MODELOS DE PROGRAMAÇÃO MATEMÁTICA PARA APRENDIZADO NÃO SUPERVISIONADO E SUAS APLICAÇÕES NA CLUSTERIZAÇÃO DE DADOS DE ESCOLAS BRASILEIRAS

MATHEMATICAL PROGRAMMING MODELS FOR UNSUPERVISED LEARNING AND THEIR APPLICATIONS IN THE CLUSTERING OF BRAZILIAN SCHOOL DATA

MODELOS DE PROGRAMACIÓN MATEMÁTICA PARA EL APRENDIZAJE NO SUPERVISADO Y SUS APLICACIONES EN LA AGRUPACIÓN DE DATOS DE ESCUELAS BRASILEÑAS

Victor Augusto do Carmo Duarte[1], Erito Marques de Souza Filho[2]

[1] Laboratório Nacional de Computação Científica (LNCC), Programa de Pós-Graduação em Modelagem Computacional, Petrópolis, RJ, Brasil.
[2] Universidade Federal Fluminense (UFF), Programa de Pós-Graduação em Ciências Cardiovasculares, Niterói, RJ, Brasil.

Data de submissão: 24 de agosto de 2024. Data de aprovação: 28 de fevereiro de 2025. Financiamento: os autores declaram não haver financiamento. Como citar: DUARTE, Victor Augusto do Carmo; SOUZA FILHO, Erito Marques de. Modelos de programação matemática para aprendizado não supervisionado e suas aplicações na clusterização de dados de escolas brasileiras. **REMAT: Revista Eletrônica da Matemática**, Bento Gonçalves, RS, v. 11, p. e301, 5 maio 2025. https://doi.org/10.35819/remat2025v11id7421.


Este artigo está licenciado sob uma licença *Creative Commons Attribution 4.0 International License*.

Resumo: A análise de dados educacionais é importante para compreender o desempenho das instituições de ensino e identificar áreas para melhorias. Nesse contexto, a clusterização de dados é um recurso amplamente utilizado, em particular com algoritmos modelados como problemas de programação matemática. Neste trabalho, é proposta a utilização e a implementação de três algoritmos de aprendizado não supervisionado, modelados com Programação Inteira Binária e Programação Linear Inteira Mista, para clusterização de dados sobre o desempenho médio de escolas brasileiras do Exame Nacional do Ensino Médio, divulgados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Tem-se por objetivo validar os modelos por investigar as características das instituições em cada cluster, contrapondo seu Indicador de Nível Socioeconômico e sua dependência administrativa a seu desempenho escolar. Os resultados encontrados apontam o desempenho superior de escolas públicas federais e escolas privadas quando comparadas a escolas públicas municipais e estaduais.
Palavras-chave: aprendizado não supervisionado; clusterização; programação inteira binária; programação linear inteira mista; dados educacionais brasileiros.

Abstract: The analysis of educational data is essential for understanding the performance of educational institutions and identifying areas for improvement. In this context, data clustering is a widely used tool, particularly with algorithms modeled as mathematical programming problems. This paper proposes the use and implementation of three unsupervised learning algorithms, modeled with Binary Integer Programming and Mixed-Integer Linear Programming, for clustering data on the average performance of Brazilian schools in the National High School Exam, published by the National Institute for Educational Studies and Research Anísio Teixeira. The aim is to validate the models by investigating the characteristics of the institutions in each cluster, comparing their Socio-Economic Level Indicator and their administrative

dependence with their school performance. The results found point to the superior performance of federal public schools and private schools when compared to municipal and state public schools.
Keywords: unsupervised learning; clustering; binary integer programming; mixed-integer linear programming; educational data.

Resumen: El análisis de datos educativos es fundamental para comprender el desempeño de las instituciones educativas e identificar áreas de mejora. En este contexto, la agrupación de datos es una herramienta ampliamente utilizada, especialmente con algoritmos modelados como problemas de programación matemática. Este trabajo propone el uso e implementación de tres algoritmos de aprendizaje no supervisado, modelados con Programación Entera Binaria y Programación Lineal Entera Mixta, para agrupar datos sobre el desempeño promedio de las escuelas brasileñas en el Examen Nacional de Educación Secundaria, publicados por el Instituto Nacional de Estudios e Investigaciones Educativas Anísio Teixeira. El objetivo es validar los modelos investigando las características de las instituciones en cada grupo, contrastando su Indicador de Nivel Socioeconómico y su dependencia administrativa con su desempeño escolar. Los resultados indican un desempeño superior de las escuelas públicas federales y privadas en comparación con las escuelas públicas municipales y estatales.
Palabras clave: aprendizaje no supervisado; agrupación; programación entera binaria; programación lineal entera mixta; datos educativos.

1 INTRODUÇÃO

A análise de dados educacionais é de extrema importância para compreendermos o panorama da educação em um país e identificar possíveis melhorias (Fonseca; Namen, 2016). No Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) disponibiliza dados abertos do Enem (Exame Nacional do Ensino Médio por Escola, que podem ser utilizados para *insights* valiosos sobre o desempenho das instituições de ensino (Brasil, 2024; Francelino; Machado, 2020). No contexto da análise de dados, a clusterização é uma técnica que visa agrupar objetos de um conjunto nos chamados *clusters* (Maia; Andrade; Fernandes, 2021), subconjuntos disjuntos e não vazios cuja união resulta no conjunto original, de modo que objetos dentro de um mesmo *cluster* sejam tão homogêneos quanto possível, e objetos de *clusters* distintos sejam tão dissimilares quanto possível. Como afirmam Ágoston e E.-Nagy (2023), o uso de Programação Matemática na modelagem de algoritmos para clusterização é uma abordagem consolidada, mas novas técnicas continuam a ser desenvolvidas (Werner, 2022). Neste artigo, apresentamos a modelagem de algoritmos de aprendizado não supervisionado com Programação Inteira Binária (PIB) e Programação Linear Inteira Mista (PLIM) e sua aplicação na clusterização de dados do INEP sobre o Enem por Escola com o objetivo de investigar a validez dos métodos utilizados contrapondo seus resultados à realidade das disparidades socioeconômicas do país e seus impactos no desempenho escolar.

Nosso objetivo principal é propor a utilização e implementar modelos de clusterização baseados em PLIM e PIB em dados abertos do Enem por Escola divulgados pelo INEP. Os modelos visam identificar padrões de desempenho das escolas e agrupá-las em *clusters* significativos, possibilitando uma análise mais detalhada e uma melhor compreensão do cenário educacional

do país. Para alcançar esse objetivo principal, são estabelecidos objetivos específicos. Primeiro, pretende-se adequar os modelos de PLIM para clusterização de Awasthi et al. (2015), Sağlam et al. (2006) e Werner (2022) ao contexto dos dados educacionais de escolas brasileiras. Em seguida, desenvolver algoritmos computacionalmente viáveis para a solução dos modelos de PLIM usando o solver de licença *Gurobi Optimizer*. Por fim, executar instâncias dos algoritmos com os dados do Enem por Escola e analisar a divisão resultante das instituições de ensino. De posse dessas informações, pode-se propor uma validação dos modelos apresentados por meio da análise crítica dos resultados obtidos, da sua contraposição à realidade socioeconômica brasileira e da comparação com abordagens tradicionais de clusterização.

