usado
Média de Alunos por Turma (ATU)	Invertido	Usado
Média de Horas-aula diária (HAD)	Mantido	Usado
Indicador de Nível Socioeconômico (INSe)	Mantido	Excluído
Percentual de Docentes com Curso Superior (DSU)	Mantido	Usado
Regularidade do Corpo Docente (IRD)	Mantido	Usado
Taxas de Distorção Idade-série (TDI)	Invertido	Usado
Taxas de Não-resposta (TNR)	Mantido	Excluído
Taxas de Rendimento Escolar (TX)	Mantido	Excluído

Fonte: Elaborado pelo autor.

Dessa forma, o processo de transformação dos dados foi conduzido de maneira criteriosa, desde o agrupamento por município até a preparação final dos indicadores, por meio da seleção de variáveis, normalização e ajuste de monotonicidade. Essas etapas foram essenciais para garantir a qualidade da clusterização e permitir a identificação de padrões significativos no desempenho educacional dos municípios analisados.

4.5 Escolha dos Parâmetros do Algoritmo

A escolha do número ideal de clusters no algoritmo K-Means é um passo fundamental para garantir a qualidade da clusterização e a representatividade dos agrupamentos. Definir um valor inadequado pode resultar em clusters pouco coesos ou em divisões arbitrárias dos dados, comprometendo a análise. Assim, para determinar o número mais adequado de clusters, foram utilizadas duas abordagens complementares, o método do cotovelo, baseado na métrica de distorção, e o Coeficiente de Silhueta.

O método do cotovelo foi aplicado utilizando a métrica de distorção. O procedimento consistiu em executar o algoritmo K-Means para diferentes valores de K (número de clusters) e calcular a distorção para cada um deles. A expectativa era que, à medida que K aumentasse, a distorção diminuísse, pois os pontos ficariam mais próximos de seus respectivos centroides. Portanto, após um certo ponto, a taxa de redução da distorção começava a diminuir significativamente, formando um "cotovelo"no gráfico. Esse ponto de inflexão indica o valor adequado para o número de clusters, equilibrando a complexidade do modelo e a qualidade da clusterização.

Além do método do cotovelo, foi utilizada a métrica do Coeficiente de Silhueta para avaliar a qualidade da clusterização. O coeficiente de silhueta mede a separação entre os clusters e a compactação dos pontos dentro de cada grupo, comparando a distância média de um ponto para os demais pontos do mesmo cluster com a distância média para os pontos do cluster mais próximo. Esse coeficiente varia de -1 a 1, sendo que valores próximos de 1 indicam clusters bem definidos, enquanto valores negativos sugerem agrupamentos incorretos. Para essa análise, o K-Means foi executado com diferentes valores de K, e o coeficiente de silhueta médio foi calculado para cada caso. O valor de K que maximizou essa métrica foi considerado um indicativo adicional da quantidade ideal de clusters.

Dessa forma, a metodologia adotada para a definição dos parâmetros do K-Means consistiu na aplicação do método do cotovelo para identificar o ponto de inflexão na curva de distorção e na análise do Coeficiente de Silhueta para validar a coesão dos clusters. A combinação dessas abordagens permitiu definir um número de clusters mais robusto, minimizando distorções e melhorando a separação entre grupos.

4.6 Execução e Validação da Clusterização

A execução e validação da clusterização são etapas importantes no processo de análise de dados, visando garantir a qualidade e a interpretação adequada dos resultados obtidos. Após a definição do algoritmo e dos parâmetros, o próximo passo foi a execução da clusterização dos dados. Após a execução da clusterização, foi importante avaliar a qualidade e a validade dos clusters obtidos. A análise visual dos clusters é uma ferramenta poderosa para entender a estrutura dos dados e validar a clusterização.

Gráficos de dispersão e visualizações multidimensionais podem ser úteis para representar os clusters em um espaço de menor dimensionalidade e identificar padrões ou agrupamentos naturais nos dados. Por isso, conforme o Código-fonte 5, foi utilizado o PCA e o t-SNE para reduzir a dimensionalidade dos dados e explorar as diferentes formas de agrupamento possíveis. Essas visualizações permitiram observar como os dados se organizam de maneira distinta em cada técnica, evidenciando diferentes padrões nos clusters gerados.

Código-fonte 5 – Exemplo de plotagem do gráfico usando PCA

```
pca_all = PCA(n_components=2, random_state=42)
pca_data_all = pca_all.fit_transform(efi[numerical_columns])

plt.figure(figsize=(6, 4))
sns.scatterplot(x=pca_data_all[:, 0], y=pca_data_all[:, 1], hue=efi['Cluster'], palette='viridis')
plt.title('Kmeans com PCA')
plt.legend(title='Cluster')
plt.savefig("pca_2021.pdf", format="pdf", bbox_inches="tight")
plt.show()
```

Fonte: Elaborado pelo autor.

4.7 Análise dos Perfis dos *Clusters*

A análise dos perfis dos *clusters* é uma etapa essencial no processo de clusterização, pois ela permite interpretar e entender as características que definem cada grupo identificado. Uma vez que os dados forem agrupados em *clusters*, é necessário analisar detalhadamente os perfis de cada um deles para extrair informações significativas, identificar padrões e, potencialmente, formular hipóteses sobre as relações entre as variáveis.

A principal importância da análise dos perfis dos *clusters* reside na sua capacidade de transformar dados brutos em informações. Essa análise permitiu identificar quais características são mais comuns em cada *cluster*, facilitando a compreensão dos fatores que diferenciam os grupos. Além disso, ao interpretar os perfis dos *clusters*, foi possível validar se os grupos formados fazem sentido no contexto do estudo e se estão alinhados com os objetivos do projeto. Por isso, neste trabalho, que utiliza técnicas de clusterização para entender perfis educacionais, a análise dos perfis dos *clusters* pode revelar padrões importantes, como a correlação entre níveis educacionais e indicadores socioeconômicos.

Com o intuito de caracterizar e diferenciar os *clusters*, optou-se por calcular um índice numérico capaz de mensurar o perfil de cada um. Para isso, foi calculada a média das entidades dentro de cada *cluster*. Em seguida, foi calculada a média geral dos *cluster* sendo ordenadas de maneira decrescente. Esse processo é essencial para avaliar a posição relativa dos clusters e para a atribuição de um perfil a cada um, com a classificação alfabética começando com a letra A para o cluster com a melhor pontuação.

Após a pontuação, foi realizada uma descrição estatística de cada *cluster*, que incluiu a análise das médias, medianas, variâncias e distribuições das variáveis dentro de cada grupo. Essa descrição quantitativa forneceu uma visão geral das características predominantes em cada *cluster*. Em seguida, foi realizada uma análise comparativa entre os *clusters*, destacando as diferenças e semelhanças entre eles, sendo acompanhada por gráficos que ilustrarão as distribuições das variáveis dentro de cada grupo. Por fim, foi realizada uma interpretação qualitativa dos resultados, relacionando os perfis identificados com o contexto do estudo.

4.8 Análise da Evolução dos *Clusters*

A análise da evolução dos clusters ao longo dos anos é uma etapa crítica para entender como os grupos identificados em uma análise de clusterização se transformam com o tempo, ao permitir observar padrões de evolução, continuidade ou mudança nos perfis dos clusters. Onde sua principal importância reside na capacidade de revelar dinâmicas temporais que podem não ser aparentes em uma análise estática. Assim, permitindo identificar diferentes tipos de mudanças, como a criação de um novo conjunto ou migração de dados entre conjuntos existentes, que os *clusters* gerados nos anos seguintes podem sofrer.

Por isso, essa análise foi realizada comparando-se os clusters identificados em diferentes períodos. Inicialmente, os dados foram segmentados por ano, e a clusterização foi realizada separadamente para cada ano de estudo. Isso permitiu a identificação de clusters específicos para cada período, respeitando as particularidades de cada ano. Após a obtenção dos clusters anuais, foi feita uma comparação detalhada entre os clusters de diferentes anos. Essa comparação envolveu a análise da composição dos clusters, verificando se os mesmos grupos de escolas, regiões ou indivíduos permaneceram juntos ao longo dos anos, ou se houve uma reconfiguração dos clusters.

Para cada cluster, foi traçada uma trajetória temporal, mostrando como o grupo evoluiu ao longo dos anos. A partir das comparações e das trajetórias dos clusters, foram identificados padrões de evolução, como a fusão de clusters, a fragmentação de grupos, ou a estabilidade de certos clusters ao longo dos anos. Esses padrões foram analisados para entender o que pode ter originado esses novos padrões. Assim, podendo mostrar como a qualidade de ensino nos grupos de escolas mudam ao longo do tempo, revelando melhorias ou declínios nos indicadores educacionais. Essas percepções são fundamentais para a formulação de políticas públicas e para a compreensão de fatores que influenciam a educação ao longo dos anos.

5 RESULTADOS

Neste capítulo, são apresentados os resultados obtidos. Na Seção 5.1, são detalhados os achados da análise exploratória. A seguir, a Seção 5.2 aborda a escolha dos parâmetros do algoritmo K-means. Na Seção 5.3, é realizada uma análise dos perfis dos clusters identificados. Além disso, a Seção 5.4 explora os perfis dos clusters. Por fim, na Seção 5.5, é apresentada uma análise da evolução de cada cluster ao longo dos anos, incluindo as migrações das suas entidades.

5.1 Análise Exploratória

Durante a análise exploratória, utilizou-se inicialmente o *Ydata Profiling* do YData (2024) para gerar uma documentação dos conjuntos de dados finais referentes aos indicadores dos anos de 2015, 2019 e 2021. Essa documentação facilita a identificação de inconsistências nos tipos de dados, a detecção de valores extremos e a observação de padrões, além de fornecer um panorama detalhado das principais características dos dados, conforme visto na Tabela 4

Quadro 4 – Estatísticas do conjunto de dados de indicadores de cada ano

Agrupamento 2015

Estatística	**Dados**
Número de Colunas	80
Número de Linhas	142.033
Células em Branco	40,2%
Colunas numericas	71
Colunas categoricas	7
Colunas de texto	2

Agrupamento 2019

Estatística	**Dados**
Número de Colunas	88
Número de Linhas	134.713
Células em Branco	36,7%
Colunas numericas	79
Colunas categoricas	7
Colunas de texto	2

Agrupamento 2021

Estatística	**Dados**
Número de Colunas	88
Número de Linhas	132.962
Células em Branco	37,1%
Colunas numericas	79
Colunas categoricas	7
Colunas de texto	2

Fonte: Elaborado pelo autor.

