UNIVERSIDADE FEDERAL DE UBERLÂNDIA – UFU
FACULDADE DE ENGENHARIA MECÂNICA – FEMEC

Alecsander Guimarães Rodrigues

Aplicação de análise de cluster em uma base de dados de escolas estaduais e municipais na região Sudeste

UBERLÂNDIA
2023

Alecsander Guimarães Rodrigues

Aplicação de análise de cluster em uma base de dados de escolas estaduais e municipais na região Sudeste

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharia Mecânica da Universidade Federal de Uberlândia, como parte dos requisitos para obtenção do título de Bacharel em Engenharia Mecatrônica.

Orientador: Prof. Dr. José Eduardo Ferreira Lopes

UBERLÂNDIA

2023

RESUMO

O presente trabalho consistiu na aplicação dos fundamentos da análise de agrupamento, através da combinação dos algoritmos hierárquico e não hierárquico. Para este fim uma base de dados contendo escolas públicas estaduais e municipais da região Sudeste do Brasil foi utilizada, além da linguagem de programação Python na parte prática do trabalho. Objetiva-se com este trabalho, aplicar os fundamentos e técnicas de análise de agrupamentos em uma base dados educacionais, a saber, escolas públicas da educação básica da região Sudeste do Brasil, gerando grupos homogêneos de escolas.. Conclui-se que é possível fazer a segmentação dos dados, tendo os grupos gerados caraterísticas diferentes entre si, contudo, o algoritmo hierárquico tem um alto custo computacional e não é viável para um grande conjunto de dados.

Palavras-chave: Análise de agrupamento; escolas públicas; desempenho escolar; segmentação de dados.

ABSTRACT

The present work consisted in applying the fundamentals of cluster analysis through the combination of hierarchical and non-hierarchical algorithms. For this purpose, a database containing state and municipal public schools from the southeastern region of Brazil was used, in addition to the Python programming language for the practical part of the work. The objective of this work is to apply the principles and techniques of cluster analysis in an educational database, namely public schools of basic education in the Southeast region of Brazil, generating homogeneous groups of schools. It is concluded that it is possible to perform the segmentation of the data, with the generated groups having different characteristics among themselves. However, the hierarchical algorithm has a high computational cost and is not feasible for a large data set.

Keywords: Cluster analysis; public schools; school performance; data segmentation.

SUMÁRIO

1 INTRODUÇÃO

A geração de informações tem sido intensificada nos últimos anos, sendo necessário até a adoção de novos modelos de armazenamento para elas, como *datalakes*. Além da engenharia e arquitetura necessária para comportar todo o novo volume gerado de dados, novas ferramentas para a análise e visualização também se popularizaram, como o Power BI e Tableau. Ainda que tais ferramentas tenham se popularizado e contem com algumas propriedades para a manipulação dos dados, linguagens de programação, como R e Python, são amplamente empregadas para grandes volumes e em casos que os softwares ainda apresentam limitação, como em análise específicas, como a de cluster, e até mesmo análise estatísticas, campo em que a linguagem R se destaca.

A gama de utilização das linguagens de programação é vasta e passa pela visualização, manipulação e limpeza de dados, contando com muitas bibliotecas para cada um desses tópicos, que facilitam seu uso. Ainda é possível contar o suporte da comunidade em caso de dúvidas quanto a sua utilização.

Os dados podem ser utilizados de diferentes formas, sendo com uma análise simples, como a sua visualização por meio de um gráfico em barras ou a confirmação de uma tendência por meio de um gráfico de dispersão. Contudo, há casos que possa ser necessário um estudo mais profundo do conjunto de dados. Uma aplicação possível é a separação de um banco de clientes em grupos, o que pode ajudar na criação de campanhas e anúncios específicos para cada um dos grupos, esse uso de dados é chamado de análise de cluster.

De maneira mais formal, pode-se definir a análise de cluster ou agrupamentos como a junção de objetos fundamentada nas características deles. O principal objetivo é a formação de grupos que tenham elementos parecidos entre si, homogeneidade interna, mas que cada conjunto possua características distintas, heterogeneidade externa. Por utilizar as relações entre as caraterísticas dos objetos em estudo para fazer a separação dos grupos, a generalização dos agrupamentos gerados, e suas caraterísticas, depende da representatividade da amostra em relação à população (HAIR Jr. et al., 2009).

A partir de dados disponibilizados pelo Inep (FNDE, 2022) e dados do censo escolar da educação básica (BRASIL, 2022c), que contém informações sobre as escolas públicas e suas caraterísticas, será aplicado as técnicas de análise de cluster, hierárquico e não-hierárquico, gerando grupos homogêneos de escolas (HAIR Jr. et al., 2009). Os dados são parte da avaliação

do Programa Dinheiro Direto na Escola (PDDE), que tem como objetivo de contribuir para o provimento das necessidades prioritárias dos estabelecimentos educacionais beneficiários que concorram para a garantia de seu funcionamento e para a promoção de melhorias em sua infraestrutura física e pedagógica, bem como incentivar a autogestão escolar e o exercício da cidadania com a participação da comunidade no controle social (BRASIL, 2022b).

O monitoramento do PDDE é feito por meio do Índice de Desempenho da Gestão Descentralizada do PDDE (IdeGES-PDDE). Ele mensura o desempenho da gestão descentralizada do PDDE em todo território nacional, com o objetivo de viabilizar iniciativas de monitoramento e avaliação, orientar a ação governamental para melhoria do desempenho do Programa, favorecer o exercício do controle social e reconhecer iniciativas exitosas de gestão (BRASIL, 2022a).

O IdeGES-PDDE agrega três indicadores relativos a dimensões representativas do desempenho do programa nos entes federados: adesão, execução e prestação de contas dos recursos. A proposta parte do pressuposto que o bom desempenho do PDDE não é alcançado apenas quando, por exemplo, as entidades recebem os recursos. Entende-se que o desempenho do programa em determinado ente federado apenas pode ser considerado satisfatório se alcança o máximo de seu público-alvo (adesão), se os recursos são utilizados (execução) e empregados nas finalidades do programa (prestação de contas (BRASIL, 2022a).

Neste contexto, objetiva-se com este trabalho, aplicar os fundamentos e técnicas de análise de agrupamentos em uma base dados educacionais, a saber, escolas públicas da educação básica da região Sudeste do Brasil, gerando grupos homogêneos de escolas.

O trabalho passará por cada uma das etapas da análise de agrupamento. Ela é composta por seis estágios. O primeiro é a definição do objetivo e definição de quais caraterísticas serão utilizadas para avaliação. O segundo é a avaliação e tratamento dos dados, além da definição da medida de similaridade. O terceiro concerne às suposições feitas durante a análise. O quarto consiste na definição do método empregado e avaliação dos grupos gerados. O quinto é a interpretação dos segmentos e o sexto é a validação dos perfis gerados (HAIR Jr. et al., 2009).

2 REVISÃO BIBLIOGRÁFICA – ANÁLISE DE AGRUPAMENTOS

De acordo com Hair Jr. et al. (2009), a análise de agrupamentos é um grupo de técnicas multivariáveis, com o principal objetivo de formar conjuntos com os objetos, baseando-se em suas caraterísticas. Como o método não rotula seus dados, trata-se de um método de aprendizagem não supervisionado. Dessa forma, assume-se que há uma relação entre os objetos, o que deve ser validado ao final do processo de formação dos *clusters.*

A principal caraterística do método é que cada grupo seja formado por objetos com caraterísticas muito similares entre si, mas que sejam diferentes quando comparados aos outros grupos. Como a formação dos *clusters* depende das caraterísticas dos objetos, é fundamental que as variáveis utilizadas durante o processo de análise sejam representativas (HAIR Jr. et al., 2009).

Segundo Hair Jr. et al. (2009), a criação dos grupos a partir da metodologia de agrupamento pode ser repartida em seis estágios, os quais serão abordados nas seções seguintes. O processo passa não só pelo processo de formação dos grupos e escolha da resolução, mas também pela interpretação dos resultados iniciais e rotulação dos dados, por último, a validação da solução final, bem como a identificação das caraterísticas dos *clusters* formados.

2.1 OBJETIVOS DA ANÁLISE DE AGRUPAMENTOS

De acordo com Hair Jr. et al. (2009), o objetivo final é a elaboração de grupos. Além disso, deve-se saber qual é o problema que está sendo abordado, ou seja, com qual fim o método está sendo utilizado. Como o resultado é altamente dependente das variáveis selecionadas, a sua seleção também é feita na primeira parte do processo.