2 REFERENCIAL TEÓRICO

Como apontam Maia, Andrade e Fernandes (2021) e Maschio et al. (2018), a clusterização é útil em diversas áreas, incluindo análise de dados educacionais, em que pode ser aplicada para a identificação de padrões de desempenho escolar e agrupamento de instituições com características semelhantes. No trabalho de Maschio et al. (2018), afirma-se que métodos para exploração de dados oriundos de ambientes educacionais podem permitir descobrir novos conhecimentos e auxiliar na investigação de questões como o risco de evasão, a formação de grupos e a análise de comportamentos. Além disso, é possível observar que, predominantemente, são utilizadas técnicas de aprendizado de máquina e de agrupamento na solução dessas questões em detrimento de abordagens alternativas como regressão logística e regressão linear (Maschio et al., 2018).

No desenvolvimento de sua pesquisa, Maia, Andrade e Fernandes (2021) investigaram correlações entre o desempenho individual no Enem e a realidade socioeconômica dos candidatos. Foram levadas em consideração variáveis como a autodeclaração étnico-racial, renda familiar mensal, quantidade de pessoas residindo no domicílio do candidato, opção por cotas, entre outras. Utilizando o algoritmo *K-means* para dois *clusters*, foi constatado que indivíduos com um melhor desempenho tendiam a estar em um cluster, enquanto que indivíduos com um pior desempenho tendiam ao outro. Com a análise de ambos os *clusters*, foram feitas observações sobre o perfil socioeconômico dos candidatos que os compunham. Da amostra de 919 alunos, 471 constituíram o *cluster* 0 (com desempenho inferior no Enem) e 448 constituíram o *cluster* 1 (com desempenho superior no Enem). Seu trabalho concluiu que a parcela negra, parda e indígena da população brasileira, que sofre de forma mais acentuada com a desigualdade socioeconômica do país, tem seu desempenho escolar afetado negativamente por ela.

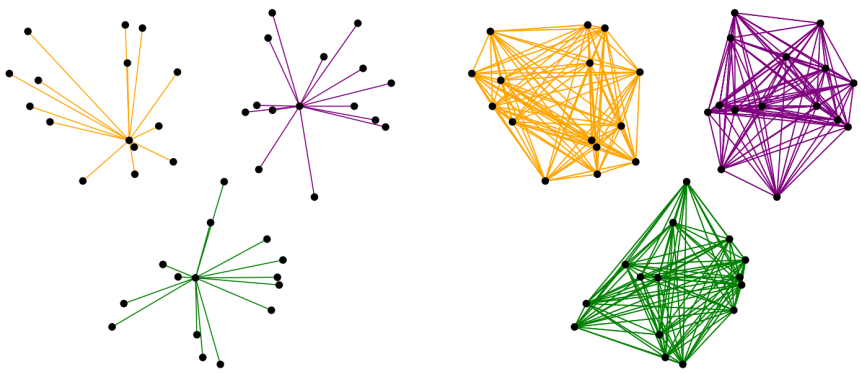
Fonseca e Namen (2016), por sua vez, utilizaram dados divulgados pelo INEP sobre o Sistema de Avaliação da Educação Básica (SAEB) para extrair “padrões que relacionam o perfil dos professores com o resultado obtido por seus alunos nas provas de Matemática”. Fazendo uso do algoritmo *Naive Bayes*, Fonseca e Namen (2016) buscaram identificar quais atributos mais

influenciavam a classificação de professores segundo o melhor ou pior desempenho de seus discentes. Foram consideradas variáveis como a carga horária semanal do docente, seu salário bruto, o tempo decorrido desde sua última formação, entre outras, além de dados a respeito do desempenho dos alunos nos testes de Língua Portuguesa e de Matemática da Prova Brasil. Uma de suas conclusões foi que um baixo salário recebido pelo educador influencia negativamente a proficiência dos alunos (Fonseca; Namen, 2016).

Com o foco no Ensino Superior, Francelino e Machado (2020) trabalharam com o algoritmo *K-means* e o aplicaram à base de dados do INEP referente ao Exame Nacional de Desempenho de Estudantes (ENADE). Explorando variáveis como estado civil, renda familiar, horas de trabalho semanal e outras, Francelino e Machado (2020) traçaram perfis dos indivíduos alocados em diferentes *clusters* construídos partindo de seus desempenhos no ENADE e puderam apontar algumas características dos estudantes que impactam seu resultado no exame. Vale destacar que os autores afirmaram que sua opção pela métrica da "distância euclidiana é devida aos dados não serem padronizados, tornando o resultado final insensível a *outliers*" (Francelino; Machado, 2020).

De acordo com Awasthi et al. (2015), modelagens com Programação Matemática de algoritmos de clusterização como o *K-means* e o *K-medians* têm sido extensivamente estudadas. Os autores apresentam modelos de PLIM para ambos os algoritmos. Enquanto seu *K-means* busca minimizar a média dos quadrados das distâncias de todos os pontos dois a dois dentro de um cluster, seu *K-medians* objetiva minimizar o somatório das distâncias de cada ponto a seu ponto representativo, como ilustrado na Figura 1. Em particular, Awasthi et al. (2015) apresentam a modelagem do *K-medians* enquanto um problema de Programação Inteira Binária que é utilizada neste trabalho.

Figura 1 – Representação da clusterização via *K-medians* (à esquerda) e *K-means* (à direita)



Fonte: Awasthi et al. (2015, p. 3).

Sağlam et al. (2006), por sua vez, descrevem e apontam os resultados de seu modelo de PLIM para o *K-means* quando aplicado a dados de uma companhia privada de telecomunica-

ções na Turquia, versão que foi adotada neste trabalho. Um desafio enfrentado pelos autores diz respeito à não linearidade de seu modelo e ao alto custo computacional envolvido em sua solução. Como resposta, a técnica de linearização proposta por Sağlam et al. (2006) melhorou drasticamente o tempo computacional de seu algoritmo. Por fim, Werner (2022) apresenta uma formulação do *K-means* com um modelo de PLIM com restrições quadráticas sujeitas a um processo de linearização proposto por Duran e Grossman (1986, *apud* Werner, 2022). Sobre os modelos de Awasthi et al. (2015), Werner (2022) e Sağlam et al. (2006) se discorrerá com mais detalhes na seção seguinte.

3 MODELOS

Uma explanação básica sobre os modelos de Awasthi et al. (2015), Werner (2022) e Sağlam et al. (2006) é apresentada nas subseções a seguir. Contudo, encontram-se sumarizadas no Quadro 1 as principais características de cada modelo, como a técnica de Programação Matemática associada, seu objetivo e seus dados de entrada.

Quadro 1 – Sumarização dos métodos implementados

Atributo	**Awasthi et al. (2015)**	**Werner (2022)**	**Sağlam et al. (2006)**
Técnica	PIB	PLIM	PLIM
Objetivo	Minimizar o somatório da distância de cada ponto a seu ponto representativo	Minimizar o somatório do quadrado da distância de cada ponto a seu centroide	Minimizar o máximo diâmetro dos *clusters*
Entrada	Número de *clusters*, tamanho da amostra e coordenadas dos pontos	Número de *clusters*, tamanho da amostra e coordenadas dos pontos	Número de *clusters*, tamanho da amostra e coordenadas dos pontos
Variáveis	Indicam quais são os pontos representativos e por qual deles cada ponto é representado	Indicam as coordenadas de cada centroide, se um ponto pertence a um *cluster* e a distância de um ponto ao centroide	Indicam o diâmetro de cada cluster, o diâmetro máximo e se um ponto pertence a um cluster

Fonte: Elaboração dos autores.