O Quadro 4 apresenta as estatísticas dos conjuntos de dados para os anos de 2015, 2019 e 2021. Observa-se que o número de colunas aumentou de 80 para 88 entre 2015 e 2019, mantendo-se constante em 2021. O número de linhas, por sua vez, diminuiu ao longo dos anos, passando de 142.033 em 2015 para 132.962 em 2021. A porcentagem de células em branco também apresentou uma redução, de 40,2% em 2015 para 37,1% em 2021, indicando uma melhoria na completude dos dados. Quanto à composição das colunas, houve um aumento no número de colunas numéricas de 71 para 79, enquanto as colunas categóricas e de texto permaneceram constantes, com 7 e 2, respectivamente. Assim, o uso da ferramenta *Ydata Profiling* forneceu informações valiosos sobre a constituição dos dados, permitindo identificar e corrigir possíveis problemas, como a presença de valores, antes de avançar para a aplicação de métodos analíticos, como a clusterização.

Quadro 5 – Estatísticas do Índice de Desenvolvimento da Educação Básica

IDEB 2015

Estatística	Dados
Média	5,24
Mediana	5,3
Desvio Padrão	1,12
Coeficiente de Variação	0,21
IDEB Máximo	9,8
IDEB Mínimo	0,8

IDEB 2019

Estatística	Dados
Média	5,59
Mediana	5,7
Desvio Padrão	1,13
Coeficiente de Variação	0,20
IDEB Máximo	9,8
IDEB Mínimo	1,4

IDEB 2021

Estatística	Dados
Média	5,51
Mediana	5,6
Desvio Padrão	0,98
Coeficiente de Variação	0,18
IDEB Máximo	9,9
IDEB Mínimo	0,6

Fonte: Elaborado pelo autor.

Analisando o Quadro 5 e focando primeiramente no IDEB por ser um dos principais indicadores que avalia a qualidade do ensino no Brasil, e por ser capaz de permitir identificar tendências de melhoria, estabilidade ou desigualdades no sistema educacional, a média passou de 5,24 em 2015 para 5,59 em 2019, mantendo-se em 5,51 em 2021. A mediana seguiu um padrão semelhante, atingindo 5,7 em 2019 e 5,6 em 2021, refletindo algum avanço no desempenho educacional. Além disso, a redução do desvio padrão, de 1,13 para 0,98 entre 2019 e 2021, indica menor dispersão dos resultados e maior consistência. O coeficiente de variação também caiu de 0,21 para 0,18, reforçando a homogeneidade dos dados.

Quanto aos valores extremos, o mínimo variou de 0,8 em 2015 para 1,4 em 2019, mas caiu para 0,6 em 2021, provavelmente fruto do impacto da pandemia por conta da COVID-19. Já o máximo manteve-se estável, em torno de 9,8 e 9,9. Essas análises sugerem que, apesar de flutuações, houve uma tendência de melhoria e estabilização no desempenho do IDEB ao longo dos anos, com uma redução na variabilidade dos dados conforme demonstrado no Quadro 5.

A Figura 5 apresenta os histogramas do IDEB para os anos de citados e permite uma análise visual da distribuição dos dados em cada período. Primeiramente, na Figura 5a, observa-se uma distribuição com uma concentração significativa de valores em torno da média de 5,25, com uma cauda mais longa à esquerda, indicando a presença de escolas com desempenho abaixo da média. A frequência máxima atinge aproximadamente 2500 escolas, sugerindo uma alta densidade de instituições com resultados próximos à mediana de 5,3.

Contudo, na Figura 5b, o histograma mostra uma ligeira mudança na distribuição, com um aumento na frequência de escolas com IDEB mais alto, refletindo a melhoria na média para 5,60 e na mediana para 5,7. A forma da distribuição permanece semelhante à de 2015, mas com uma redução na quantidade dos valores do IDEB entre 2 e 4, indicando uma diminuição no número de escolas com desempenho muito baixo. A frequência máxima continua próxima de 2500 escolas, mas com um deslocamento para valores mais altos do IDEB.

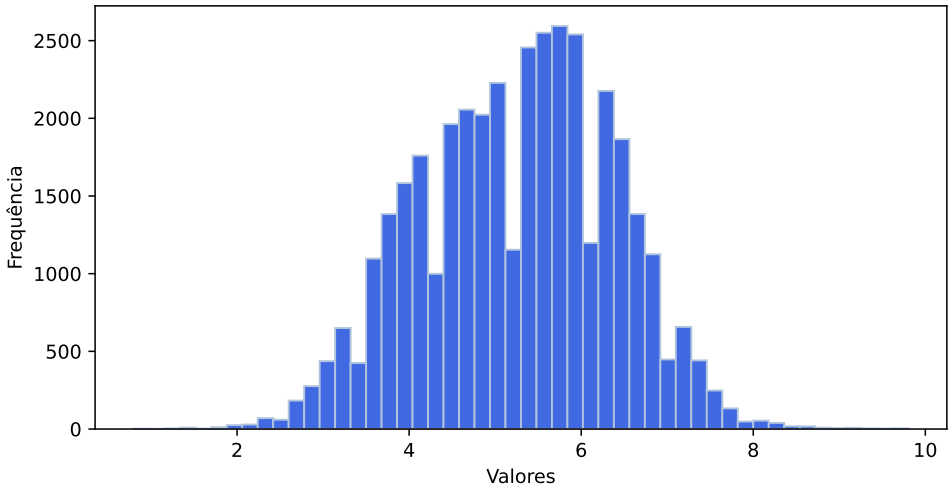
Já em 2020, na Figura 5c, o histograma revela uma distribuição mais concentrada em torno da média de 5,51 e da mediana de 5,6, com uma redução na dispersão dos dados, conforme indicado pelo menor desvio padrão de 0,98. A frequência máxima atinge aproximadamente 2500 escolas, mas com uma forma mais simétrica e menos assimétrica do que nos anos anteriores. Isso sugere uma maior homogeneidade nos resultados, com menos escolas apresentando desempenho extremamente baixo ou alto.

Comparando os três anos, é possível notar uma tendência de melhoria no desempenho médio e mediano do IDEB, com uma redução na variabilidade dos dados e uma maior concentração de escolas em torno dos valores centrais. Essa evolução indica um progresso na qualidade da educação básica, embora ainda existam desafios que precisam ser superados para que tais avanços se consolidem de maneira equitativa em todo o território nacional, especialmente em relação às escolas com desempenho abaixo da média. Além disso, o efeito da pandemia de COVID-19 se reflete na diminuição dos resultados apresentados e reforçam a importância de políticas públicas que priorizam uma redistribuição de recursos para escolas mais fragilizadas.

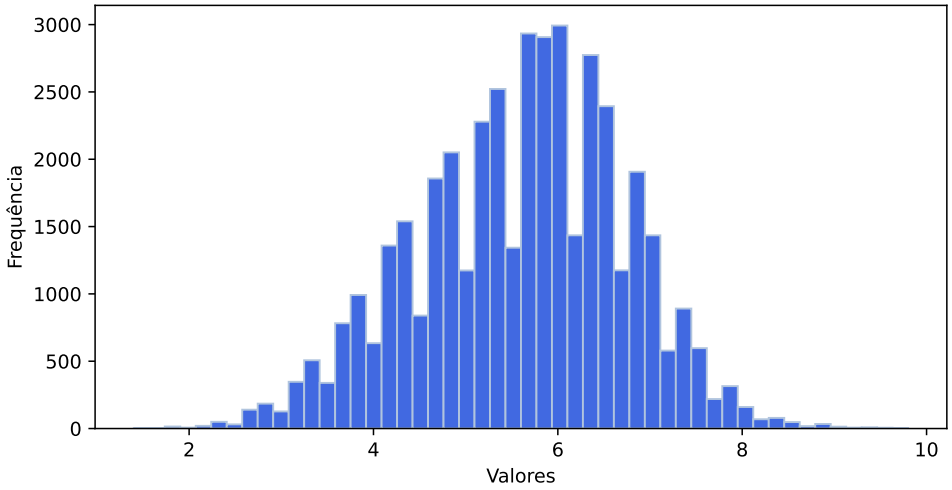
Ao analisar a correlaçao do IDEB, é possível observar que o mesmo se relaciona tanto com a Taxa de Rendimento como a Taxa de Distorção Idade-Série, uma vez que a progressão dos alunos ao longo dos anos letivos influencia diretamente o resultado desse índice. Esse padrão evidencia a importância da permanência e do sucesso acadêmico dos estudantes para a melhoria do desempenho educacional geral. Além disso, a taxa de distorção idade-série, que mede a proporção de alunos com idade superior à esperada para determinada série, também se mostra um fator relevante, pois altos índices de distorção indicam dificuldades no fluxo escolar, impactando negativamente o IDEB.

Outro aspecto relevante é a correlação do IDEB com a projeção desse próprio indicador e com o Índice de Nível Socioeconômico (INSE). A projeção do IDEB busca estimar o desempenho futuro com base em tendências históricas e fatores associados à aprendizagem e à permanência escolar, sendo um parâmetro essencial para o planejamento educacional. A relação entre o IDEB projetado e o INSE reforça a influência do contexto socioeconômico na qualidade do ensino, uma vez que escolas localizadas em regiões com melhores condições socioeconômicas tendem a apresentar tanto um IDEB mais alto quanto projeções mais otimistas. Essa conexão destaca a necessidade de políticas públicas que reduzam desigualdades educacionais, garantindo que escolas em contextos mais vulneráveis recebam suporte adequado para elevar seus índices de desempenho e, consequentemente, melhorar a equidade no sistema educacional.