A análise é comumente empregada para fins exploratórios, mas pode-se agrupá-los em três objetivos, entretanto a análise não necessita de ficar limitada a apenas um deles. O primeiro é descrição taxonômica, ou seja, a segmentação dos dados em grupos. O segundo é a simplificação dos dados, ao invés de usar cada observação para se ter compreensão do problema, pode-se utilizar o agrupamento o qual ela pertence. Por último, a análise ainda pode ser usada para a identificação de relações, estas podem não ser percebidas quando com as observações individuais, mas quando os dados são vistos com a estrutura simplificada que foi obtida pelo método, novas relações podem ser descobertas (HAIR Jr. et al., 2009).

Segundo Hair Jr. et al. (2009), a seleção das variáveis deve ser feita de maneira que elas caracterizem os objetos que serão agrupados, bem como tenha relação com os fins da análise. Como já mencionado, os grupos são formados com base nas variáveis de cada observação presente no conjunto de dados, assim, o resultado depende dessa seleção.

2.2 AVALIAÇÃO DOS DADOS

As observações podem ser formadas por diferentes tipos de dados, métricos ou não-métricos. Além disso, pode não ser interessante utilizar todas as variáveis durante o processo de agrupamento, bem como a quantidade de observações utilizadas e sua padronização. Outro ponto a ser observado é a presença de valores discrepantes e como eles devem ser tratados durante a análise. Por último, a definição de como as observações devem ser comparadas entre si (HAIR Jr. et al., 2009).

2.2.1 Tipo e Número das Variáveis

De acordo com Hair Jr. et al. (2009), as variáveis presentes no conjunto de dados podem ser tanto métricas quanto não-métricas, contudo, o método é melhor empregado quando são as variáveis selecionadas são de apenas um modelo, sendo mais comum a utilização das métricas.

Quanto ao número de variáveis empregadas no método, recomenda-se a utilização de, no máximo, 20 variáveis. Devido ao cálculo de similaridade, que se torna insustentável, tal problema é chamado maldição da dimensionalidade (HAIR Jr. et al., 2009).

Segundo Hair Jr. et al. (2009), o método não possui uma ferramenta para o auxílio na seleção das variáveis. Contudo, pode-se observar quais das variáveis dos agrupamentos resultantes e fazer a replicação do método sem elas.

2.2.2 Tamanho Amostral

A quantidade de observações selecionadas está relacionada à representatividade dos grupos da população. Caso seja de interesse da análise a identificação de grupos pequenos, deve-se utilizar uma amostra maior para facilitar a identificação. Além disso, é importante definir o tamanho mínimo dos grupos, para que seja possível realizar a distinção entre um grupo pequeno e observações discrepantes. Todavia, o aumento do número de amostras também

aumenta o cálculo necessário para as medidas de similaridade e, por isso, deve-se ter cuidado na seleção (HAIR Jr. et al., 2009).

2.2.3 Discrepâncias

Segundo Hair Jr. et al. (2009), os *outliers* podem ser de três tipos: observações que realmente não são representativas da população, representações que formam grupos pequenos, mas que não formam uma segmentação interessante para análise, e, por último, grupos, que devidos a amostragem, não amostragem, mas que representam um segmento importante da população.

Alguns métodos podem ser empregados para a detecção dessas observações, um deles é o gráfico, contudo, seu emprego pode se tornar difícil conforme o número de observações, e de variáveis, aumenta. Uma outra opção são os métodos empíricos, mas cabe ressaltar, que a análise não deve ficar limitada a comparação univariada, utilizando-se também a multivariada. Além disso, pode-se empregar as medidas de similaridade entre as observações e os resultados dos agrupamentos, observando os grupos que tiverem poucas observações. Os últimos devem ser avaliados com cuidado, pois, como mencionado, podem ser um segmento importante, mas que não foi bem representado pela amostra (HAIR Jr. et al., 2009).

2.2.4 Medidas de similaridade

De acordo com Hair Jr. et al. (2009), a comparação entre as observações é fundamental para a formação dos agrupamentos, pois é a partir da similaridade que os objetos são unidos em grupos. Dentre os métodos para realizar essa comparação, três se destacam: medida de correlação, medida de distância e medida de associação. As duas primeiras são usadas para dados métricos, enquanto a última para dados não-métricos.

2.2.4.1 Medidas de correlação

Segundo NETTLETON (2014), essas medidas representam a relação entre duas variáveis, observando o padrão de comportamento entre elas. A relação será positiva se quando uma aumentar, a outra também aumentar, e negativa se uma aumentar enquanto a outra diminuir. Também pode não haver relação entre elas.

O coeficiente de correlação mais comum é o de Pearson, que atribui um valor de -1 a 1 para a relação entre as variáveis, onde -1 é a correlação total negativa, 1 é a correlação total negativa e 0 para quando não há correlação (NETTLETON, 2014). O coeficiente é dado pela equação (1):

$$r_{ij} = \frac{\sum_{k=1}^{p}(x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j})}{\sqrt{\sum_{k=1}^{p}(x_{ik} - \overline{x_i})^2 \sum_{k=1}^{p}(x_{jk} - \overline{x_j})^2}} \quad (1)$$

Sendo:

- x_{ik} = valor da variável k para a observação i;
- x_{ij} = valor da variável k para a observação j;
- $\overline{x_i}$ = representa a média de todas as variáveis para o indivíduo i;
- $\overline{x_j}$ = representa a média de todas as variáveis para o indivíduo j;
- p = representa o número de variáveis.

Contudo, medidas de correlação não costumar ser empregadas na criação dos agrupamentos, devido ao foco da análise não ser as relações dos objetos, mas, sim, suas magnitudes (NETTLETON, 2014).

2.2.4.2 Medidas de distância

Segundo Hair Jr. et al. (2009), as medidas de distância representam a proximidade entre as observações com base em suas variáveis. Todavia, os métodos são medidas de dissimilaridade devido a aumentarem para distâncias maiores, dessa forma, observações similares são representadas por valores baixos.

A medida de distância mais comum é distância euclidiana, ela é pode ser compreendida geometricamente como a linha reta entre dois pontos em um plano, ou, dado um triângulo retângulo, como sua hipotenusa (HAIR Jr. et al., 2009). A distância para duas observações, i e j, e para p variáveis pode ser vista na equação (2):

$$d_{i,j} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \quad (2)$$

De acordo com Hair Jr. et al. (2009), pode-se utilizar também a distância euclidiana quadrática, que é dada pela soma dos quadrados da diferença entre as observações, i e j, e para p variáveis, ou seja, a mesma que a distância euclidiana, mas sem extrair a raiz quadrada. A equação (3) representa essa distância:

$$d_{i,j} = \sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2 \quad (3)$$

Outra medida que pode ser usada na análise de *cluster* é a *City-Block* ou *Manhattan,* ela é dada pela soma da diferença absoluta das variáveis, ou, geometricamente, a soma dos lados do triângulo retângulo. Ela não é indicada para agrupamentos que possuem variáveis altamente correlacionadas, pois tende a gerar resultados inválidos. Contudo, tende a ser eficiente quando o número de variáveis aumenta (HAIR Jr. et al., 2009). Ela é uma derivação da distância de Minkowski, dada pela equação (4):

$$d_{i,j} = \left(\sum_{k=1}^{p} \left(\left|x_{ik} - x_{jk}\right|\right)^n\right)^{\frac{1}{n}} \quad (4)$$

Se aplicarmos $n = 2$, chegamos na distância euclidiana, já para $n = 1$, temos a distância *Manhattan,* dada pela equação (5):

$$d_{i,j} = \sum_{k=1}^{p} \left|x_{ik} - x_{jk}\right| \quad (5)$$

A distância Mahalanobis é outra distância utilizada, ela não é uma distância entre dois pontos distintos, mas entre um ponto e uma distribuição. Ela tem como diferencial um termo que leva em conta as correlações entre as variáveis de uma forma que pesa cada variável igualmente. Dessa maneira, quando há uma forte correlação entre as variáveis, a distância será reduzida pela correlação, já quando a correlação é fraca, o peso da covariância também será.

Além disso, a distância também realiza a padronização das variáveis (PRABHAKARAN, 2019).

A equação da distância é dada por (6):

$$d_{i,j} = \sqrt[2]{(x_i - x_j)^T S^{-1} (x_i - x_j)} \tag{6}$$

A escolha de uma medida de distância passa pela avaliação dos resultados gerados por ela, pois cada medida pode gerar uma solução diferente. Contudo, havendo uma alta correlação, tanto de forma positiva quanto negativa, entre as variáveis, recomenda-se a utilização da distância de Mahalanobis, pois consegue ajustar as correlações presentes (HAIR Jr. et al., 2009).