3.1 FORMULAÇÃO DE AWASTHI ET AL. (2015) PARA O *K-MEDIANS*

Neste modelo, desejamos encontrar k pontos representativos dentre o conjunto P de modo que uma partição de P em k subconjuntos (disjuntos e não vazios) seja tal que a soma das

distâncias de cada ponto ao ponto representativo do *cluster* ao qual ele foi alocado seja a menor possível. Note que $d(p, q)$ representa a distância de Manhattan (como Awasthi et al. (2015) não fixam uma métrica para análise, neste momento adota-se a distância de Manhattan, mas a euclidiana é utilizada em seções futuras deste trabalho em testes variando a função objetivo do modelo) entre os pontos p e q, y_p é uma variável de decisão binária que denota se o ponto p é um ponto representativo (se é, $y_p = 1$; caso contrário, $y_p = 0$) e $z_{p,q}$ indica se p representa q (se representa, $z_{p,q=1}$; caso contrário, $z_{p,q=0}$). Vale ressaltar que a solução deste problema induz à matriz de adjacências de um grafo composto por k estrelas disjuntas cujos nós internos correspondem aos pontos representativos (Awasthi et al., 2015). Considere o modelo a seguir:

$$\text{Minimizar} \sum_{p,q \in P} d(p, q) \cdot z_{p,q}$$

$$\text{Sujeito a} \begin{cases} \sum_{p \in P} z_{p,q} = 1 & \forall q \in P \\ z_{p,q} \leq y_p & \forall p, q \in P \\ \sum_{p \in P} y_p = k & \forall p \in P \\ z_{p,q}, y_p \in \{0, 1\} & \forall p, q \in P \end{cases}$$

3.2 FORMULAÇÃO DE WERNER (2022) PARA O *K-MEANS*

Neste modelo, considere um conjunto P de n pontos $p_1, \ldots, p_n$. Queremos determinar k centroides c_j, $j \in \{1, \ldots, k\}$, dentre o conjunto P de modo que a alocação dos pontos de P aos k *clusters* seja tal que a soma do quadrado da distância euclidiana de cada ponto ao centroide de seu *cluster* seja minimizada. Note que, para um ponto i, $i \in \{1, \ldots, n\}$, o quadrado de sua distância a qualquer centroide é majorado por r_i e M_i denota uma constante suficientemente grande (*Big M constant*) para assegurar a igualdade de r_i ao quadrado da distância do ponto i ao centroide do seu cluster. Para cada $j \in \{1, \ldots, k\}$, c_j representa o j-ésimo centroide e, para todo $i \in \{1, \ldots, n\}$, define-se uma variável binária b_j^i que indica se o ponto i pertence ao *cluster* de centroide c_j (se pertence, $b_j^i = 1$; caso contrário, $b_j^i = 0$). Desse modo, considere o modelo a seguir (Werner, 2022):

$$\text{Minimizar} \sum_{i=1}^{n} r_i$$

$$\text{Sujeito a} \begin{cases} \|p_i - c_j\|_2^2 \leq r_i + M_i \cdot (1 - b_j^i) & \forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, k\} \\ \sum_{j=1}^{k} b_j^i = 1 & \forall i \in \{1, \ldots, n\} \\ b_j^i \in \{0, 1\} & \forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, k\} \\ r_i \geq 0 & \forall i \in \{1, \ldots, n\} \end{cases}$$

Devido ao cálculo do quadrado do módulo do vetor diferença na primeira restrição, o modelo perde sua linearidade pois, como se vê, dados dois vetores $x, y \in \mathbb{R}^n$ vale que:

$$\begin{aligned} \|x - y\|_2^2 &= \|(x_1, x_2, \ldots, x_n) - (y_1, y_2, \ldots, y_n)\|_2^2 \\ &= \|(x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n)\|_2^2 \\ &= \left[\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}\right]^2 \\ &= (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2 \\ &= (x_1^2 - 2x_1y_1 + y_1^2) + (x_2^2 - 2x_2y_2 + y_2^2) + \ldots + (x_n^2 - 2x_ny_n + y_n^2) \end{aligned}$$

implicando a não linearidade do modelo e a natureza quadrática de sua restrição. A descrição do processo de linearização dessa restrição foge ao escopo deste trabalho, mas se encontra em Duran e Grossman (1986, *apud* Werner, 2022).

3.3 FORMULAÇÃO DE SAĞLAM ET AL. (2006) PARA O *K-MEANS*

Neste caso, considere um conjunto P de n pontos. Queremos encontrar uma partição de P em k *clusters* de modo a minimizar o diâmetro máximo dos *clusters*. Note que não nos preocupamos com a identidade do centroide de cada cluster, mas com o conjunto de pontos que compõe cada *cluster*. Considere que d_{ij} representa a distância euclidiana entre dois pontos i e j, D_l indica o diâmetro do *cluster* l e majora a distância entre quaisquer dois pontos i e j pertencentes ao *cluster* l, D_{max} é o maior dos diâmetros e x_{il} denota se o ponto i pertence ao *cluster* l (se pertence, $x_{il} = 1$; caso contrário, $x_{il} = 0$). Isso posto, considere o modelo a seguir (Sağlam et al., 2006):

$$\text{Minimizar } D_{max}$$

$$\text{Sujeito a} \begin{cases} D_l \geq d_{ij} \cdot x_{il} \cdot x_{jl} & \forall i, j \in \{1, \ldots, n\}, \forall l \in \{1, \ldots, k\} \\ \sum_{j=1}^{k} x_{il} = 1 & \forall i \in \{1, \ldots, n\} \\ D_{max} \geq D_l & \forall l \in \{1, \ldots, k\} \\ x_{il} \in \{0, 1\} & \forall i \in \{1, \ldots, n\}, \forall l \in \{1, \ldots, k\} \\ D_l \geq 0 & \forall l \in \{1, \ldots, k\} \end{cases}$$

Por conta do produto de variáveis binárias na primeira restrição discriminada, tal modelo

perde sua linearidade. Adicionalmente, tal modelagem é computacionalmente ineficiente e resulta em modelos insolúveis em tempo razoável, inclusive para instâncias com poucos dados. Contudo, segundo Bisschop (2006), é possível garantir a completa linearidade do modelo reescrevendo essa restrição de um modo conveniente. E, conforme Sağlam et al. (2006), ainda há vantagens em termos de desempenho computacional. Sejam x_1 e x_2 duas variáveis binárias e considere que seu produto $x_1 \cdot x_2$ compõe alguma restrição ou a função objetivo do modelo, tornando-o não linear. O método de linearização de Bisschop (2006) consiste em criar uma variável y sujeita a:

$$y \in \{0,1\} \text{ tal que } \begin{cases} y & \leq x_1 \\ y & \leq x_2 \\ y & \geq x_1 + x_2 - 1 \end{cases}$$

e substituir o produto $x_1 \cdot x_2$ por y em toda restrição do modelo e/ou em sua função objetivo. Isso posto, os próprios autores, Sağlam et al. (2006), apresentam uma alternativa a essa restrição problemática, garantindo a linearidade do modelo e levando a uma modelagem do *K-means* estritamente com PLIM. Essa restrição é tal que:

$$D_l \geq d_{ij} \cdot (x_{il} + x_{jl} - 1) \forall i, j \in \{1, \ldots, n\}, \forall l \in \{1, \ldots, k\}$$

garantindo que ela se ative apenas se os pontos em análise pertencem ao mesmo *cluster*.