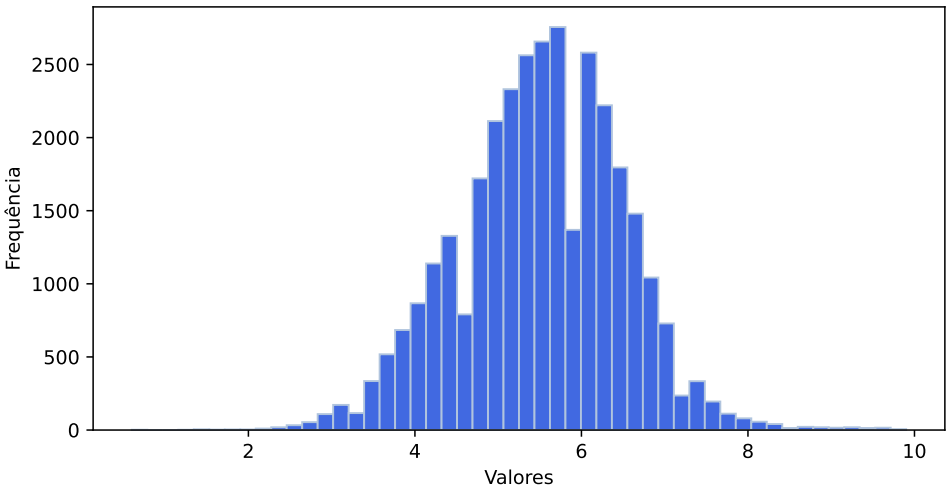
Figura 5 – Histograma do Índice de Desenvolvimento da Educação Básica



(a) IDEB 2015



(b) IDEB 2019



(c) IDEB 2020

Fonte: Elaborada pelo autor.

Quadro 6 – Estatísticas da Taxa de Distorção Idade-série

TDI 2015

Estatística	Dados
Média	16,55
Mediana	11,9
Desvio Padrão	16,42
Coeficiente de Variação	0,99
TDI Máximo	100
TDI Mínimo	0

TDI 2019

Estatística	Dados
Média	13,13
Mediana	8,6
Desvio Padrão	14,41
Coeficiente de Variação	1,10
TDI Máximo	100
TDI Mínimo	0

TDI 2021

Estatística	Dados
Média	9,88
Mediana	6,0
Desvio Padrão	12,44
Coeficiente de Variação	1,26
TDI Máximo	100
TDI Mínimo	0

Fonte: Elaborado pelo autor.

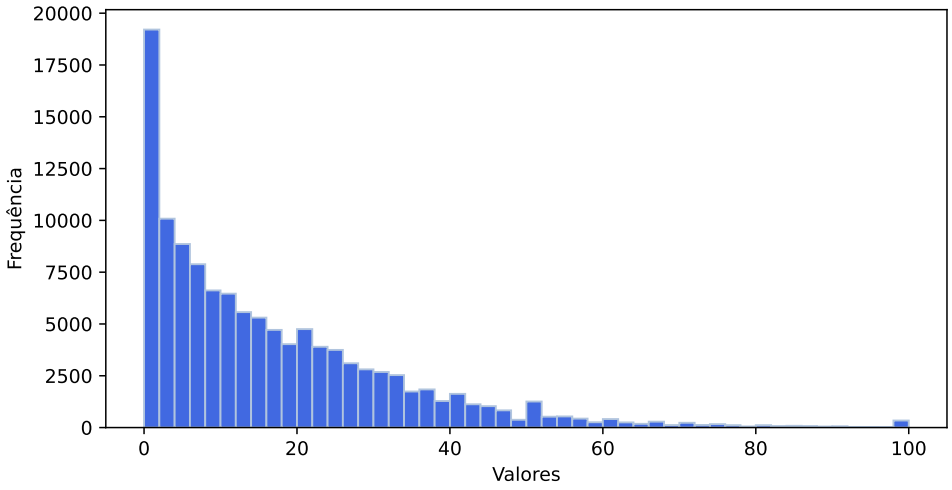
Analisando o TDI, conforme a Seção 2.1.1.2, esse indicador reflete a presença de alunos que possivelmente repetiram de ano, além de possuir correlação com os dados do IDEB. Em 2015, como visto no Quadro 6, a média do TDI foi de 16,56, com uma mediana de 11,9, indicando uma taxa de distorção idade-série moderada na maioria das escolas. No entanto, o desvio padrão de 16,43 sugere uma grande variabilidade, com algumas escolas enfrentando problemas severos de defasagem.

Em 2019, a média do TDI caiu para 13,13 e a mediana para 8,6, refletindo uma melhora geral na distorção idade-série. O desvio padrão reduziu para 14,41, mas o coeficiente de variação aumentou para 1,10, indicando que a variabilidade relativa cresceu. Já em 2020, a média caiu para 9,88, e a mediana para 6,0, mantendo a tendência de redução da distorção. O desvio padrão reduziu para 12,44, mas o coeficiente de variação subiu para 1,26, sugerindo uma maior dispersão relativa.

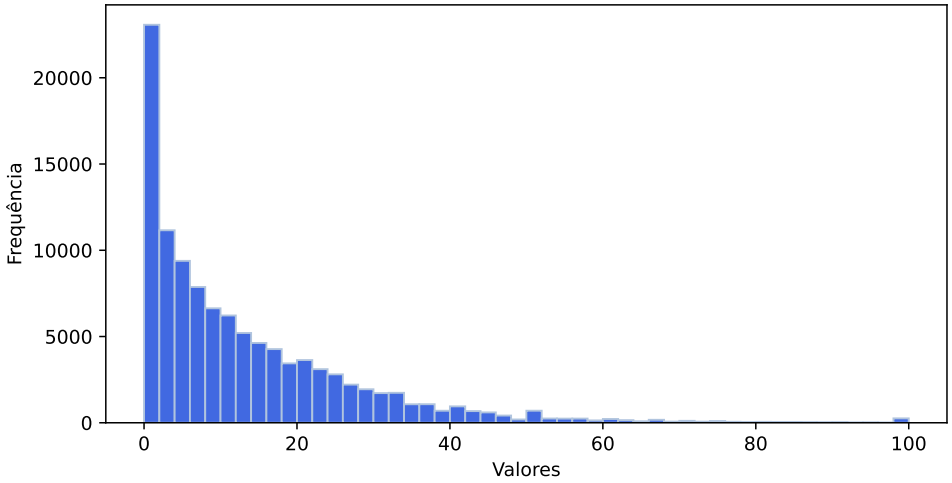
Na Figura 6, observa-se que a maioria das escolas apresenta taxas de distorção próximas a 0%, mas com uma longa cauda à direita, indicando a existência de instituições com valores elevados. No ano de 2019, representado na Figura 6b, há uma redução geral no TDI, com mais escolas concentradas em faixas mais baixas. Já em 2020, representado pela Figura 6c, essa tendência se acentua, mas a persistência de valores extremos corrobora o aumento do coeficiente de variação observado nas estatísticas descritivas.

Por fim, ao comparar os três anos, é possível notar uma tendência de redução na média e mediana do TDI, indicando melhora na distorção idade-série. No entanto, o aumento do coeficiente de variação sugere maior variabilidade nos dados, possivelmente devido a escolas com taxas muito altas de distorção. A presença de valores extremos de 100 em todos os anos evidencia a necessidade de políticas específicas para lidar com as escolas que ainda enfrentam problemas de defasagem, mostrando que, embora tenha ocorrido progresso, desafios persistem para garantir mais equidade no sistema educacional.

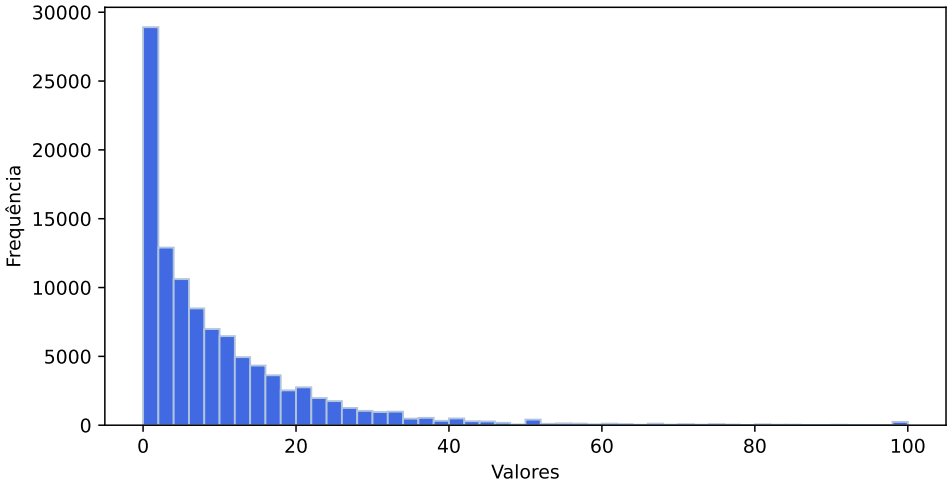
Figura 6 – Histograma da Taxa de Distorção Idade-Série



(a) TDI 2015



(b) TDI 2019



(c) TDI 2020

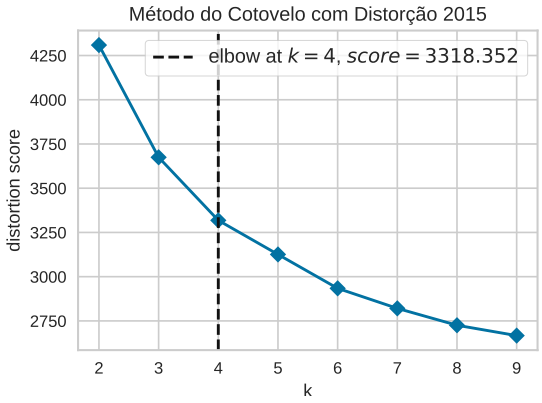
Fonte: Elaborada pelo autor.