2.2.4.3 Medidas de associação

Segundo Everitt et al. (2011), uma forma simples de representar a semelhança é através da porcentagem de concordância entre as observações, dada por:

$$S_{ij} = \frac{a + d}{a + b + c + d} \tag{7}$$

Sendo que:

- a representa o número de atributos presentes em ambos os indivíduos;
- b representa o número de atributos presentes em i, mas não em j;
- c representa o número de atributos presentes em j, mas não em i;
- d representa o número de atributos ausentes em ambos os indivíduos.

A equação (7) acima representa a medida de semelhança do coeficiente de correspondência simples, já a medida de distância é dada por seu complementar, ou seja, dada pela equação (8):

$$d_{ij} = \frac{b + c}{a + b + c + d} \tag{8}$$

De acordo com Everitt et al. (2011), outro coeficiente é o de Jaccard, no qual os atributos ausentes comuns não são considerados, dada pela equação (9):

$$S_{ij} = \frac{a}{a + b + c} \tag{9}$$

2.2.5 Padronização dos dados

Segundo Hair Jr. et al. (2009), as variáveis presentes nos dados podem apresentar escalas diferentes, além de grande dispersão. Tais ocorrências podem influenciar a solução e é recomendável que a padronização seja feita antes de calcular as similaridades.

O método mais comum de padronização é a subtração da média e divisão pelo desvio padrão para cada variável. Dessa forma, a variável convertida terá média 0 e desvio de padrão, o que facilita a análise dos dados. Esse método é chamado de *score* padrão ou z-*score*. Outros métodos comuns de padronização são: a divisão da variável pela amplitude, de forma que os valores finais irão de -1 a 1, a divisão pela média, resultado na variável padronizada ter média 1, e a divisão pelo desvio padrão, resultando na variável padronizada ter desvio padrão 1. Os métodos (HAIR Jr. et al., 2009).

2.3 SUPOSIÇÕES PARA A ANÁLISE DE AGRUPAMENTOS

De acordo com Hair Jr. et al. (2009), ao se trabalhar como a análise de agrupamento deve-se sempre lembrar que ela não é método estatístico, mas, sim, uma forma de avaliação das caraterísticas das observações com base nas variáveis escolhidas. Dessa forma, o cuidado com a representatividade da amostra, a pressuposição de uma estrutura e multicolinearidade entre as variáveis, são fundamentais para a obtenção de um bom resultado.

2.3.1 Existência de estrutura

Como a análise tem como objetivo a separação dos dados em grupos, a primeira suposição feita é que há uma diferença significativa entre os dados que permita separá-los em *clusters* e que cada um destes possua caraterísticas distintas (HAIR Jr. et al., 2009).

2.3.2 Representativa da amostra

Segundo Hair Jr. et al. (2009), de forma geral, uma amostra é usada durante a análise. Assim, os agrupamentos gerados no processo refletem os dados de entrada e só terão validade se a amostra for representativa do que é encontrado na população. Os dados utilizados podem conter *outliers,* que podem ser tanto discrepâncias verdadeiras ou apenas falhas durante o processo de amostragem, e sua remoção pode introduzir viés ao resultado e exclusão de um grupo existente na população.

2.3.3 Multicolinearidade

Se houver uma forte relação entre as variáveis, deve-se fazer uma seleção de forma que elas não distorçam o resultado, pois caso um grupo de variáveis tenha uma grande correlação e seja numeroso, ele irá funcionar como um peso na criação dos agrupamentos das observações. Dessa forma, é necessário que as variáveis sejam escolhidas em número igual de acordo com o grau de correlação existente entre elas. Outra forma, é usar uma medida de distância que faça a compensação da correlação existente entre elas (HAIR Jr. et al., 2009).

2.4 FORMAÇÃO DOS AGRUPAMENTOS E AVALIAÇÃO

De acordo com Hair Jr. et al. (2009), a primeira decisão para a realização do agrupamento é qual o método a ser utilizado. A partir do resultado obtido, a próxima etapa é a avaliação de possíveis *outliers* gerados e a reaplicação do método, caso decida-se pela sua eliminação. Por último, a escolha do número de *clusters* da solução, de forma similar, pode-se realizar testes e observar a mudança no resultado.

2.4.1 Algoritmos de agrupamento

Os dois métodos mais comuns de agrupamento são o hierárquico e o não hierárquico. Ambos têm o mesmo objetivo, mas fazem isso de forma diferente. O primeiro começa com todas as observações sendo agrupamentos separados e em cada etapa realiza a junção de dois deles, até que haja apenas um. Já no segundo algoritmo, há a definição prévia do número de agrupamentos e o conjunto de dados é formado com base nessa decisão (HAIR Jr. et al., 2009).

2.4.1.1 Algoritmos hierárquicos

Segundo Hair Jr. et al. (2009), o algoritmo hierárquico envolve uma série de etapas, n – 1, na qual n é o número de observações do conjunto de dados. O processo pode tanto acontecer de forma aglomerativa, em que cada etapa envolve a junção de duas observações em um grupo, até que a obtenção de um único grupo. Também pode-se usar o método divisível, que é o inverso do método aglomerativo, parte-se de um único grupo até chegar em grupos que contenha apenas uma observação.

O processo de aglomeração tem início com todas as observações em agrupamentos separados, a próxima etapa é a junção dos dois agrupamentos mais similares de acordo com a medida escolhida, o ciclo se repete até que um único seja obtido (HAIR Jr. et al., 2009).

De acordo com Hair Jr. et al. (2009), contudo, quando os agrupamentos a serem comparados possuem mais que uma observação, deve-se decidir de que maneira a comparação deve ser realizada. Alguns métodos que podem ser utilizados são: método da ligação individual, método da ligação completa, método da ligação média, método do centroide e método de Ward.

O método da ligação individual, também conhecido como método do vizinho mais próximo, utiliza a distância entre os elementos mais próximos de *cluster* como a similaridade entre os agrupamentos. Esse método tende a gerar grupos encadeados e desiquilibrados (EVERITT, B.S. et al., 2011).

O método da ligação completa, ou método do vizinho mais distante, utiliza a menor distância entre os elementos mais distantes de cada conjunto como a similaridade entre os agrupamentos. Esse método tende a gerar *clusters* compactos, no qual a similaridade dentro do grupo é dada pelo seu diâmetro (HAIR Jr. et al., 2009).

O método da ligação média, utiliza a média das distâncias entre todos os pares de observações de cada grupo como similaridade entre os agrupamentos. Tende a juntar agrupamentos com pequenas variações (EVERITT et al., 2011).

O método do centroide usa a distância entre os centroides dos agrupamentos como medida de similaridade. Como o seu valor é recalculado a cada junção, pode gerar resultados confusos, mas são menos afetados por discrepâncias (HAIR Jr. et al., 2009).

De acordo com Hair Jr. et al. (2009), o método de Ward seleciona os dois aglomerados que tiverem a menor soma de quadrados feita com todas as variáveis como medida de similaridade. Em cada etapa, o par que apresentar a menor soma de quadrados será combinado. Essa alternativa tende a gerar agrupamentos de mesmo tamanho, pois como o método visa a

redução da variação interna, os *clusters* que tiverem poucas observações serão favorecidos. Contudo, a identificação de grupos pequenos torna-se difícil.

$$\sum_{j=1}^{k}\sum_{i\in G_j}\left(x_{i1}-\overline{X_i}\right)^2=\sum_{j=1}^{k}n_j\left(\overline{X}_{ji}-\overline{X}_1\right)^2+\sum_{j=1}^{k}\sum_{i\in G_j}\left(x_{i1}-\overline{X_{j1}}\right)^2 \quad (10)$$

2.4.1.2 Algoritmos não hierárquicos

Esses algoritmos têm como base a definição do número de agrupamentos de maneira que eles possuam coesão interna e diferença externa. O processo consiste na definição de um ponto inicial para o grupo, a semente. A próxima etapa é a atribuição de uma observação a um *cluster* com base na similaridade, conforme o decorrer do processo, as observações podem ser redistribuídas entre as partições (BUSSAB et al., 1990).

Segundo Hair Jr. et al. (2009), a definição da semente pode acontecer tanto de maneira aleatória quanto ser predefinida. O primeiro método tem como vantagem sua simplicidade, mas ao custo de baixo desempenho. Já o segundo envolve certo conhecimento que pode vir tanto de um conhecimento sobre os perfis dos agrupamentos, quanto da análise exploratória dos dados. Um dos meios possíveis é a utilização do método hierárquico, por meio do qual pode-se usar os centroides dos agrupamentos resultantes como sementes.

Dentre as opções de métodos de atribuição, a mais popular é o *k-means*. Ele divide os dados no número de grupos predefinidos e redistribui as observações entre as partições até que algum critério previamente estabelecido seja alcançado (HAIR Jr. et al., 2009).