4 IMPLEMENTAÇÃO

A implementação da solução dos problemas de PLIM associados às modelagens descritas foi feita na linguagem de programação *Python*. Entre outras, foi utilizada a biblioteca *gurobipy*, vinculada ao solver *Gurobi Optimizer*, o qual demandou a aquisição de uma licença acadêmica gratuita. Explorando a base de dados abertos do INEP (Brasil, 2024), na categoria Enem por Escola, tem-se acesso aos microdados no formato CSV referentes às edições de 2005 a 2015 da modalidade, e dados da edição de 2015, que incluem todas as variáveis relevantes (listadas adiante), os quais foram utilizados. Um conjunto de 27.152 escolas compõe a base de dados, e cada uma delas apresenta dados de 1 a 11 edições do Enem, totalizando 172.305 registros.

A entrada de dados para o algoritmo de solução é feita por meio de tais arquivos CSV – sistema de registro utilizado pelo INEP. Há nos arquivos colunas referentes à nota média das escolas nas seguintes áreas de conhecimento do Enem: Ciências da Natureza e suas Tecnologias, Ciências Humanas e suas Tecnologias, Linguagens, Códigos e suas Tecnologias, Matemática e suas Tecnologias e Redação. A última coluna, por sua vez, se refere ao Indicador de Nível Socioeconômico (INSE) das instituições (alto ou baixo) ou à sua dependência administrativa (municipal, estadual, federal ou privada). Em particular, as cinco primeiras variáveis

são utilizadas pelos algoritmos de clusterização, que associam a cada escola um vetor de cinco dimensões, uma para cada variável. Os dois últimos atributos, por sua vez, foram selecionados para fundamentar o contraponto dos resultados da clusterização para $k = 2$ com a realidade socioeconômica das escolas.

Uma etapa de pré-processamento foi necessária para tratar irregularidades na base de dados. Como mencionado, edições do Enem anteriores à de 2015 não foram consideradas por omitirem os valores de variáveis pertinentes, a saber, o INSE das instituições. Com isso, o número inicial de registros, 172.305, foi reduzido para 15.597. Em seguida, descartando toda entrada em que se omitia a variável referente à dependência administrativa da escola ou as variáveis referentes às notas médias por área de conhecimento no Enem, a base de dados foi reduzida a registros de 2015 de 15.497 escolas. Em razão de as cinco variáveis supracitadas estarem na mesma escala (pois são notas nas competências do Enem), não foi aplicada uma técnica de normalização dos dados.

Sobre bibliotecas do *Python* utilizadas, algumas foram importadas para auxiliar em operações com vetores e matrizes (*numpy*), na leitura de arquivos (*pandas*), na solução dos problemas de otimização (*gurobipy*), na plotagem de gráficos (*matplotlib*), nas operações com grafos (*networkx*) e nos algoritmos de clusterização (*sklearn*).

De posse dessas informações, inicialmente, o solver *Gurobi Optimizer* é utilizado para solucionar o modelo de Programação Matemática desenvolvido, construindo *clusters* válidos, não vazios e disjuntos. Os experimentos computacionais realizados se encerram com a completa clusterização do conjunto de dados ou com a exaustão dos recursos computacionais disponíveis ao ponto da inviabilidade. Para cada um dos *clusters* encontrados, são analisadas as características das escolas alocadas em termos de suas notas médias nas áreas do conhecimento do Enem. Em particular, uma planilha de dados associando cada escola ao *cluster* em que ela foi alocada é construída ao final.

5 RESULTADOS

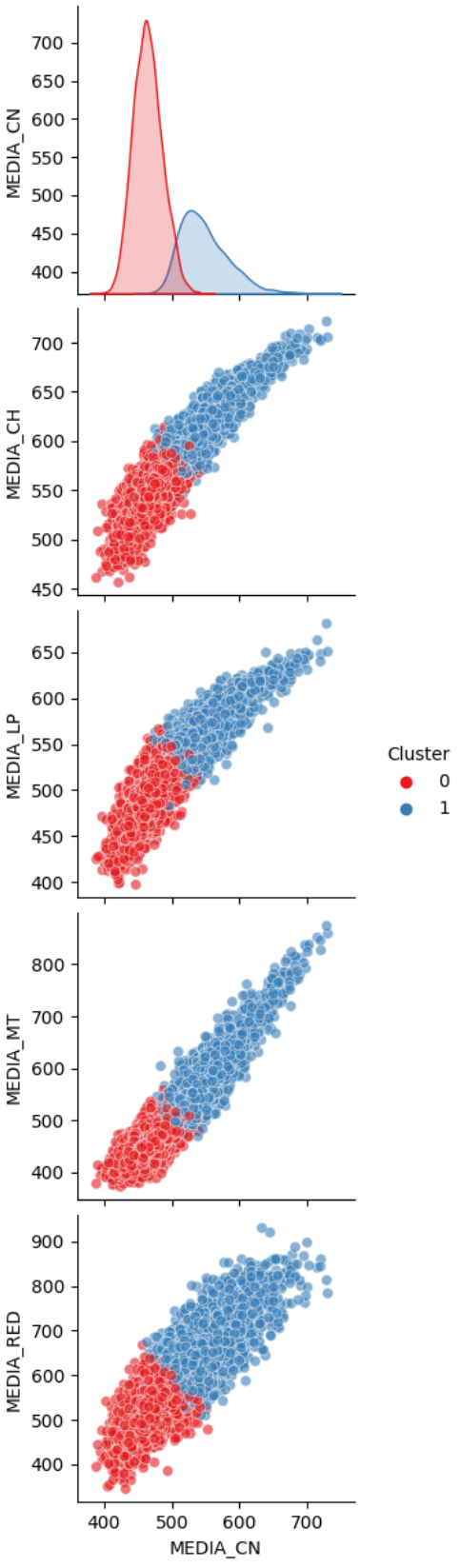
Nesta seção, serão apresentados os resultados dos algoritmos de clusterização baseados nos modelos de Awasthi et al. (2015), Werner (2022) e Sağlam et al. (2006). A base de dados adotada consiste nas planilhas divulgadas pelo INEP a respeito do Enem por Escola. Para fins de comparação, os dados serão clusterizados com algoritmos próprios da biblioteca *sklearn* do *Python*, a saber, *K-means*, *Weighted K-means* e *K-medians*, e os resultados encontrados serão contrapostos àqueles dos algoritmos implementados neste trabalho. Foram realizados testes computacionais com 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 e 20 *clusters*, bem como modificações nas funções objetivo e das restrições dos modelos desenvolvidos, a fim de se verificar nuances em seu desempenho.

Nas Figuras 2–6, representam-se os resultados do algoritmo de clusterização *K-means* da

biblioteca *sklearn* do *Python* quando aplicado aos dados do Enem por Escola definindo-se dois *clusters*. Cada figura, apresenta cinco gráficos exibindo a relação das variáveis duas a duas: nota média em Ciências da Natureza e suas Tecnologias (MEDIA_CN), Ciências Humanas e suas Tecnologias (MEDIA_CH), Linguagens, Códigos e suas Tecnologias (MEDIA_LP), Matemática e suas Tecnologias (MEDIA_MT) e Redação (MEDIA_RED). Adicionalmente, é possível observar os histogramas das distribuições de cada variável. Note que, em cada caso, as instituições com desempenho inferior ocupam o *cluster* 0. A Figura 7, por sua vez, apresenta um gráfico de setores com os resultados do algoritmo *K-medians*, da biblioteca *sklearn*, quando aplicado ao mesmo conjunto de dados com os mesmos parâmetros.