5.2 Escolha dos Parâmetros do Algoritmo

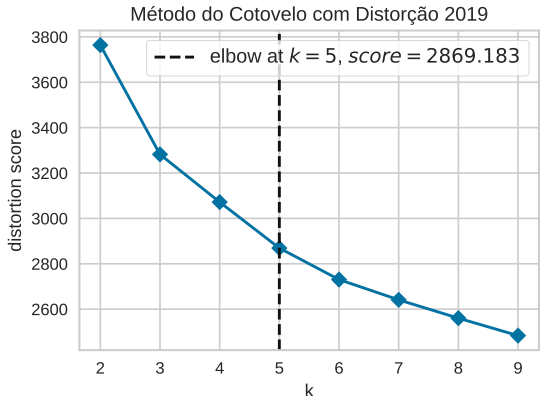
Foi gerado um gráfico do método do cotovelo para cada ano. Em 2015 (Figura 7a) e 2021 (Figura 7b), os resultados indicam que 4 clusters são ideais, com scores de distorção de 3318,352 e 3157,345, respectivamente. Isso sugere que essa segmentação é suficiente para esses dados. Enquanto em 2019 (Figura 7c), o método apontou 5 clusters como ideal, com um score de distorção de 2869,183.

Contudo, como mostrado na Figura 7c, a inclinação do cotovelo em 2019 não é evidente, dificultando a identificação do ponto ideal. Além disso, esse método, embora amplamente usado, apresenta limitações, como subjetividade na interpretação e sensibilidade à distribuição dos dados. Para validar k, utilizou-se o coeficiente de silhueta, gerando também seu gráfico para os mesmos anos, com k variando de 3 a 5, a fim de aprimorar a escolha do número ideal de clusters.

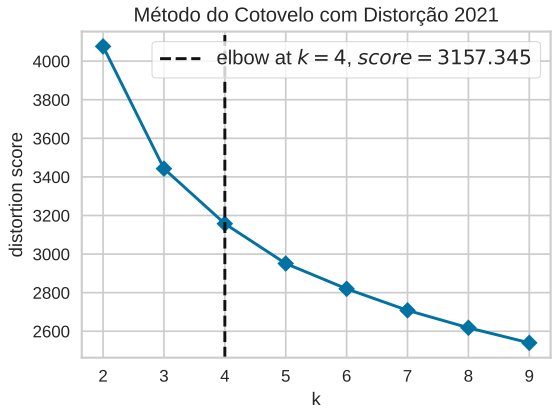
Figura 7 – Método do Cotovelo dos Diferentes Anos



(a) Agrupamento 2015



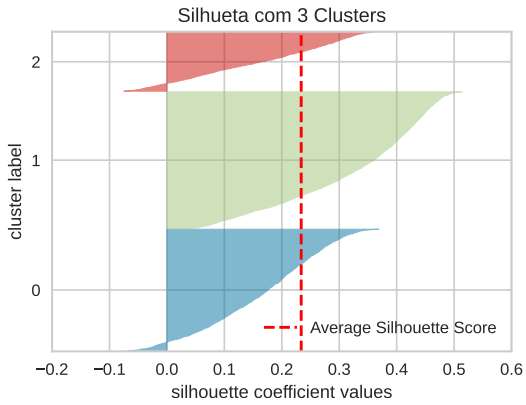
(b) Agrupamento 2019



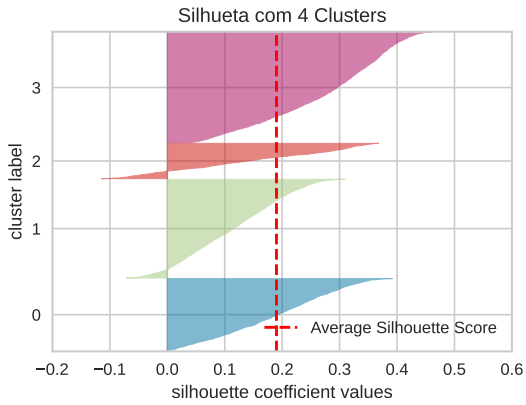
(c) Agrupamento 2021

Fonte: Elaborada pelo autor.

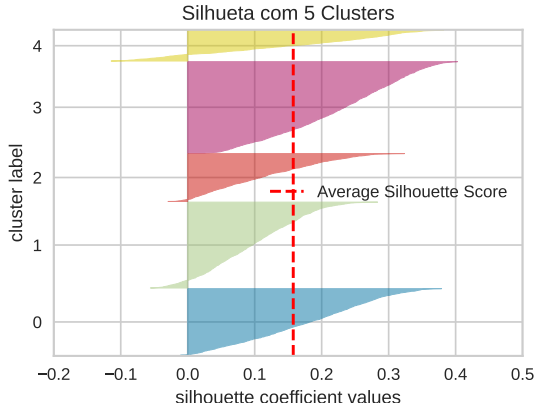
Figura 8 – Silhueta do Agrupamento de 2015



(a) Agrupamento 3 Clusters



(b) Agrupamento 4 Clusters



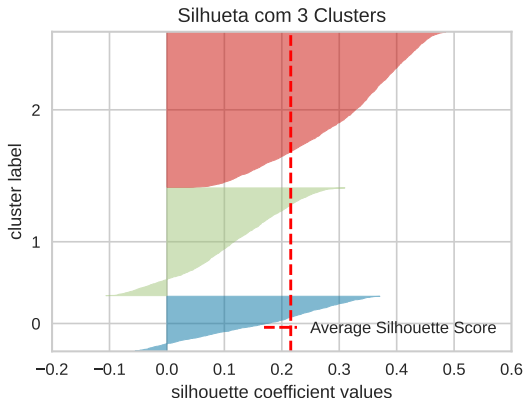
(c) Agrupamento 5 Clusters

Fonte: Elaborada pelo autor.

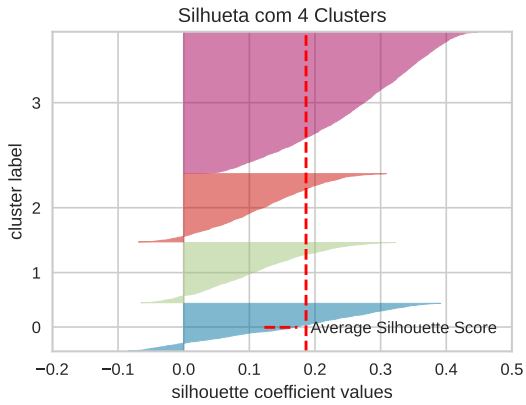
Para o ano de 2015, observa-se que, com 3 *clusters*, Figura 8a, a silhueta não apresenta uma distribuição uniforme, destacando-se um valor significativamente maior, como no cluster 1. Já com 5 *clusters*, há uma grande presença de valores negativos, além de uma alta variação entre as silhuetas, o que indica uma segmentação menos eficiente. Contudo, a silhueta com 4 *clusters*, na Figura 8b se mostra a melhor opção, pois apresenta uma distribuição mais equilibrada, sem valores excessivamente discrepantes ou negativos. Além disso, esse resultado está alinhado com o método do cotovelo para o ano de 2015 Figura 7a.

De forma semelhante, no ano de 2019, a silhueta para 3 *clusters*, visto na Figura 9a, também não é uniforme, apresentando um valor significativamente maior em relação aos demais. No caso de 5 *clusters*, observa-se novamente um excesso de valores negativos, reforçando a inadequação dessa segmentação. Dessa forma, a opção com 4 *clusters*, da Figura 9c, se mostra mais adequada, pois mantém uma distribuição mais equilibrada e sem variações extremas, além de estar em consonância com os resultados da análise do método do cotovelo dos outros anos.

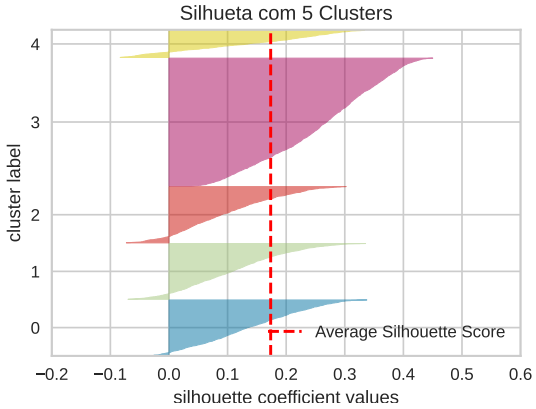
Figura 9 – Silhueta do Agrupamento de 2019



(a) Agrupamento 3 Clusters



(b) Agrupamento 4 Clusters



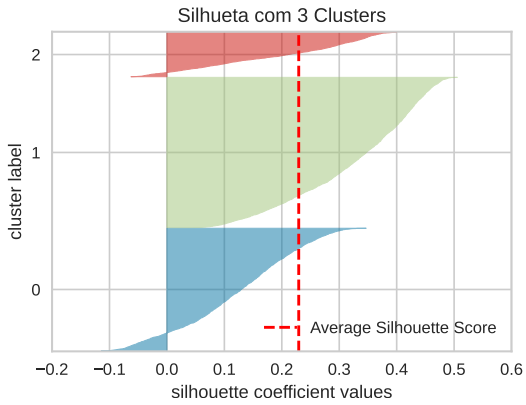
(c) Agrupamento 5 Clusters

Fonte: Elaborada pelo autor.