Existem três opções quanto a atribuição de uma observação a um agrupamento. A primeira é o método sequencial, que consiste na definição de uma semente e a inclusão de todas as observações que estão a uma determinada distância da semente e assim sucessivamente. A segunda, paralela, considera todas as sementes na definição da observação a um grupo, dessa forma, a observação fará parte do grupo da semente com a qual tiver maior proximidade. Por último, o método de otimização permite que a observação mude de agrupamento caso ela se torne mais próxima de uma partição do que a qual foi originalmente direcionada (HAIR Jr. et al., 2009).

2.4.1.3 Decisão sobre a aplicação de cada algoritmo

Segundo Hair Jr. et al. (2009), a escolha entre os métodos passa pela avaliação do problema, bem como as vantagens e desvantagens de cada algoritmo.

O método hierárquico tem como vantagens a sua simplicidade, a ampla gama de medidas de similaridade e a velocidade, é possível analisar diferentes soluções de agrupamentos, seja avaliar outros métodos de ligação, medidas de similaridade ou o número de agrupamentos. Contudo, o método também possui suas desvantagens, incluindo a combinação permanente, não sendo possível a mudança de uma observação para outro agrupamento, a suscetibilidade à presença de valores discrepantes, bem como o alto custo de processamento conforme o número de amostras aumenta (HAIR Jr. et al., 2009).

De acordo com Hair Jr. et al. (2009), o algoritmo não hierárquico, por sua vez, é menos vulnerável a discrepâncias, medidas de distância utilizadas e amostragem incorreta. Outra vantagem é a possibilidade de se trabalhar com um grande volume de dados e a mudança de agrupamento das observações durante o processo. No entanto, para obtenção de bons resultados, melhores do que quando comparado ao hierárquico, é necessário que as sementes sejam definidas, e, mesmo que elas sejam definidas, não há garantias que o resultado obtido seja ótimo, o resultado deve ser validado ao final do processo. Outra desvantagem é a de que o algoritmo *k-means* tende a produzir agrupamentos esféricos e de mesmo tamanho. Por último, o método não hierárquico não é eficiente na criação de diferentes soluções, pois cada solução é uma análise separada, diferentemente do que acontece no hierárquico.

Uma opção é combinação dos dois métodos. Utiliza-se o algoritmo hierárquico para a definição de um conjunto de soluções e do número de agrupamentos, valores discrepantes também podem ser removidos durante essa etapa. As amostras restantes são empregas no método não hierárquico tendo como base as observações da etapa anterior, incluindo a geração das sementes em função dos centroides dos agrupamentos do método hierárquico (HAIR Jr. et al., 2009).

2.4.2 Avaliação dos resultados

Segundo Hair Jr. et al. (2009), antes da avaliação de uma possível solução, é necessário observar as caraterísticas dos agrupamentos formados. Grupos com apenas uma ou poucas observações devem ser analisados, se não representarem um segmento importante dentro do conjunto de dados, devem ser descartados e o algoritmo deve ser aplicado novamente. O processo pode se repetir algumas vezes até a identificação de todos os objetos a serem descartados.

A definição do número de grupos a serem formados é outra questão importante na definição de uma solução. No algoritmo hierárquico pode não haver um indício claro de qual seria o melhor número de agrupamentos, já para o não hierárquico, tal número deve ser previamente definido (HAIR Jr. et al., 2009).

De acordo com Hair Jr. et al. (2009), esse critério é conhecido como regra de parada e tem como base o aumento de heterogeneidade, isto é, a diferença entre as observações presentes em determinado agrupamento. Se houver um aumento elevado nesse valor, o conjunto anterior é um bom candidato para solução final.

As regras de parada podem ser divididas em duas classes, uma delas é a medida de mudança de similaridade e a outra é a medida direta de heterogeneidade. A primeira utiliza algum critério de similaridade e observa a sua variação entre o número de agrupamentos gerados. Caso haja uma alteração nessa medida entre as soluções, tem-se um indício que a versão anterior à junção é uma boa candidata como solução final. Contudo, é comum que haja mais de um ponto com grande mudança de heterogeneidade, para esses casos, a decisão cabe ao pesquisador. A regra mais comum para essa classe é a de mudança percentual na heterogeneidade. Na segunda classe usada para fazer a avaliação, visa-se fazer a medição de heterogeneidade para cada solução de agrupamento para então tomar uma decisão quanto ao número final de *clusters* (HAIR Jr. et al., 2009).

2.5 INTERPRETAÇÃO DOS AGRUPAMENOS

Segundo Hair Jr. et al. (2009), a etapa de interpretação dos grupos gerados tem como objetivo traçar um perfil das soluções geradas de maneira que seja possível comparar as suas caraterísticas entre os *clusters,* bem como entre outras possíveis soluções. Uma opção de análise é uma observação dos centroides dos conjuntos. Caso os dados tenham passado por processamento, os resultados a serem analisados são os das variáveis originais. Contudo, conforme aumenta-se o número de variáveis observadas ou de agrupamentos, a avaliação por meio pode ser difícil. Assim, pode-se empregar um meio visual, como um gráfico para o estudo das particularidades de cada solução.

A interpretação dos agrupamentos também pode servir de auxílio na decisão tomar quando a regra de parada indica mais de uma solução possível. Também é possível utilizar os resultados obtidos para avaliar uma correspondência com os resultados propostos por teoria ou experiência prática. Por último, caso haja a necessidade de uma diferença significativa entre os

conjuntos de variáveis, pode-se expandir a solução até que essa distinção seja evidenciada (HAIR Jr. et al., 2009).

2.6 VALIDAÇÃO E PERFILAMENTO DOS AGRUPAMENTOS

A validação dos agrupamentos é uma etapa fundamental do método, pois a partir dela é possível afirmar que o foi observado dentro da solução representa a população. Um dos métodos para chegar a essa conclusão é a validação cruzada. Por meio da análise de duas amostras distintas, espera-se que a solução para cada uma delas gere resultados similares. Contudo, pode ser custoso a geração de duas ou mais amostras, dificultando esse método. Uma opção, é a divisão da amostra em grupos e a avaliação de cada um deles. Ainda é possível testar diferentes soluções para uma mesma amostra, bem como utilizar os centroides de uma solução na geração de outra para efeito comparativo (HAIR Jr. et al., 2009).

Segundo Hair Jr. et al. (2009), uma solução é dita muito estável se menos de dez por cento das observações são designadas para outro agrupamento. Já para resultados entre dez e vinte por cento as soluções são ditas como estáveis, e para valores entre vinte e vinte e cinco por cento são razoavelmente estáveis.

Outro método de validação é o emprego de outras variáveis, não usadas durante a solução. Nesse caso, elas também devem ter comportamentos diferentes para cada agrupamento e esse comportamento deve ter uma base teórica ou prática (HAIR Jr. et al., 2009).

De acordo com Hair Jr. et al. (2009), uma última etapa, o perfilamento, busca observar as caraterísticas dos agrupamentos através das variáveis não utilizadas, seja durante a solução ou na validação. As opções mais comuns na elaboração dos perfis são o uso de variáveis descritivas, como medidas comportamentais, padrões de consumo ou caraterísticas demográficas, ou variáveis preditivas, que são razões hipotéticas para o agrupamento. A primeira opção baseia-se na relevância prática, já a segunda deve ter um suporte teórico, pois é uma tentativa de entender o impacto daquelas variáveis durante a formação. Seja qual for o método escolhido, espera-se encontrar, por meio de análise de discriminante ou alguma outra técnica, diferenças no comportamento das variáveis entre os *clusters*.

Dessa forma, a finalidade da análise de perfil de cada agrupamento é entender as caraterísticas que compõem cada um por meio de suas variáveis e qual a diferença de seu comportamento através dos *clusters*, o que poderia ser usado para a previsão da atribuição de uma determinada observação a um agrupamento (HAIR Jr. et al., 2009).

3 METODOLOGIA

Este capítulo diz a respeito à fonte de dados que será utilizada durante o estudo, os detalhes de processamento e preparação desses dados, bem como as ferramentas utilizadas para a realização dessas atividades.

3.1 FONTE DE DADOS

Os dados utilizados durante este trabalho são fornecidos pelo Inep (FNDE, 2022), além de dados do censo escolar da educação básica (BRASIL, 2022c), que contém informações sobre as escolas públicas e suas caraterísticas. O conjunto de dados utilizado possui um total de 437 colunas e 35011 linhas. Devido ao grande número de variáveis, optou-se pela seleção de variáveis para que elas representassem os elementos básicos presentes em uma escola, assim, foram empregadas 6 para a realização do agrupamento, conforme a Tabela 1, além de 12 para a validação dos resultados, conforme Tabela 2.