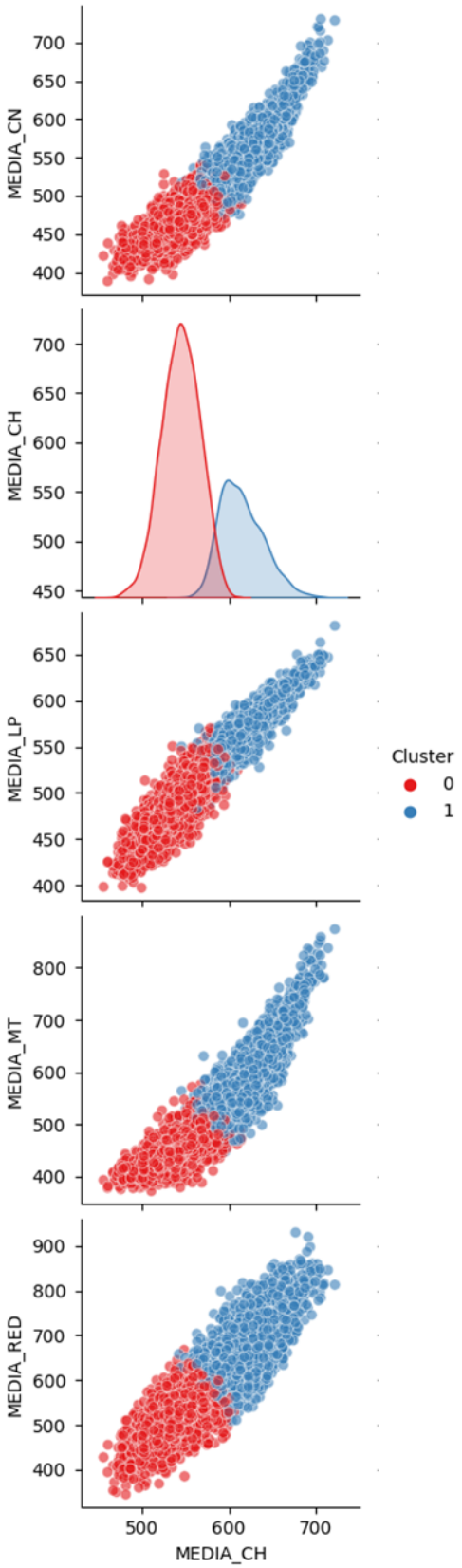
Nesse caso, foram comparados os *clusters* atribuídos a cada instituição (dentre duas opções) com sua dependência administrativa (1 em escolas municipais e estaduais e 2 para escolas federais e privadas). A Figura 8, por sua vez, exibe a matriz de confusão resultante do uso do algoritmo *Weighted K-means*, da biblioteca *sklearn*, no mesmo conjunto de dados, tendo como referência a mesma variável.

Figura 2 – Resultados do *K-means* dois a dois com a MEDIA_CN



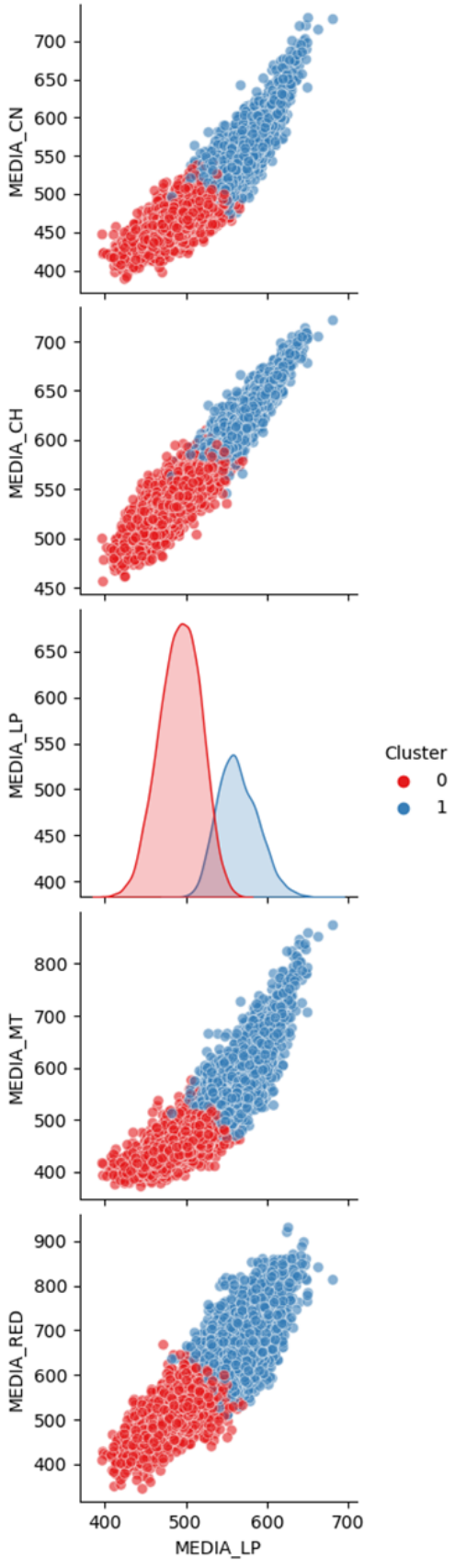
Fonte: Elaboração dos autores.

Figura 3 – Resultados do *K-means* dois a dois com a MEDIA_CH



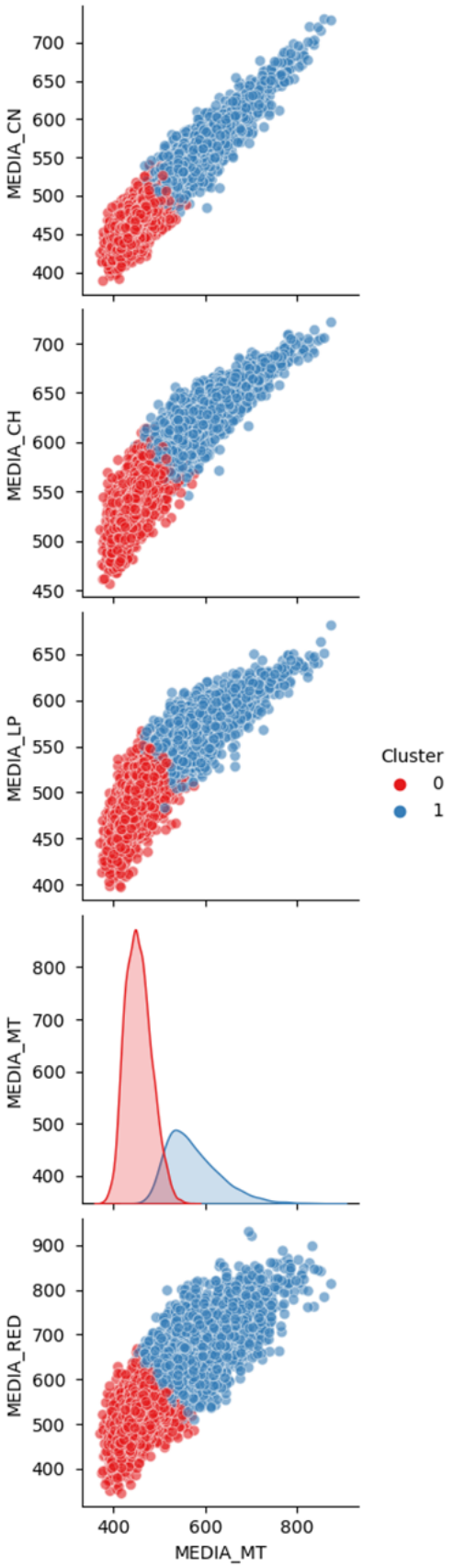
Fonte: Elaboração dos autores.