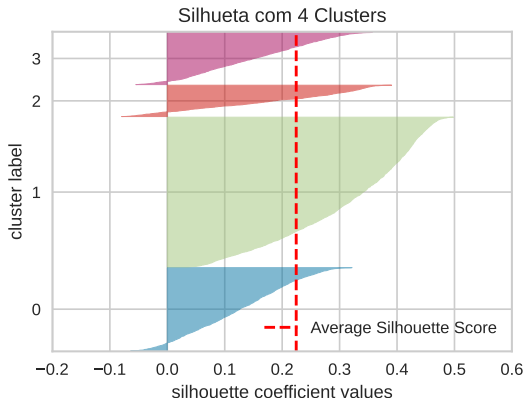
Por fim, é possível observar o mesmo padrão no ano de 2021 na Figura 10. A silhueta para 3 clusters não apresenta uma distribuição uniforme, com um valor significativamente maior em relação aos demais. Já no caso de 5 clusters, há um excesso de valores negativos e uma grande discrepância no tamanho das silhuetas, reforçando a inadequação dessa segmentação. Dessa forma, a opção com 4 clusters se mostra mais adequada, pois apresenta uma distribuição mais equilibrada em comparação com as outras alternativas, além de estar conforme o método do cotovelo na Figura 7c.

Dessa maneira, a escolha de 4 clusters se mostra a opção mais consistente ao longo dos anos analisados. Esse resultado reflete uma segmentação mais equilibrada e homogênea, evitando a presença excessiva de valores negativos e discrepantes observada em outras configurações. Além disso, a convergência entre os critérios avaliados — análise da silhueta e método do cotovelo — reforça a robustez dessa escolha, garantindo uma divisão coerente dos dados e uma melhor interpretação dos perfis dos clusters.

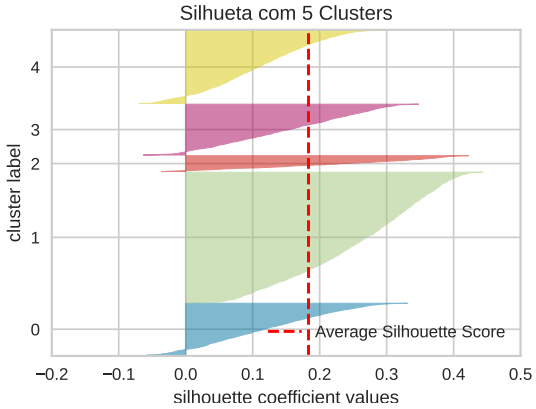
Figura 10 – Silhueta do Agrupamento de 2021



(a) Agrupamento 3 Clusters



(b) Agrupamento 4 Clusters



(c) Agrupamento 5 Clusters

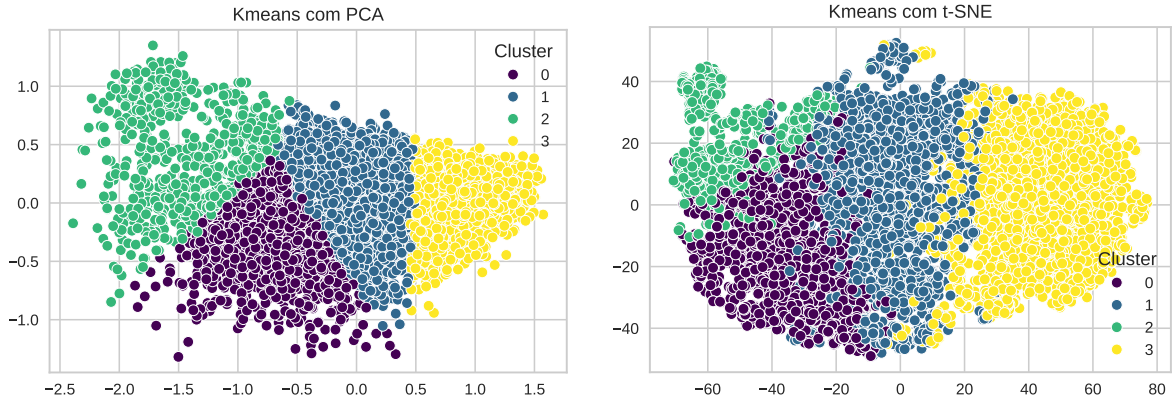
Fonte: Elaborada pelo autor.

5.3 Execução e Validação da Clusterização

Para cada um dos períodos selecionados, foram aplicadas técnicas de agrupamento utilizando o algoritmo K-means em combinação com os métodos de redução de dimensionalidade Principal Component Analysis (PCA) e t-Distributed Stochastic Neighbor Embedding (t-SNE), para realizar sua plotagem no gráfico.

Os resultados da clusterização mostram variações na distribuição dos dados ao longo do tempo, o que sugere mudanças nas características das amostras analisadas. No caso da redução de dimensionalidade via PCA, observa-se uma segmentação bem definida dos grupos, com limites relativamente claros entre os clusters. A distribuição espacial dos pontos revela que as diferentes categorias possuem certa homogeneidade dentro de cada grupo, embora pequenas sobreposições possam ser notadas em algumas áreas da projeção.

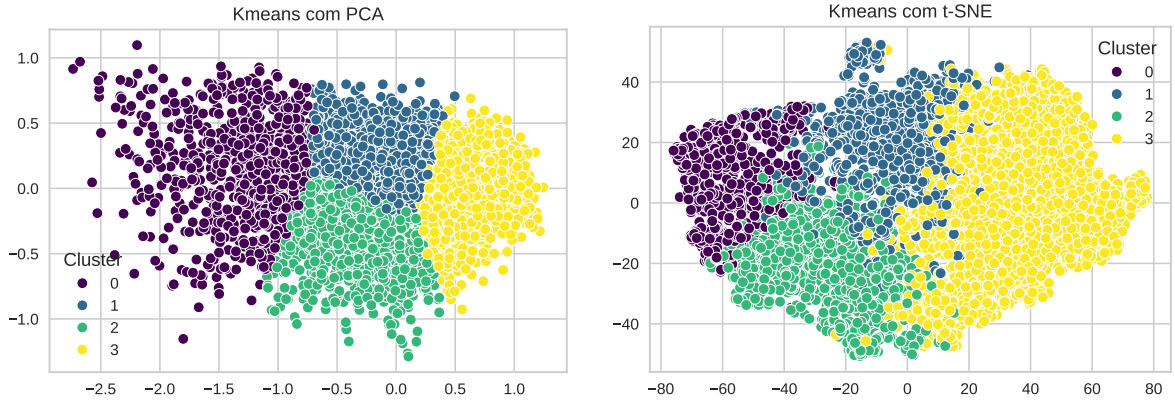
Figura 11 – Clusterização de 2015



(a) Agrupamento usando PCA (b) Agrupamento usando t-SNE

Fonte: Elaborada pelo autor.

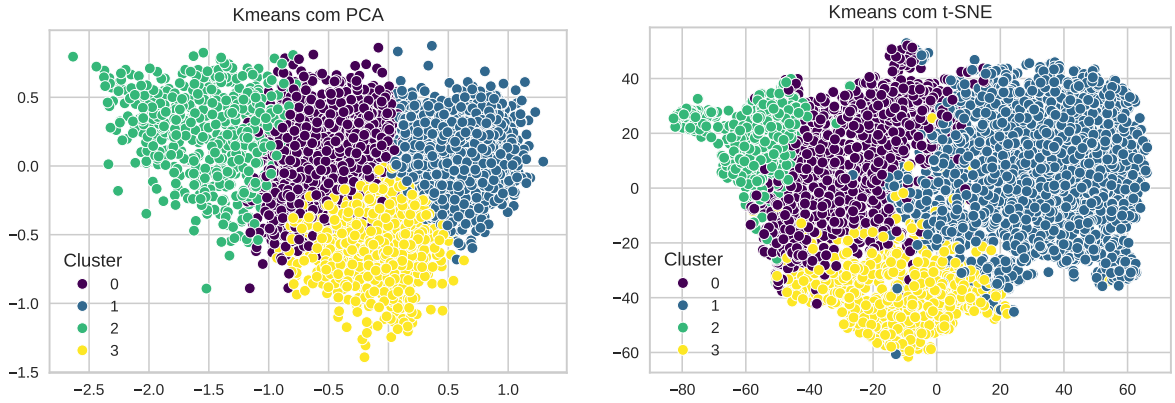
Figura 12 – Clusterização de 2019



(a) Agrupamento usando PCA (b) Agrupamento usando t-SNE

Fonte: Elaborada pelo autor.

Figura 13 – Clusterização de 2021



(a) Agrupamento usando PCA (b) Agrupamento usando t-SNE

Fonte: Elaborada pelo autor.

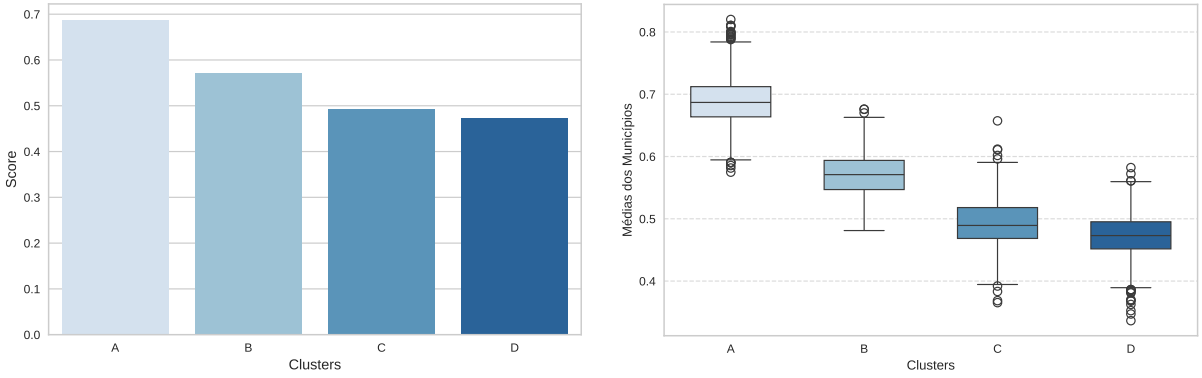
Por outro lado, ao empregar o método t-SNE, a visualização dos *clusters* não apresentou uma separação clara, com os agrupamentos frequentemente sobrepostos no espaço bidimensional. A estrutura resultante sugere que a técnica do PCA consegue capturar melhor as relações locais entre os pontos, organizando-os de forma que a proximidade dos dados reflete suas semelhanças intrínsecas.