Tabela 1 - Variáveis para agrupamento

Variável	Descrição
MIRD	Média da Regularidade do Corpo Docente
IdeGES	Índice de Desempenho da Gestão Descentralizada
QT_MAT_BAS	Número de Matrículas na Educação Básica
QT_DOC_BAS	Número de Docentes da Educação Básica
QT_TUR_BAS	Número de Turmas de Educação Básica
Eletrônicos	Soma da quantidade de computadores e *tablets*

Fonte: FNDE 2022, BRASIL 2022c

Tabela 2 - Variáveis de validação

Variável	Descrição
NO_UF	Nome da Unidade da Federação
TP_DEPENDENCIA	Dependência Administrativa

TP_LOCALIZACAO	Localização
IN_BIBLIOTECA_SALA_LEITURA	Dependências físicas existentes e utilizadas na escola - Biblioteca e/ou Sala de leitura
IN_ALIMENTACAO	Alimentação escolar para os alunos - PNAE/FNDE
IN_QUADRA_ESPORTES	Dependências físicas existentes e utilizadas na escola - Quadra de esportes coberta ou descoberta
IN_SALA_MULTIUSO	Dependências físicas existentes e utilizadas na escola - Sala multiuso (música, dança e artes)
IN_ENERGIA_REDE_PUBLICA	Abastecimento de energia elétrica - Rede pública
IN_AGUA_POTAVEL	Fornece água potável para o consumo humano
IN_BANHEIRO	Dependências físicas existentes e utilizadas na escola - Banheiro
IN_ESGOTO_REDE_PUBLICA	Esgoto sanitário - Rede pública
IN_INTERNET	Acesso à Internet

Fonte: FNDE 2022, BRASIL 2022c

3.2 FERRAMENTAS

A linguagem de programação Python foi utilizada durante a parte prática do projeto em conjunto com algumas bibliotecas. A biblioteca Pandas permite a fácil manipulação de dados estruturados, que tem como base o NumPy (HARRIS et al, 2020), e pode ser combinada com a Matplolib (HUNTER, 2007) que possibilita a construção de gráficos. Ainda foi usada a biblioteca Statsmodels (SEABOLD; PERKTOLD, 2010), que fornece modelos e testes estatísticos, além da Scipy (VIRTANEN et al, 2020) e da Sklearn (PEDREGOSA et al, 2011) para a implementação do algoritmo de agrupamento.

3.3 APLICAÇÃO

A partir dos dados selecionados, apenas da região Sudeste e que não continham valores nulos, o método hierárquico foi empregado de duas formas diferentes. A primeira foi a combinação da distância euclidiana quadrática com o método de agrupamento completo, já a segunda usou a distância euclidiana com o método de agrupamento de Ward. Em ambos os

casos foram utilizados a padronização dos dados por meio do z-*score*, além de terem sido gerados dois agrupamentos em cada caso, um com três grupos e outro com quatro.

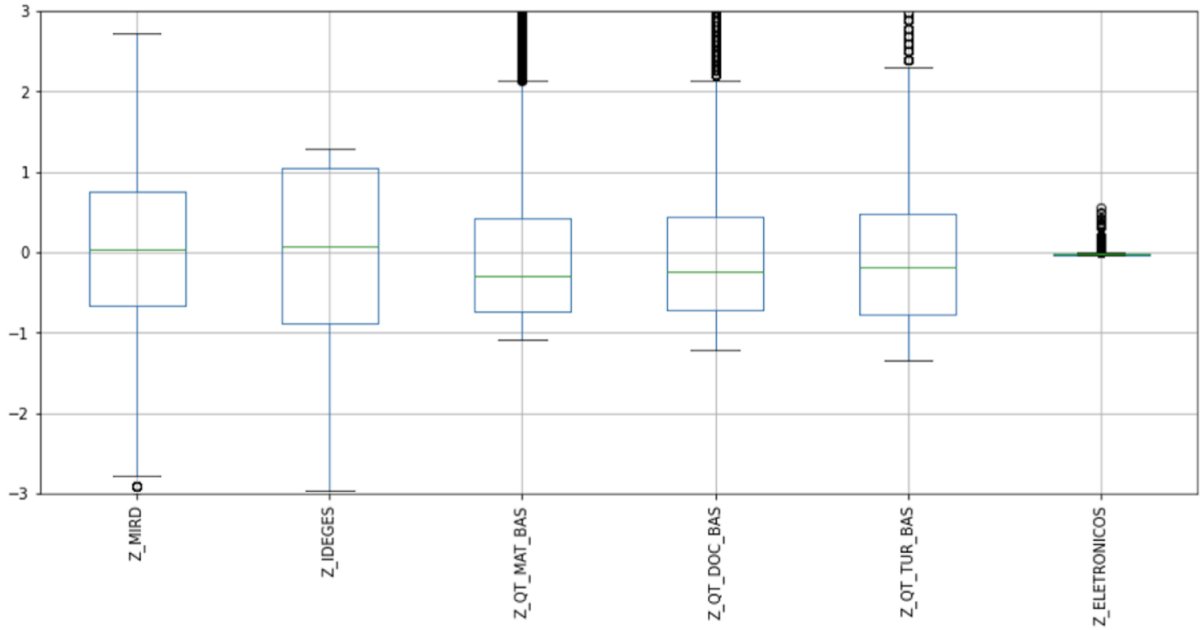
O método não hierárquico foi aplicado em seguida, apenas com os resultados dos agrupamentos de quatro grupos, usando os centroides dos agrupamentos obtidos como pontos iniciais. Dessa forma, obteve-se duas soluções possíveis para o agrupamento dos dados.

Tendo os dois resultados dos agrupamentos, buscou-se fazer a interpretação dos resultados e verificar se os grupos gerados em cada um dos agrupamentos eram diferentes um dos outros com base nas caraterísticas definidas para o emprego do método. Em seguida, foi feita a análise de perfil com base em variáveis que não foram utilizadas no agrupamento. A última etapa foi a utilização do método ANOVA para a avaliação da importância das variáveis durante a formação dos *clusters*.

4 RESULTADOS

Sabendo do objetivo do trabalho, aplicar os fundamentos e técnicas de análise de agrupamentos em uma base dados educacionais, a saber, escolas públicas da educação básica da região Sudeste do Brasil, gerando grupos homogêneos de escolas, utilizou-se das variáveis presente na Tabela 1 no agrupamento. Contudo, antes de usá-las, é necessário fazer uma análise.

Figura 1 - Distribuição das variáveis padronizadas



Fonte: autoria própria

Conforme visto na Figura 1, a média da regularidade do corpo docente apresenta metade dos seus valores concentrados entre o intervalo de -1 a 1, assim como as demais variáveis, excetuando a de eletrônicos. Contudo, a outra metade fica bem distribuída no restante do intervalo, além de conter poucas discrepâncias, estas são encontradas apenas no limite inferior. O índice de desempenho da gestão descentralizada possui o maior intervalo entre o primeiro quartil e o terceiro e de forma similar à MIRD possui as observações que vão até o primeiro quartil bem distribuídas, mas as acima do terceiro quartil são concentradas em um pequeno intervalo. Além disso, não possui *outliers*.

Para a quantidade de matrículas, de docentes e de turmas, o comportamento é bem similar entre si, o intervalo interquartil é concentrado entre -1 e 0,5, possuindo metade das suas observações um pouco abaixo da média. O restante das observações fica concentradas no limite superior e possui muitas discrepâncias. A quantidade de eletrônicos possui uma distribuição bastante concentrada, ficando próxima à média, excetuando-se apenas os *outliers*.

Entendendo a distribuição das variáveis, optou-se pela preservação de todas as observações devido à possibilidade de os valores com maior discrepância em relação à média formarem grupos com características únicas e, dessa maneira, um perfil importante poderia vir a ser descartado.

Além disso, assume-se que as observações fazem parte de determinados grupos a serem descobertos, que a quantidade de observações é representativa da população e que não há multicolinearidade entre as variáveis.

Os dois primeiros agrupamentos foram gerados a partir do método hierárquico utilizando a distância euclidiana quadrática com o método de agrupamento completo. O primeiro agrupamento, mostrado na Tabela 3, gerou três grupos com comportamentos distintos. O primeiro grupo, 0, é caraterizado por ser centrado na média, apresentada como 0 devido à padronização, e tem 34988 observações. Já o segundo, 1, é caracterizado pelo alto número de eletrônicos e é composto de 18 observações. O último tem como caraterística o alto número de docentes e tem apenas 5 observações. A Figura 2 oferece outro a mudança do comportamento em cada grupo.