Figura 4 – Resultados do *K-means* dois a dois com a MEDIA_LP



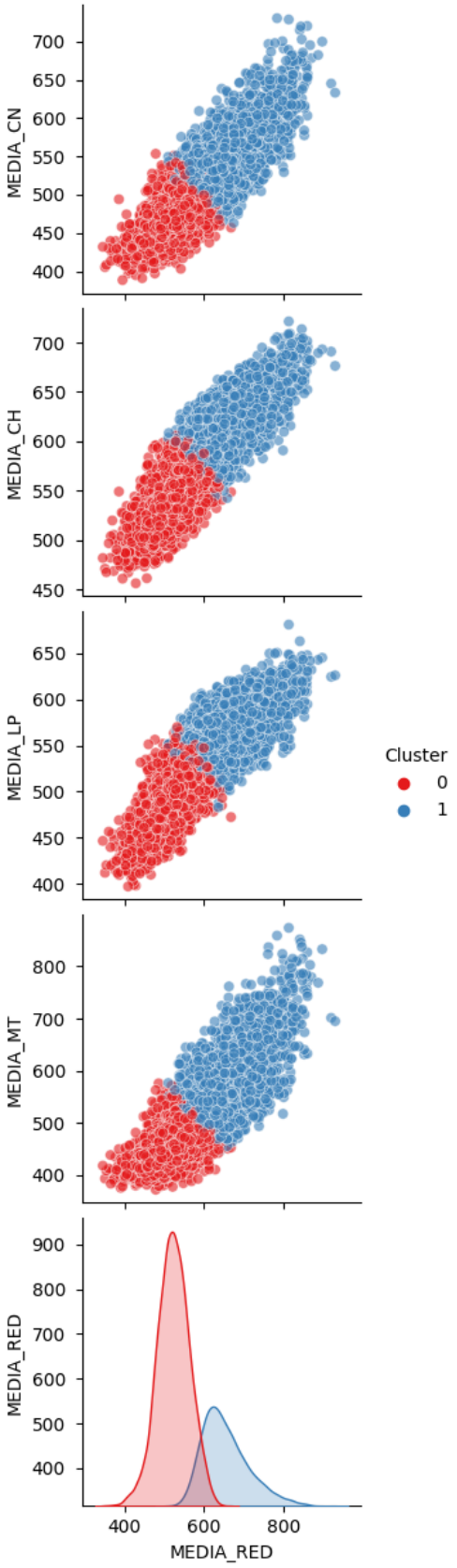
Fonte: Elaboração dos autores.

Figura 5 – Resultados do *K-means* dois a dois com a MEDIA_MT



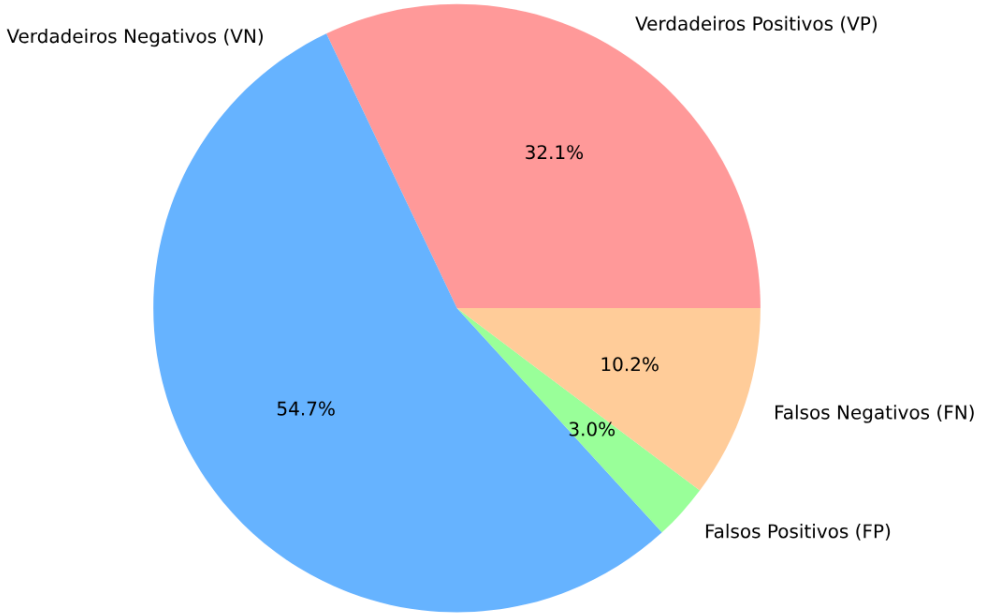
Fonte: Elaboração dos autores.

Figura 6 – Resultados do *K-means* dois a dois com a MEDIA_RED



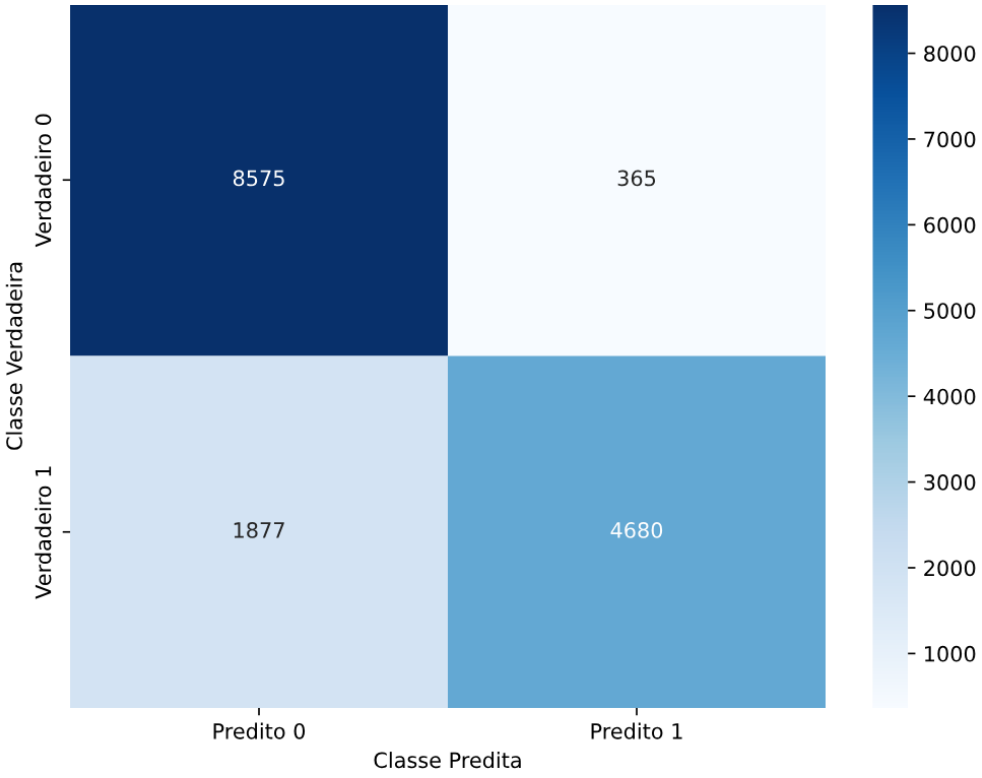
Fonte: Elaboração dos autores.

Figura 7 – Relação entre os *clusters* do *K-medians* e a dependência administrativa da instituição



Fonte: Elaboração dos autores.