Ao comparar os diferentes anos analisados, percebe-se que os padrões de agrupamento apresentam algumas semelhanças estruturais, mas também diferenças significativas que indicam mudanças ao longo do tempo. Em particular, a evolução na formação dos *clusters* sugere possíveis transformações nos fatores que influenciam a distribuição dos dados. Pois, por mais que a quantidade de cluster se mantenha assim como a forma principal, ainda é possível notar visualmente alguma mudança no tamanho dos *clusters* ao longo dos anos.

5.4 Análise dos Perfis dos Clusters

Ao analisar a pontuação dos *clusters* de 2015, observa-se uma variação significativa entre grupos que refletem diferentes níveis educacionais, onde o *Cluster* A representa os municípios com as escolas de melhor desempenho e o *Cluster* D, as de pior desempenho, com os demais grupos posicionados intermediariamente. O *Cluster* A atinge aproximadamente 0,68, enquanto os *Clusters* B, C e D apresentam médias distintas, evidenciando uma distribuição heterogênea. Além disso, a análise da distorção revela uma alta variabilidade no *Cluster* A, com valores extremos mais expressivos, enquanto os *Clusters* C e D exibem distorções semelhantes, indicando perfis mais homogêneos.

Figura 14 – Pontuação dos clusters de 2015

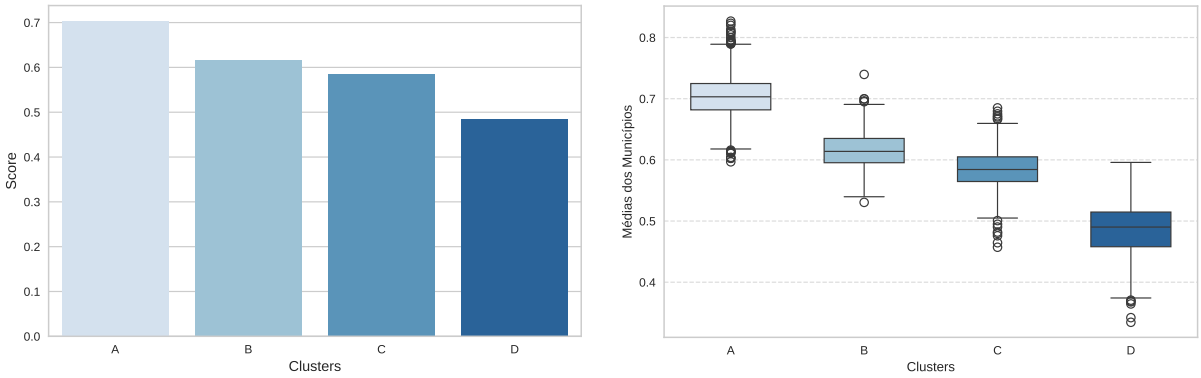


(a) Média da pontuação dos cluster

(b) Distorção dos cluster

Fonte: Elaborada pelo autor.

Figura 15 – Pontuação dos clusters de 2019

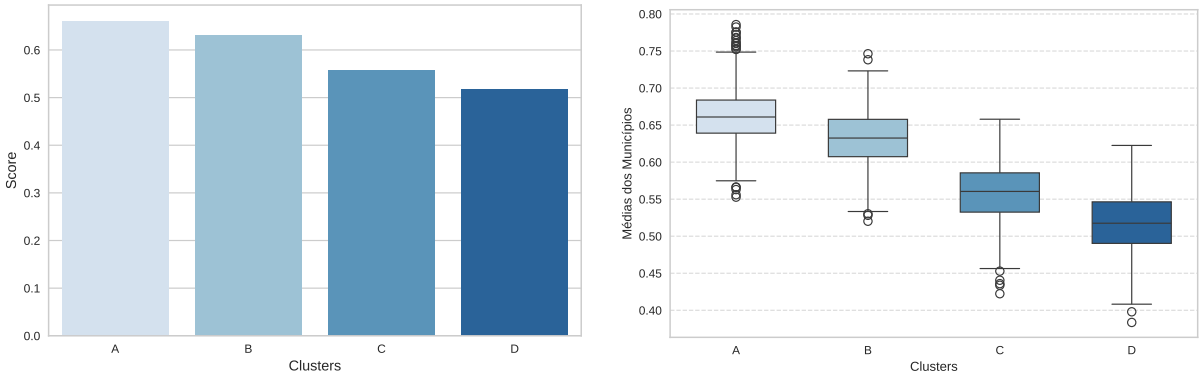


(a) Média da pontuação dos cluster (b) Distorção dos cluster

Fonte: Elaborada pelo autor.

Em seguida, no ano de 2019, os escores dos *clusters* demonstram uma tendência de melhoria em relação a 2015, com todos os grupos apresentando pontuações superiores às do período anterior. O *cluster* A se destaca, atingindo aproximadamente 0,70, seguido pelos *clusters* B e C, que alcançam valores de 0,61 e 0,58, respectivamente. A principal diferença entre os anos analisados está na maior aproximação entre os *clusters* B e C, sugerindo uma redução na disparidade entre esses grupos. A análise da distorção revela que, embora o *cluster* A continue apresentando uma quantidade significativa de valores extremos, há uma leve redução da dispersão no *cluster* D em comparação ao ano anterior. Esse comportamento pode indicar um aumento na estabilidade e na homogeneidade dos grupos.

Figura 16 – Pontuação dos clusters de 2021



(a) Média da pontuação dos cluster (b) Distorção dos cluster

Fonte: Elaborada pelo autor.

Em 2021, observa-se uma continuidade na evolução dos escores dos *clusters*, com exceção do *Cluster* A, que apresenta uma leve redução para 0,66. Por outro lado, os *Clusters* B e D demonstram crescimento, atingindo 0,63 e 0,53, respectivamente, enquanto o *Cluster* C apresenta uma leve queda para 0,55. A análise da distorção em 2021 indica uma estabilização, com menor dispersão dos valores em comparação aos anos anteriores. Diferentemente dos anos anteriores, os *Clusters* A e B demonstram maior semelhança na distorção, indicando possíveis mudanças na composição desses grupos ao longo do tempo.

Essas alterações podem refletir as desigualdades educacionais ampliadas pela pandemia, afetando de maneira distinta os diferentes níveis de desempenho escolar. Pois, o *Cluster* A continua apresentando uma quantidade significativa de valores extremos, mesmo com a redução na pontuação. Essas variações podem estar relacionadas ao impacto da pandemia de COVID-19, que afetou o desempenho educacional no Brasil (Bof; Moraes, 2022). Além disso, a suspensão das atividades presenciais em 99,3% das escolas brasileiras e a adoção de estratégias de ensino não presenciais contribuíram para desafios na aprendizagem (INEP, 2021).

5.5 Análise da Evolução dos Clusters

Além das mudanças nas características intrínsecas dos clusters ao longo do tempo, observa-se também uma significativa reconfiguração na distribuição dos municípios dentro desses agrupamentos. Essa dinâmica reflete não apenas transformações internas nos atributos que definem cada cluster, mas também a migração de entidades entre os diferentes grupos, indicando possíveis alterações socioeconômicas, políticas ou institucionais que impactaram a composição e a classificação das entidades analisadas. A análise dessa redistribuição permite compreender melhor as tendências de evolução e os fatores que influenciaram a movimentação dos municípios entre os clusters ao longo dos anos de 2015, 2019 e 2021. Logo, ao examinar essas mudanças, é possível identificar padrões de comportamento durante o período observado.

Quadro 7 – Quantidade de entidades nos clusters

Clusters 2015

Clusters	Quantidade	Pontuação
A	1.941	0,688033
B	1.737	0,570959
C	622	0,493403
D	1.270	0,472214

Clusters 2019

Clusters	Quantidade	Pontuação
A	2.474	0,702837
B	1.060	0,615062
C	1.201	0,583990
D	835	0,485258

Clusters 2021

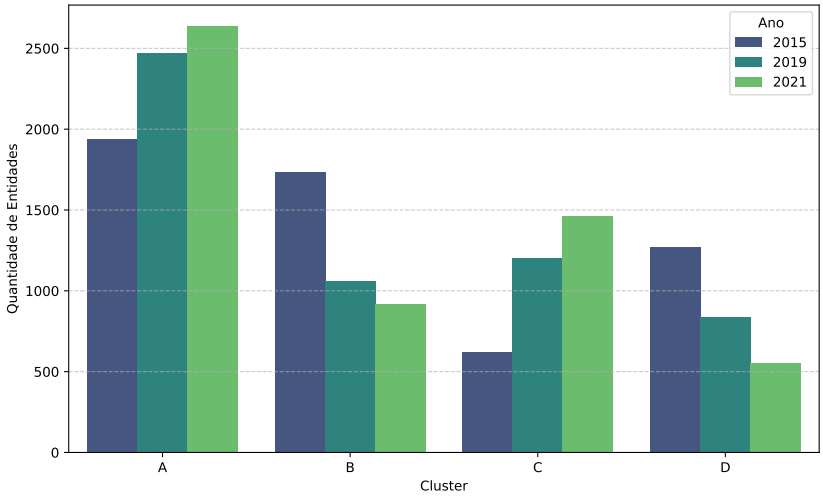
Clusters	Quantidade	Pontuação
A	2.636	0,661250
B	919	0,631496
C	1.463	0,557952
D	552	0,518104

Fonte: Elaborado pelo autor.

A análise da evolução dos clusters ao longo dos anos revela mudanças significativas na distribuição das entidades dentro de cada grupo, sugerindo possíveis transformações estruturais nos dados. Conforme ilustrado no Quadro 7, o cluster A apresentou um crescimento contínuo, passando de 1.941 entidades em 2015 para 2.474 em 2019 e atingindo 2.636 em 2021. Além disso, a pontuação média do cluster aumentou de 0,68 em 2015 para 0,70 em 2019, mas sofreu uma queda para 0,66 em 2020, possivelmente afetado pela COVID-19. Esse crescimento no número de entidades indica que as características que definem esse agrupamento tornaram-se mais predominantes ao longo do tempo, o que pode refletir tanto uma tendência de homogeneização quanto uma possível melhoria nos atributos associados a esse grupo.