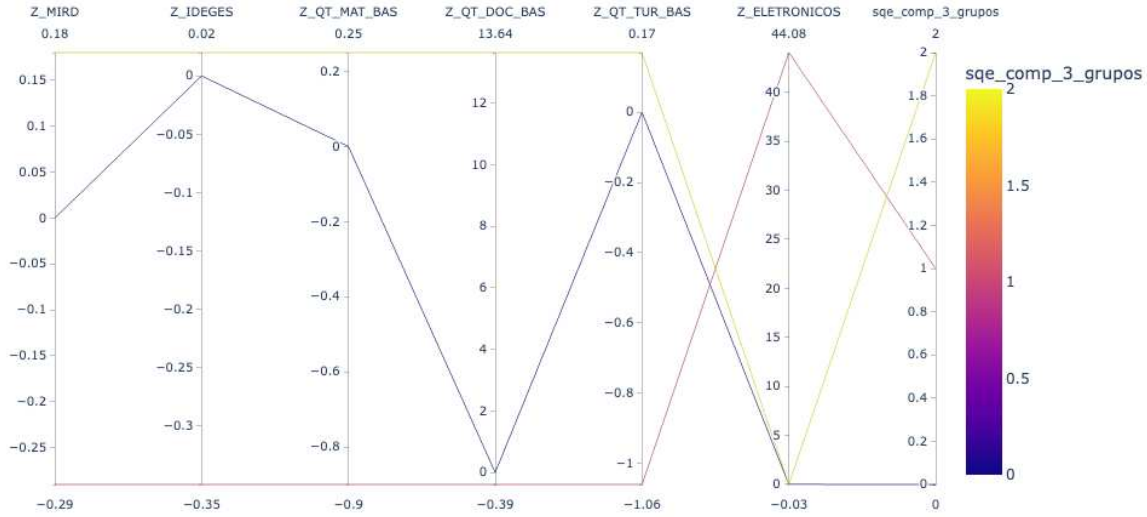
Tabela 3 - Centroides dos três grupos a partir do método completo

GRUPO	Z_MIRD	Z_IDEGES	Z_QT_MAT_BAS	Z_QT_DOC_BAS	Z_QT_TUR_BAS	Z_ELETRONICOS

0	0.00	0.00	0.00	-0.00	0.00	-0.02
1	-0.29	-0.35	-0.90	-0.39	-1.06	44.08
2	0.18	0.02	0.25	13.64	0.17	-0.03

Fonte: autoria própria.

Figura 2 - Centroides dos três grupos a partir do método completo



Fonte: autoria própria

De forma similar, foi feito um agrupamento composto de quatro grupos. O *cluster* 0 da etapa anterior foi dividido em dois novos *clusters*, o primeiro, 0, tem como característica o alto número de matrículas e é composto de 33 observações. O *cluster 1,* ficou com o restante das observações, 34955, e tem o mesmo comportamento do *cluster 0* gerado pelo agrupamento com três grupos. Os *clusters* 2 e 3, não foram alterados e são os mesmos representados pelos *clusters* 1 e 2 do agrupamento anterior. Os centroides podem ser vistos no formato tabular, Tabela 4, e pela Figura 3.

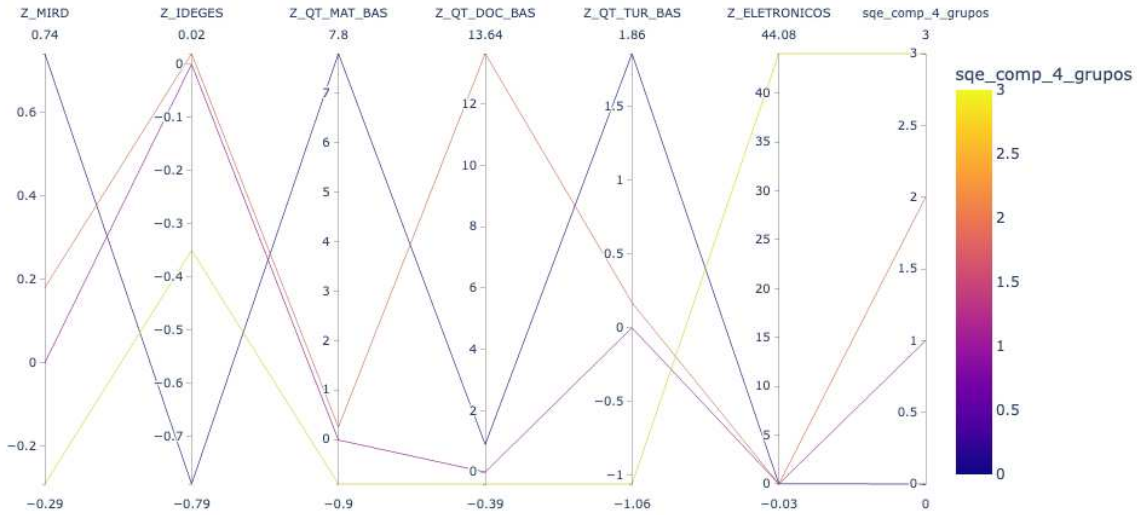
Tabela 4 - Centroide dos quatros grupos a partir do método completo

GRUPO	Z_MIRD	Z_IDEGES	Z_QT_MAT_BAS	Z_QT_DOC_BAS	Z_QT_TUR_BAS	Z_ELETRONICOS
0	0.74	-0.79	7.80	0.90	1.86	-0.01
1	-0.00	0.00	-0.01	-0.00	-0.00	-0.02
2	0.18	0.02	0.25	13.64	0.17	-0.03

3	-0.29	-0.35	-0.90	-0.39	-1.06	44.08

Fonte: autoria própria.

Figura 3 - Centroide dos quatros grupos a partir do método completo



Fonte: autoria própria.

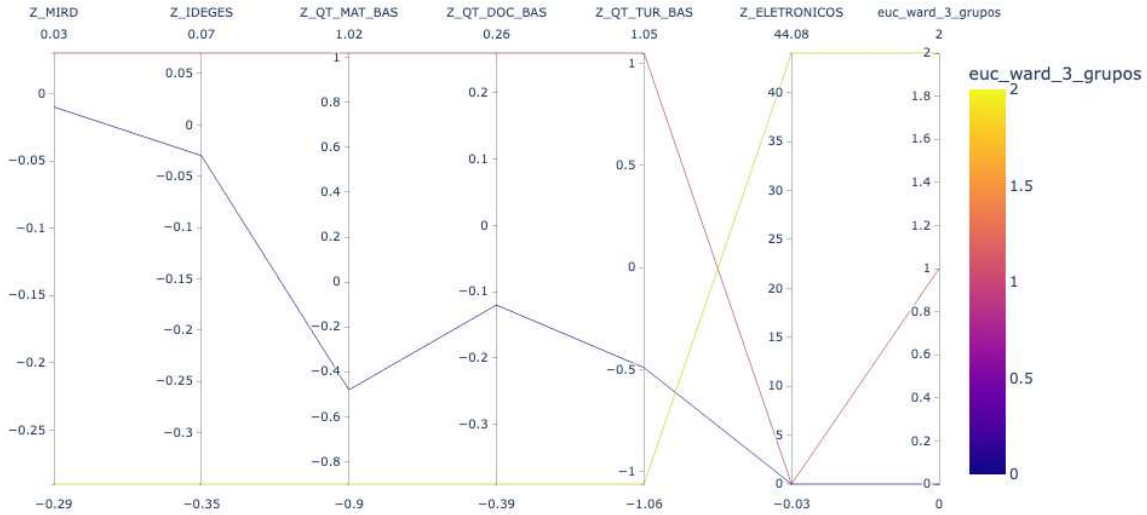
Os dois próximos agrupamentos foram gerados a partir do método hierárquico utilizando a distância euclidiana e método de Ward. O primeiro agrupamento, mostrado na Tabela 5, gerou três grupos com comportamentos distintos. O primeiro grupo, 0, é caraterizado por ter um menor número de matrículas e de turmas quando comparado com a média, e tem 23864 observações. Já o segundo, 1, tem comportamento oposto, apresentando um maior número de matrículas e de turmas em relação à média, tendo 11129 observações. O último é caracterizado pelo alto número de eletrônicos e é composto de 18 observações. A Figura 4 mostra o comportamento em cada grupo.

Tabela 5 - Centroide dos três grupos a partir do método de Ward

GRUPO	Z_MIRD	Z_IDEGES	Z_QT_MAT_BAS	Z_QT_DOC_BAS	Z_QT_TUR_BAS	Z_ELETRONICOS
0	-0.01	-0.03	-0.48	-0.12	-0.49	-0.03
1	0.03	0.07	1.02	0.26	1.05	-0.02
2	-0.29	-0.35	-0.90	-0.39	-1.06	44.08

Fonte: autoria própria.