Figura 8 – Relação entre os *clusters* do *Weighted K-medians* e a dependência administrativa da instituição



Fonte: Elaboração dos autores.

De modo geral, pode-se avaliar o desempenho do *Weighted K-means* com métricas como acurácia (86%), precisão (82%), sensibilidade (96%) e F1-score (88%).

5.1 RESULTADOS DO ALGORITMO *K-MEDIANS* DE AWASTHI ET AL. (2015)

Quando testado em uma instância com 1.000 instituições de ensino e $k = 2$, foi possível comparar os *clusters* aos quais cada escola foi alocada (de acordo com seu desempenho médio no Enem) com sua dependência administrativa [1 (para escolas públicas municipais e estaduais) ou 2 (para escolas públicas federais e escolas privadas)] e seu INSE [1 (baixo nível socioeconômico) ou 2 (alto nível)]. No primeiro caso, tem-se: acurácia de 88%, precisão de 95%, sensibilidade de 86% e F1-score de 90%. No segundo caso, analisando o INSE, encontra-se 85% de acurácia, 85% de precisão, 93% de sensibilidade e 89% de F1-score. Considere que a distância de Manhattan calculada entre os pontos na função objetivo deste modelo seja substituída pela distância euclidiana. Note que tal modificação na função objetivo preserva a linearidade da formulação. Assumindo essa nova abordagem, pode-se comparar o desempenho dessa variação do algoritmo, para o caso do INSE, segundo as mesmas métricas: 85% para acurácia, 85% para precisão, 93% para sensibilidade e 89% para F1-score, e, para o caso da dependência administrativa: 88% de acurácia, 95% de precisão, 87% de sensibilidade e 91% de F1-score.

5.2 RESULTADOS DO ALGORITMO *K-MEANS* DE SAĞLAM ET AL. (2006)

Seu custo computacional tornou proibitiva uma amostra com o mesmo tamanho da analisada no modelo anterior. Entretanto, dentre o mesmo conjunto de dados, para uma instância com 500 instituições de ensino e $k = 2$, foi possível comparar os *clusters* aos quais cada escola foi alocada, segundo os mesmo critérios, com sua dependência administrativa e seu INSE. No primeiro caso, os valores encontrados são 52% para acurácia, 75% para precisão, 68% para sensibilidade e 65% para F1-score. No segundo caso, tem-se acurácia de 59%, precisão de 78%, sensibilidade de 67% e F1-score de 72%. Considere uma variação do modelo de Sağlam et al. (2006) de modo a modificar suas restrições e calcular a distância de Manhattan entre os pontos em vez de a distância euclidiana. Seguem os resultados. No caso da dependência administrativa, vale 52% para acurácia, 75% para precisão, 58% para sensibilidade e 65% para F1-score. No caso do INSE, tem-se acurácia de 43%, precisão de 39%, sensibilidade de 63% e F1-score de 48%.

5.3 RESULTADOS DO ALGORITMO *K-MEANS* DE WERNER (2022)

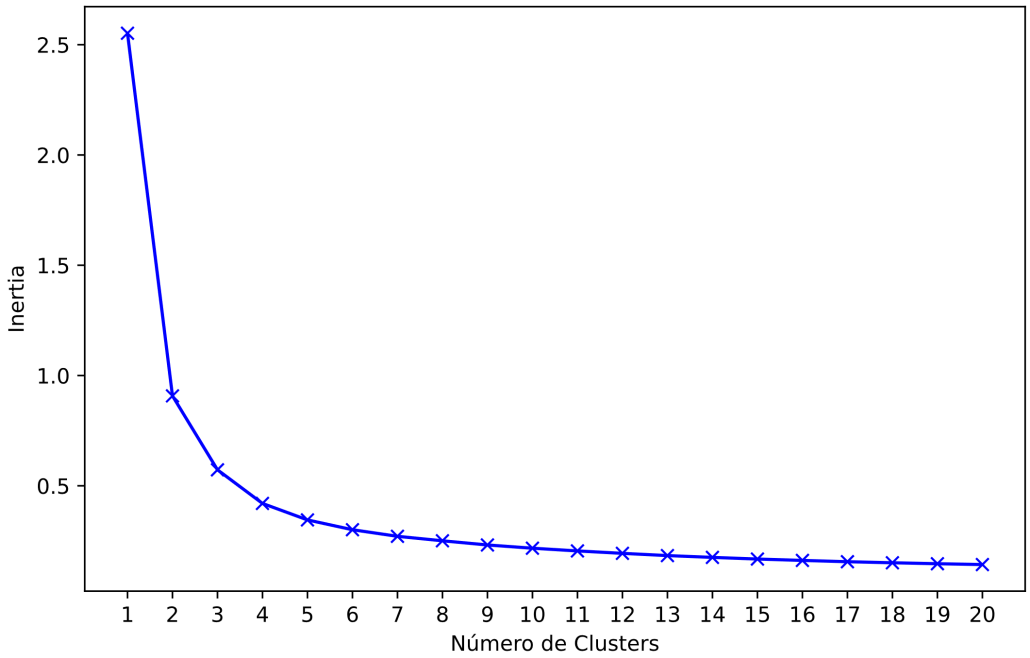
Conforme apontado adiante, a complexidade computacional deste algoritmo não permite sua testagem para amostras de dados de tamanho compatível com os modelos anteriores. Por conta disso, testando uma amostra de 200 escolas e fazendo $k = 2$, também foram comparadas as

alocações em *clusters* com a dependência administrativa e o INSE das escolas. Segundo as métricas de avaliação calculadas, no primeiro caso, a acurácia vale 51%, a precisão vale 71%, a sensibilidade vale 58% e o F1-score vale 64%. No segundo caso, tem-se acurácia de 58%, precisão de 73%, sensibilidade de 68% e F1-score de 70%

5.4 RESULTADOS VARIANDO O NÚMERO DE *CLUSTERS*

Devido ao desempenho computacional superior do algoritmo *K-medians* de Awasthi et al. (2015), ele será a base para comparação com o *K-means* da biblioteca *sklearn* do *Python* nos sucessivos testes para distintos valores de $k = 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15$ e 20, uma vez que $k = 2$ foi exaustivamente testado nas seções anteriores. A Figura 9 apresenta o gráfico da aplicação do Método do Cotovelo na base de dados. Com base nos resultados encontrados, espera-se que o número ideal de *clusters* seja 2 ou 3. A partir do 4, com a diminuição da inclinação visualmente perceptível na Figura 9, a adição de mais *clusters* não proporciona uma diminuição significativa na soma das distâncias dentro dos *clusters* (a chamada *inertia*).

Figura 9 – Aplicação do Método do Cotovelo aos dados do Enem por Escola do INEP



Fonte: Elaboração dos autores.

Para verificar empiricamente esse processo, serão calculadas, para cada *cluster* no qual forem alocadas escolas, as médias das notas das instituições nas áreas de conhecimento do Enem, à medida que o valor de k cresce (Tabelas 1–14).

Tabela 1 – Médias para $k = 3$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	473,66	485,93
2	520,14	546,70
3	587,05	621,07

Fonte: Elaboração dos autores.

Tabela 2 – Médias para $k = 4$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	469,34	474,85
2	497,11	514,18
3	539,60	569,58
4	604,07	636,32

Fonte: Elaboração dos autores.

Tabela 3 – Médias para $k = 5$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	462,94	470,94
2	485,05	505,38
3	521,05	548,42
4	563,46	594,44
5	641,03	652,94

Fonte: Elaboração dos autores.