Por outro lado, o *cluster* B experimentou uma redução significativa, diminuindo de 1737 entidades em 2015 para 1060 em 2019 e posteriormente para 919 em 2021. Esse declínio pode indicar que muitas entidades anteriormente pertencentes a esse grupo migraram para outros *clusters*, o que sugere uma reestruturação na composição das categorias analisadas. O *cluster* C, diferentemente, demonstrou um crescimento expressivo ao longo do período analisado. Inicialmente com 622 entidades em 2015, esse número, em 2019, atingiu 1201, e continuou a expandir-se, chegando a 1463 em 2021. Em contraste, o *cluster* D apresentou um decréscimo considerável, passando de 1270 entidades em 2015 para 835 em 2019 e reduzindo-se ainda mais para 552 em 2021. Essa redução acentuada pode indicar que as entidades desse grupo foram absorvidas por outros *clusters*, sugerindo uma perda de relevância das características que definiam esse agrupamento.

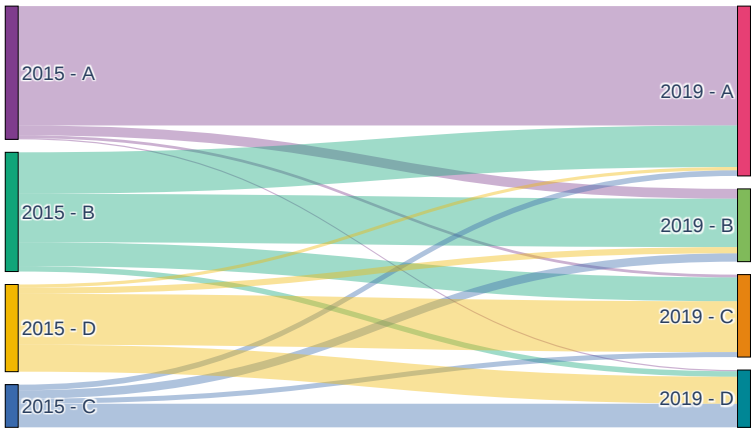
Figura 17 – Gráfico da quantidade de entidades por cluster



Fonte: Elaborada pelo autor.

De maneira geral, não houve a morte ou criação de um novo *cluster*, mas observou-se um aumento significativo no número de entidades nos *clusters* A e C, enquanto os demais reduziram sua participação. Isso sugere uma migração de entidades entre os *clusters*, possivelmente associada a mudanças socioeconômicas, políticas ou institucionais que impactaram o desempenho das escolas. O crescimento do *cluster* A pode indicar uma melhoria nos indicadores de qualidade educacional, fazendo com que entidades de *clusters* inferiores migrassem para *clusters* superiores. Já o aumento do *cluster* C sugere que algumas escolas podem ter tido uma queda na qualidade, migrando para esse *cluster* a partir de classificações mais altas.

Figura 18 – Gráfico da migração das entidades de 2015 para 2019

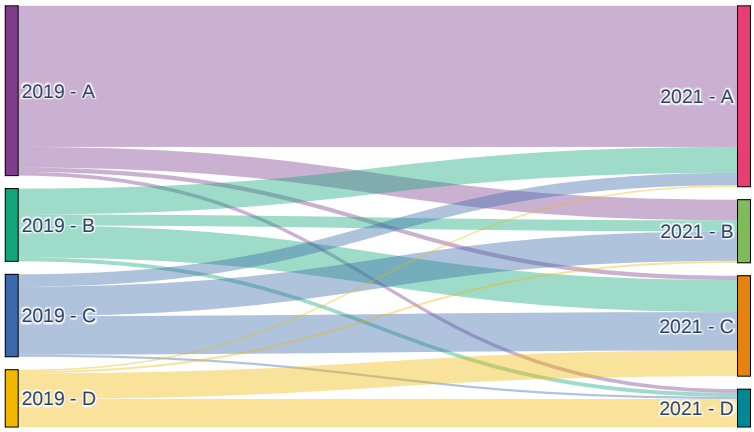


Fonte: Elaborada pelo autor.

A análise da migração de entidades entre os *clusters* da Figura 18 destacou o *cluster* A como o principal receptor de entidades, recebendo 600 entidades do *cluster* B, 83 do *cluster* C e 47 do *cluster* D. Contudo, o *cluster* B apresentou um comportamento mais equilibrado, recebendo 143 entidades do *cluster* A, 122 do *cluster* C e 90 do *cluster* D. O *cluster* C também demonstrou uma dinâmica significativa, recebendo 41 entidades do *cluster* A, 347 do *cluster* B e 73 do *cluster* D. Por fim, o *cluster* D recebeu um número menor de entidades, com 18 provenientes do *cluster* A, 80 do *cluster* B e 344 do *cluster* C.

No período subsequente, Figura 19, O *cluster* A continuou a receber um número significativo de entidades, com 377 provenientes do *cluster* B, 180 do *cluster* C e 23 do *cluster* D. O *cluster* B, por sua vez, recebeu 301 entidades do *cluster* A, 429 do *cluster* C e 29 do *cluster* D. O *cluster* C apresentou uma dinâmica semelhante, recebendo 63 entidades do *cluster* A, 466 do *cluster* B e 374 do *cluster* D. Já o *cluster* D recebeu um volume menor de entidades, com 54 do *cluster* A, 57 do *cluster* B e 32 do *cluster* C.

Figura 19 – Gráfico da migração das entidades de 2019 para 2021



Fonte: Elaborada pelo autor.

Esses padrões de migração sugerem que fatores como investimentos em infraestrutura escolar, políticas de valorização do magistério e programas de incentivo à permanência dos alunos na escola podem ter desempenhado um papel fundamental na ascensão de municípios aos *clusters* superiores. Por outro lado, desigualdades regionais, redução de investimentos e dificuldades de adaptação às novas metodologias de ensino, somadas aos impactos da pandemia de COVID-19, podem ter contribuído para a regressão de algumas entidades, resultando em sua migração para *clusters* de menor desempenho.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve como objetivo principal analisar a evolução da qualidade de ensino das escolas públicas no Brasil por meio da clusterização de indicadores educacionais, utilizando técnicas avançadas de análise de dados. A partir da aplicação do algoritmo K-Means e de métodos de redução de dimensionalidade, como PCA e t-SNE, foi possível identificar padrões e tendências nos dados educacionais dos anos de 2015, 2019 e 2021. A análise exploratória dos dados revelou uma melhoria gradual no IDEB ao longo dos anos, com uma redução na variabilidade dos resultados, indicando uma maior equidade no sistema educacional. No entanto, a persistência de valores extremos em indicadores como a Taxa de Distorção Idade-Série pode evidenciar que problemas significativos ainda precisam ser superados, especialmente em regiões com condições socioeconômicas mais vulneráveis.

A escolha do número ideal de clusters foi realizada com base no método do cotovelo e no coeficiente de silhueta, resultando em quatro clusters para todos os anos analisados. Essa segmentação permitiu uma análise detalhada dos perfis dos clusters, revelando uma migração significativa de entidades entre os grupos ao longo do tempo. O cluster A, que representa as escolas com melhores indicadores educacionais, apresentou um crescimento contínuo, enquanto os clusters B e D sofreram reduções significativas. O cluster C, por sua vez, mostrou um aumento expressivo, sugerindo que algumas escolas podem ter experimentado uma deterioração em suas condições educacionais.

A análise da evolução dos clusters destacou a importância de políticas públicas que promovam a equidade educacional, especialmente em regiões com maiores desafios socioeconômicos. A migração de entidades entre os clusters ao longo dos anos sugere que investimentos direcionados e estratégias específicas podem contribuir para a melhoria do desempenho escolar e a redução das desigualdades educacionais. Além disso, a aplicação de técnicas de clusterização e análise de dados mostrou-se uma ferramenta valiosa para a compreensão dos fatores que influenciam a qualidade da educação no Brasil, fornecendo subsídios para a formulação de políticas mais eficazes e equitativas. Por exemplo, escolas no cluster C, que demonstraram uma deterioração nos indicadores, podem se beneficiar de investimentos direcionados em programas de reforço escolar, capacitação de professores e infraestrutura. Já as escolas no cluster A, que apresentam os melhores resultados, podem servir como modelos e centros de excelência, compartilhando suas melhores práticas com outras instituições.

Em síntese, este trabalho contribui para a compreensão da evolução da qualidade de ensino no Brasil, destacando a importância de uma abordagem baseada em dados para a tomada de decisões no setor educacional. A análise dos clusters e sua evolução ao longo do tempo oferece percepções valiosos para a identificação de áreas prioritárias de intervenção, visando a melhoria da qualidade da educação e a redução das disparidades regionais.

Futuros trabalhos podem expandir essa análise, incorporando indicadores adicionais, como demográficos, e explorando técnicas mais avançadas de aprendizado de máquina para uma compreensão ainda mais aprofundada dos desafios educacionais no país. Além de poderem usar indicadores demográficos para buscar uma relação direta entre os valores investidos na educação por parte do governo e o nível da qualidade da educação. Além disso, seria interessante conduzir uma análise detalhada dos fatores que influenciam a migração das entidades entre os diferentes *clusters* ao longo dos anos, identificando as principais causas dessas mudanças e como elas refletem as transformações socioeconômicas e políticas no cenário educacional.

Adicionalmente, seria valioso realizar uma análise regional, considerando as desigualdades educacionais no Brasil, a fim de entender como as disparidades geográficas impactam a distribuição e a evolução dos clusters. Por fim, a validação externa dos resultados com especialistas em educação ou a comparação com classificações oficiais poderia fortalecer a confiabilidade e a aplicabilidade prática do estudo, garantindo que as conclusões estejam alinhadas com a realidade do sistema educacional brasileiro.