Figura 4 - Centroide dos três grupos a partir do método de Ward



Fonte: autoria própria.

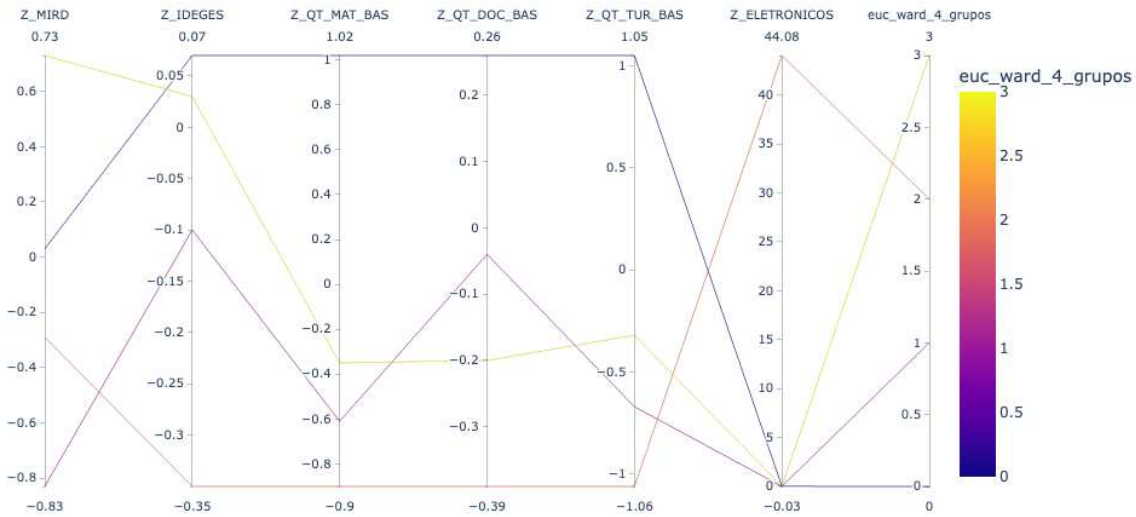
De modo similar foi feito um agrupamento com quatro *clusters*. O primeiro *cluster*, 0, é o mesmo do *cluster* 1 gerado no agrupamento anterior. Já os *clusters* 1 e 3, foram divididos a partir do *cluster* 0 anterior, ainda possuindo um menor número de matrículas e turmas, mas o primeiro tem uma MIRD menor em relação à média, enquanto o segundo possui uma maior, o número de observações é de 11358 e 12506 para os *clusters* 1 e 3, respectivamente. O *cluster* 2 não foi alterado. Os valores dos centroides podem ser vistos na Tabela 6 e o comportamento geral na Figura 5.

Tabela 6 - Centroide dos quatro grupos a partir do método de Ward

GRUPO	Z_MIRD	Z_IDEGES	Z_QT_MAT_BAS	Z_QT_DOC_BAS	Z_QT_TUR_BAS	Z_ELETRONICOS
0	0.03	0.07	1.02	0.26	1.05	-0.02
1	-0.83	-0.10	-0.61	-0.04	-0.67	-0.03
2	-0.29	-0.35	-0.90	-0.39	-1.06	44.08
3	0.73	0.03	-0.35	-0.20	-0.32	-0.02

Fonte: autoria própria.

Figura 5 - Centroide dos quatro grupos a partir do método de Ward



Fonte: autoria própria.

Todos os quatro agrupamentos gerados são formados por grupos heterogêneos. Contudo, quando o método completo foi aplicado houve a concentração das observações em um único grupo. Já quando o método de Ward foi usado, a distribuição foi mais igualitária. Ademais, um grupo que tem como caraterística o alto número de eletrônicos em relação à média foi gerado em todos os agrupamentos.

Os agrupamentos com quatro *clusters* foram escolhidos para serem usados no método não hierárquico por apresentarem uma melhor distribuição das observações entre os grupos. O resultado da aplicação do método *k-means* gerou uma significativa melhoria nessa repartição no caso do método completo, mas piorou quando se avalia o método de Ward.

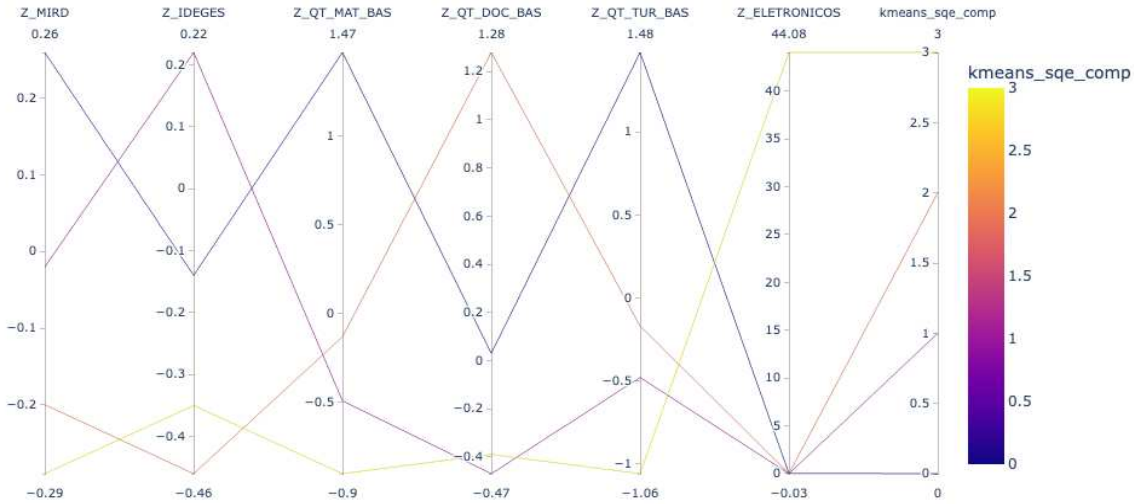
Para o primeiro caso, usando o método completo, 7445 observações passaram do antigo grupo 1 para o novo grupo 0, enquanto 7209 mudaram para o novo grupo 2, sendo que 20301 permaneceram no de número 1 e o de número 3 não foi alterado. O novo grupo 0 tem características bem distintas, tendo uma MIRD maior que a média, além da maior quantidade de matrículas e turmas, quando comparado com a média. O grupo 1 possui o menor número de docentes quando comparado à média e o maior IdeGES, também possui um número relativamente baixo de matrículas e turmas. Já o grupo 2 tem o maior número de docentes e o menor IdeGES. O grupo 3 tem como caraterística o alto número de eletrônicos. Os valores dos centroides podem ser vistos na Tabela 7 e a comparação entre os grupos pode ser vista na Figura 6.

Tabela 7 - Centroide dos grupos a partir do método completo e do *k-means*

GRUPO	Z_MIRD	Z_IDEGES	Z_QT_MAT_BAS	Z_QT_DOC_BAS	Z_QT_TUR_BAS	Z_ELETRONICOS
0	0.26	-0.14	1.47	0.03	1.48	-0.01
1	-0.02	0.22	-0.49	-0.47	-0.48	-0.03
2	-0.20	-0.46	-0.13	1.28	-0.17	-0.02
3	-0.29	-0.35	-0.90	-0.39	-1.06	44.08

Fonte: autoria própria.

Figura 6 - Centroide dos grupos a partir do método completo e do *k-means*



Fonte: autoria própria.

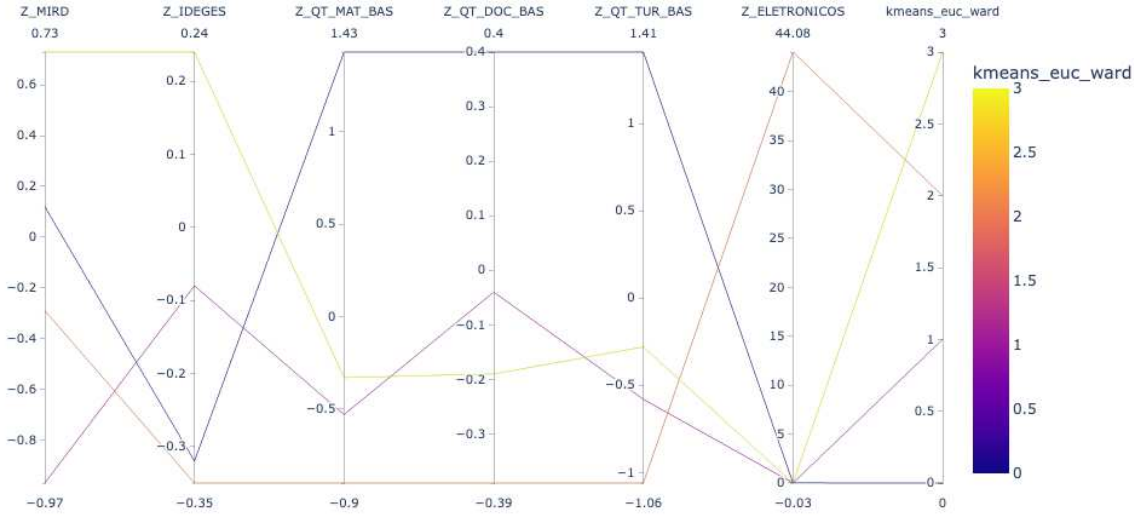
No segundo caso, usando os centroides do agrupamento com o método de Ward, houve algumas alterações das observações entre os grupos. O *cluster* 0 teve seu número reduzido para 7953 observações, antes era 11129, enquanto os de número 1 e 3 passaram para 12144 e 14986, tendo 11358 e 12506, anteriormente e respectivamente. O comportamento geral de cada *cluster* sofreu poucas mudanças. O de número 0 teve seus centroides aumentados, exceto pelo IdeGES, que foi reduzido e chegou a ficar abaixo da média. Mudança similar ocorreu no *cluster* 1, em que a quantidade de matrículas, a quantidade de turmas e o IdeGES sofreram um leve aumento, tendo apenas a MIRD reduzida. O *cluster* 2 continuou o mesmo. Já o último, 3, sofreu mudanças similares ao *cluster* 1, sendo o único que possui o IdeGES acima da média, além de ter a MIRD bem acima dos demais *clusters*. Os valores dos centroides podem ser vistos na Tabela 8 e a comparação entre os grupos pode ser vista na Figura 7.

Tabela 8 - Centroide dos grupos a partir do método de Ward e do *k-means*

GRUPO	Z_MIRD	Z_IDEGES	Z_QT_MAT_BAS	Z_QT_DOC_BAS	Z_QT_TUR_BAS	Z_ELETRONICOS
0	0.03	0.07	1.02	0.26	1.05	-0.02
1	-0.83	-0.10	-0.61	-0.04	-0.67	-0.03
2	-0.29	-0.35	-0.90	-0.39	-1.06	44.08
3	0.73	0.03	-0.35	-0.20	-0.32	-0.02

Fonte: autoria própria.

Figura 7 - Centroide dos grupos a partir do método de Ward e do *k-means*



Fonte: autoria própria.

A validação dos agrupamentos gerados foi feita com variáveis que não foram utilizadas anteriormente. Para isso, foram utilizadas as variáveis que indicavam o nome da unidade da federação, o tipo de dependência administrativa, o tipo de localização, se havia biblioteca ou sala de leitura, cozinha, quadra de esportes, sala multiuso, alimentação, energia oriunda da rede pública, água potável, banheiro, esgoto oriundo da rede pública e internet.

Em ambos os agrupamentos, as variáveis de alimentação, sala multiuso, energia, água potável, banheiro e internet tiveram comportamentos bem similares entre os grupos. Nas unidades federativas, houve uma maior diferença entre os grupos baseado no método de Ward do que os no método completo. O comportamento do tipo de dependência foi similar, o método completo possui dois *clusters* de composição similar, enquanto o método de Ward teve todos os *clusters* diferentes, inclusive um com maioria estadual, o que não aconteceu no anterior. O tipo de localização gerou um *cluster* quase completamente urbano em ambos os agrupamentos,

além do *cluster* dos eletrônicos ser composto de maioria rural. Os dois grupos restantes possuem composição diferentes entre si em ambos os métodos. Uma maior diferença é percebida na variável que indica se há quadra de esporte, enquanto o método de Ward cria apenas um grupo em que ela existe na maioria das observações, já o método completo gera este grupo e mais um. A presença de esgoto é bastante distinta entre os grupos, contudo em todos eles a maioria das observações o possuem. Além disso, três dos quatro grupos em cada agrupamento possuem comportamento igual, apenas um sendo diferente. Ainda cabe ressaltar que o grupo dos eletrônicos é o que possui com o menor índice.

A última análise feita no trabalho foi a avaliação da importância das variáveis para a formação dos agrupamentos por meio do método ANOVA. A variável de maior importância em ambos os casos foi a que indicava a presença de eletrônicos, inclusive gerando um grupo caracterizado pela sua alta presença, mas contendo poucas observações. Duas outras variáveis tiveram grande importância nos dois agrupamentos, as que indicavam a quantidade de matrículas e de turmas. Cada agrupamento ainda possuiu uma variável específica, o método completo deu uma maior importância para a quantidade de docentes, enquanto o método de Ward teve maior influência da média da regularidade do corpo docente (MIRD). O IdeGES teve pouco impacto nos dois casos, mas para o método completo ela teve maior impacto que a MIRD.

5 CONSIDERAÇÕES FINAIS

O presente trabalho tinha como objetivo aplicar os fundamentos e técnicas de análise de agrupamentos em uma base dados educacionais, a saber, escolas públicas da educação básica da região Sudeste do Brasil, gerando grupos homogêneos de escolas. Para isso, a combinação do algoritmo hierárquico com o não hierárquico foi utilizada, sendo que na utilização do primeiro optou-se pela utilização de dois métodos diferentes, o completo e o de Ward, e a avaliação das diferenças apresentadas pelos gerados por cada um dos métodos.

Os dois agrupamentos gerados cumpriram com o objetivo do trabalho, sendo possível observar diferenças nos perfis por meio das variáveis que foram utilizadas no algoritmo, mas também com as variáveis de validação. Dessa forma, os dois métodos podem ser utilizados para a simplificação dos dados escolares e o aprofundamento do estudo das características que os compõem para compreender quais as diferenças existentes entre eles.

Porém, uma limitação é o custo computacional do algoritmo hierárquico, sendo que o volume de dados foi relativamente pequeno, apenas 35011 linhas. Uma opção que pode ser

analisada em outros trabalhos é comparação de resultados entre a combinação dos dois algoritmos e a utilização de apenas o algoritmo não hierárquico com pontos iniciais aleatórios.

REFERÊNCIAS

BRASIL. Ministério da Educação. Fundo Nacional de Desenvolvimento da Educação. **Monitore o PDDE**, 2022a. Disponível em: <https://www.gov.br/fnde/pt-br/acesso-a-informacao/acoes-e-programas/programas/ monitore-o-pdde>. Acesso em: 20 jun. 2022.

BRASIL. Ministério da Educação. Fundo Nacional de Desenvolvimento da Educação. **PDDE**, 2022b. Disponível em: <https://www.gov.br/fnde/pt-br/acesso-a-informacao/acoes-e-programas/programas/pdde>. Acesso em: 20 jun. 2022.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP. **Censo Escolar**: Microdados do Censo Escolar da Educação Básica, 2022c. Disponível em: < https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar>. Acesso em: 20 jun. 2022.

FNDE. **Programa Dinheiro Direto na Escola – PDDE**: Dados e Recursos. 2022. Disponível em: <https://www.fnde.gov.br/dadosabertos/dataset/basico_escola>. Acesso em: 09 mai. 2022.

HAIR Jr., J. F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009.

METZ, Jean. **Interpretação de clusters gerados por algoritmos de clustering hierárquico**. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2006.

NETTLETON, David. **Commercial data mining**: processing, analysis and modeling for predictive analytics projects. Waltham: Elsevier, 2014.

PRABHAKARAN, Selva. Mahalanobis Distance: Understanding the math with examples (python). **Machine learning plus**, 2019. Disponível em: <https://www.machinelearningplus.com/statistics/mahalanobis-distance/>. Acesso em: 28 jun. 2022.

EVERITT, B.S. et al. **Cluster Analysis**. 5. ed. Londres: Wiley, 2011.

BUSSAB, W.O. et al. **Introdução à Análise de Agrupamentos**. 9º Simpósio Brasileiro de Probabilidade e Estatística. São Paulo: IME – USP, 1990.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, [s. l.], v. 9, n. 3, p. 90 – 95, 2007.

HARRIS, C.R. et al. Array programming with NumPy. **Nature**, [s. l.], v. 585, n. 7825, p. 357 – 362, 2020.

SEABOLD, Skipper; PERKTOLD, Josef. statsmodels: Econometric and statistical modeling with python. **Proceedings of the 9th Python in Science Conference**, 2010.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. **Nature Methods**, [s. l.], v. 17, n. 3, p. 261 – 272, 2020.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, [s. l], v. 12, p. 2825 – 2830, 2011.