REFERÊNCIAS

BADILLO, S.; BANFAI, B.; BIRZELE, F.; DAVYDOV, I. I.; HUTCHINSON, L.; KAM-THONG, T.; SIEBOURG-POLSTER, J.; STEIERT, B.; ZHANG, J. D. An introduction to machine learning. **Clinical pharmacology & therapeutics**, Wiley Online Library, v. 107, n. 4, p. 871–885, 2020.

BHOLOWALIA, P.; KUMAR, A. Ebk-means: A clustering technique based on elbow method and k-means in wsn. **International Journal of Computer Applications**, Citeseer, v. 105, n. 9, 2014.

BI, Q.; GOODMAN, K. E.; KAMINSKY, J.; LESSLER, J. What is machine learning? a primer for the epidemiologist. **American journal of epidemiology**, Oxford University Press, v. 188, n. 12, p. 2222–2239, 2019.

BOF, A. M.; MORAES, G. H. Impactos da pandemia no aprendizado dos estudantes brasileiros. **Cadernos de Estudos e Pesquisas em Políticas Educacionais**, v. 7, 2022.

CNN. **Brasil Tem Baixo desempenho E estagna em ranking Mundial da Educação Básica**. 2023. Dados publicados na CNN Brasil. Disponível em: https://www.cnnbrasil.com.br/nacional/brasil-estaciona-em-ranking-de-avaliacao-internacional-de-educacao-basica/. Acesso em: 12 jul. 2024.

CUI, M. *et al.* Introduction to the k-means clustering algorithm based on the elbow method. **Accounting, Auditing and Finance**, Clausius Scientific Press, v. 1, n. 1, p. 5–8, 2020.

CUTLER, D. M.; LLERAS-MUNEY, A. Education and health: insights from international comparisons. National Bureau of Economic Research, 2012.

DHAR, V. Data science and prediction. **Communications of the ACM**, ACM New York, NY, USA, v. 56, n. 12, p. 64–73, 2013.

FERNÁNDEZ, R.; CORREAL, J. F.; D'AYALA, D.; MEDAGLIA, A. L. A decision-making framework for school infrastructure improvement programs. **Structure and Infrastructure Engineering**, Taylor & Francis, p. 1–20, 2023.

GONÇALVES, T. G. G. L.; SANTO, S. C. do; SANTOS, N. G. dos. Indicadores educacionais brasileiros: limites e perspectivas. **Educação Em Perspectiva**, v. 8, n. 3, p. 444–461, 2017.

GRABUSTS, P. The choice of metrics for clustering algorithms. In: **ENVIRONMENT. TECHNOLOGIES. RESOURCES. Proceedings of the International Scientific and Practical Conference**. [*S. l.*: *s. n.*], 2011. v. 2, p. 70–76.

HARTAMA, D.; ANJELITA, M. Analysis of silhouette coefficient evaluation with euclidean distance in the clustering method (case study: Number of public schools in indonesia). **Jurnal Mantik**, v. 6, n. 3, p. 3667–3677, 2022.

INEP. **Estatísticas revelam os impactos da pandemia na educação**. 2021. Site contendo depate sobre o impacto da pandemia. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/institucional/estatisticas-revelam-os-impactos-da-pandemia-na-educacao. Acesso em: 11 mar. 2025.

INEP. **Indicador apresenta distorção idade-série para ensino fundamental e médio**. 2024. Descrição do INEP sobre o Indice de Distorcao idade-série e seus resultados. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/indicador-apresenta-distorcao-idade-serie-para-ensino-fundamental-e-medio. Acesso em: 16 set. 2024.

INEP. **Taxas de Rendimento Escolar**. 2024. Publicação do INEP que apresenta as taxas de rendimento escolar no Brasil. Disponível em: https://www.gov.br/inep/pt-br/centrais-de-conteudo/acervo-linha-editorial/publicacoes-institucionais/estatisticas-e-indicadores-educacionais/taxas-de-rendimento-escolar. Acesso em: 16 set. 2024.

INEP. **Índice de Desenvolvimento da Educação Básica (IDEB)**. 2024. Site contendo descrição e dados do IDEB. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb. Acesso em: 15 set. 2024.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern recognition letters**, Elsevier, v. 31, n. 8, p. 651–666, 2010.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. **Electronic Markets**, Springer, v. 31, n. 3, p. 685–695, 2021.

KRIEGEL, H.-P.; KRÖGER, P.; SANDER, J.; ZIMEK, A. Density-based clustering. **Wiley interdisciplinary reviews: data mining and knowledge discovery**, Wiley Online Library, v. 1, n. 3, p. 231–240, 2011.

LI, X.; ZHANG, Y.; CHENG, H.; ZHOU, F.; YIN, B. An unsupervised ensemble clustering approach for the analysis of student behavioral patterns. **Ieee Access**, IEEE, v. 9, p. 7076–7091, 2021.

LIMA, S. P.; CRUZ, M. D. A genetic algorithm using calinski-harabasz index for automatic clustering problem. **Revista Brasileira de Computação Aplicada**, v. 12, n. 3, p. 97–106, 2020.

LIU, H.-H.; ONG, C.-S. Variable selection in clustering for marketing segmentation using genetic algorithms. **Expert systems with applications**, Elsevier, v. 34, n. 1, p. 502–510, 2008.

MACQUEEN, J. *et al.* Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [*S. l.*], 1967. v. 1, n. 14, p. 281–297.

MAHESH, B. Machine learning algorithms-a review. **International Journal of Science and Research (IJSR).[Internet]**, v. 9, n. 1, p. 381–386, 2020.

MUGHNYANTI, M.; EFENDI, S.; ZARLIS, M. Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation. In: IOP PUBLISHING. **IOP Conference Series: Materials Science and Engineering**. [*S. l.*], 2020. v. 725, n. 1, p. 012128.

MURTAGH, F.; CONTRERAS, P. Algorithms for hierarchical clustering: an overview. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 2, n. 1, p. 86–97, 2012.

NAFURI, A. F. M.; SANI, N. S.; ZAINUDIN, N. F. A.; RAHMAN, A. H. A.; ALIFF, M. Clustering analysis for classifying student academic performance in higher education. **Applied Sciences**, MDPI, v. 12, n. 19, p. 9467, 2022.

NIKITA SACHDEVA. **Top 12 Clustering Algorithms in Machine Learning**. 2023. Dados publicados no daffodil – Os 12 algorimtos mais populares de clusterização. Disponível em: https://insights.daffodilsw.com/blog/top-5-clustering-algorithms-in-machine-learning. Acesso em: 13 set. 2024.

NUGENT, R.; MEILA, M. An overview of clustering applied to molecular biology. **Statistical methods in molecular biology**, Springer, p. 369–404, 2010.

PITAFI, S.; ANWAR, T.; SHARIF, Z. A taxonomy of machine learning clustering algorithms, challenges, and future realms. **Applied sciences**, MDPI, v. 13, n. 6, p. 3529, 2023.

QEDU. **Distorção Idade-Série**. 2024. Artigo da QEdu que aborda o conceito de distorção idade-série e sua importância para a educação no Brasil. Disponível em: https://conteudos.qedu.org.br/academia/distorcao-idade-serie/. Acesso em: 16 set. 2024.

QUINTERO, Y.; ARDILA, D.; AGUILAR, J.; CORTES, S. Analysis of the socioeconomic impact due to covid-19 using a deep clustering approach. **Applied Soft Computing**, Elsevier, v. 129, p. 109606, 2022.

SHAULSKA, L.; DORONINA, O.; NAUMOVA, M.; HONCHARUK, N.; BONDAREVSKA, K.; TOMCHUK, O. Cross-country clustering of labor and education markets in the system of strategic economic. **REICE: Revista Electrónica De Investigación En Ciencias Económicas**, v. 8, n. 16, p. 166–196, 2020.

SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. In: IEEE. **2018 Fourth international conference on computing communication control and automation (ICCUBEA)**. [*S. l.*], 2018. p. 1–6.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. [*S. l.*]: Pearson Education India, 2016.

USAMA, M.; QADIR, J.; RAZA, A.; ARIF, H.; YAU, K.-L. A.; ELKHATIB, Y.; HUSSAIN, A.; AL-FUQAHA, A. Unsupervised machine learning for networking: Techniques, applications and research challenges. **IEEE access**, IEEE, v. 7, p. 65579–65615, 2019.

VALLES-CORAL, M. A.; SALAZAR-RAMÍREZ, L.; INJANTE, R.; HERNANDEZ-TORRES, E. A.; JUÁREZ-DÍAZ, J.; NAVARRO-CABRERA, J. R.; PINEDO, L.; VIDAURRE-ROJAS, P. Density-based unsupervised learning algorithm to categorize college students into dropout risk levels. **Data**, MDPI, v. 7, n. 11, p. 165, 2022.

WANG, X.; XU, Y. An improved index for clustering validation based on silhouette index and calinski-harabasz index. In: IOP PUBLISHING. **IOP Conference Series: Materials Science and Engineering**. [*S. l.*], 2019. v. 569, n. 5, p. 052024.

XU, R.; WUNSCH, D. **Clustering**. [*S. l.*]: John Wiley & Sons, 2008.

YANG, M.-S.; SINAGA, K. P. A feature-reduction multi-view k-means clustering algorithm. **IEEE Access**, IEEE, v. 7, p. 114472–114486, 2019.

YData. **YData Profiling Documentation**. 2024. Documentação oficial da ferramenta YData Profiling para análise de dados. Disponível em: https://docs.profiling.ydata.ai/latest/. Acesso em: 16 set. 2024.

YUAN, C.; YANG, H. Research on k-value selection method of k-means clustering algorithm. **J**, MDPI, v. 2, n. 2, p. 226–235, 2019.

ZHANG, T.; OLES, F. The value of unlabeled data for classification problems. In: CITESEER. **Proceedings of the Seventeenth International Conference on Machine Learning,(Langley, P., ed.)**. [*S. l.*], 2000. v. 20, n. 0, p. 0.