Tabela 4 – Médias para $k = 6$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	462,15	464,82
2	481,37	494,55
3	503,33	525,60
4	532,12	564,49
5	567,53	607,11
6	641,03	662,89

Fonte: Elaboração dos autores.

Tabela 5 – Médias para $k = 7$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	457,01	460,42
2	477,37	487,34
3	480,97	513,27
4	503,22	545,96
5	532,12	580,58
6	567,53	620,33
7	641,02	673,17

Fonte: Elaboração dos autores.

Tabela 6 – Médias para $k = 8$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	457,01	454,53
2	476,84	479,42
3	480,43	501,65
4	498,13	526,75
5	524,83	556,62
6	550,65	587,66
7	582,13	624,89
8	644,57	676,46

Fonte: Elaboração dos autores.

Tabela 7 – Médias para $k = 9$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	457,01	454,00
2	476,84	478,78
3	480,43	500,67
4	497,23	525,00
5	518,86	550,67
6	537,75	569,69
7	561,88	591,50
8	589,40	627,65
9	648,27	679,50

Fonte: Elaboração dos autores.

Tabela 8 – Médias para $k = 10$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	447,39	454,33
2	465,74	478,01
3	479,45	498,82
4	480,21	515,18
5	497,79	535,32
6	518,86	553,38
7	537,74	572,56
8	561,88	594,78
9	589,40	630,43
10	648,27	681,68

Fonte: Elaboração dos autores.

Tabela 9 – Médias para $k = 11$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	447,39	453,40
2	465,74	476,88
3	479,45	497,34
4	480,21	513,76
5	497,79	531,53
6	518,86	546,65
7	537,75	568,54
8	561,88	584,43
9	585,01	605,12
10	627,32	634,98
11	668,44	684,07

Fonte: Elaboração dos autores.

Tabela 10 – Médias para $k = 12$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	447,50	453,74
2	465,94	475,24
3	469,61	487,83
4	480,36	501,65
5	485,22	521,27
6	499,33	531,53
7	519,59	550,30
8	537,75	571,61
9	561,88	586,00
10	585,01	608,09
11	627,32	636,29
12	668,44	684,85

Fonte: Elaboração dos autores.

Tabela 11 – Médias para $k = 13$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	447,5	453,51
2	465,94	475,29
3	469,61	484,1
4	480,36	499,26
5	485,22	513,47
6	499,33	526,85
7	518,54	543,13
8	528,64	558,04
9	541,56	581,96
10	561,72	585,38
11	585,01	614,46
12	627,32	645,52
13	668,44	694,43

Fonte: Elaboração dos autores.

Tabela 12 – Médias para $k = 14$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	444,80	453,45
2	459,89	475,10
3	469,54	483,87
4	473,61	498,69
5	484,59	512,51
6	486,69	525,35
7	501,85	541,85
8	518,91	555,91
9	528,64	578,60
10	541,56	583,59
11	561,72	615,52
12	585,01	615,90
13	627,32	653,41
14	668,44	703,75

Fonte: Elaboração dos autores.

Tabela 13 – Médias para $k = 15$

Cluster	**Awasthi et al. (2015)**	***Sklearn***
1	444,80	451,47
2	459,89	471,10
3	469,54	485,34
4	473,61	496,40
5	483,95	503,74
6	485,94	520,87
7	503,75	531,47
8	504,53	547,75
9	523,41	559,58
10	537,23	579,09
11	542,44	588,96
12	563,44	614,74
13	585,01	624,49
14	627,32	655,22
15	668,44	703,94

Fonte: Elaboração dos autores.

Tabela 14 – Médias para $k = 20$

Cluster	Awasthi et al. (2015)	*Sklearn*
1	444,8	444,61
2	459,89	463,05
3	468,79	480,03
4	472,37	480,87
5	479,11	496,99
6	482,11	498,93
7	486,05	512,00
8	496,74	526,59
9	500,37	526,95
10	517,33	548,34
11	522,35	548,63
12	530,92	567,96
13	537,3	576,16
14	547,49	581,53
15	549,97	593,37
16	565,32	615,23
17	585,77	621,07
18	615,46	645,1
19	637,57	666,92
20	673,05	713,52

Fonte: Elaboração dos autores.

6 DISCUSSÃO

Como apontado por Sağlam et al. (2006), ainda que seu modelo linear seja mais eficiente do ponto de vista computacional que sua versão não linear, seus resultados apresentam uma qualidade inferior à dos demais modelos encontrados na literatura. Esse comportamento esperado se verificou nos testes computacionais realizados neste trabalho. O modelo de Werner (2022), por sua vez, apresenta complexidade exponencial no número de pontos analisados e, por conta disso, é computacionalmente ineficiente para grandes amostras. O segundo fator de maior impacto na performance do algoritmo é a quantidade pré-definida de *clusters*. Por fim, a autora conclui que a dimensionalidade dos pontos analisados é o parâmetro que menos influencia a complexidade.

Por outro lado, o algoritmo *K-medians*, baseado no modelo de Awasthi et al. (2015), com base nos testes feitos neste trabalho, se mostrou o mais eficiente tanto no tempo computacional

quanto na qualidade das classificações. Comparando-se as métricas de avaliação, tem-se que o algoritmo de Awasthi et al. (2015) e o *Weighted K-means* da biblioteca *sklearn* apresentaram valores equiparáveis de acurácia (88% e 86%), precisão (95% e 82%), sensibilidade (86% e 96%) e F1-score (90% e 88%). Se forem analisadas as métricas para o *K-medians* da *sklearn*, as observações são semelhantes: acurácia de 87%, precisão de 91%, sensibilidade de 76% e F1-score de 83%.

Em relação às variações construídas para os modelos, nota-se que a modificação nas restrições do algoritmo de Sağlam et al. (2006) levou a resultados piores em cada teste realizado nas bases de dados com as dependências administrativas e com o INSE. Por outro lado, a modificação da função objetivo do modelo de Awasthi et al. (2015) levou a métricas com resultados equiparáveis, senão ligeiramente superiores. Sobre as variações no número de *clusters*, como previsto pelo Método do Cotovelo e atestado pelos resultados, o aumento arbitrário de k não leva a uma divisão mais coerente das escolas. Como é possível observar a cada incremento no valor de k, passam a surgir conjuntos *clusters* cuja heterogeneidade é posta em xeque, pois a média dos valores de cada um são virtualmente iguais. Um valor para $k \in \{2, 3\}$ é possivelmente o ótimo, conforme a Figura 8.

De modo geral, os resultados dos algoritmos mais consistentes apontam para dados da realidade socioeconômica e educacional brasileira. Em cada caso da alocação das escolas entre dois *clusters*, o critério das notas médias nas áreas de conhecimento do Enem era o principal fator para a homogeneidade intracluster e a heterogeneidade extracluster. A diferença entre notas médias das escolas de cada *cluster* chega a 80 pontos. Dito isso, quando contrapostos os *clusters* aos quais as instituições foram alocadas ao seu perfil socioeconômico (aqui representado por seu INSE e por sua dependência administrativa), os resultados indicam que escolas públicas municipais e estaduais tendem a pertencem ao *cluster* de pior desempenho, assim como escolas de nível socioeconômico baixo, enquanto que escolas públicas federais e escolas privadas tendem a pertencer ao *cluster* de melhor desempenho, tal como as escolas de nível socioeconômico alto.

7 CONSIDERAÇÕES FINAIS

Neste trabalho, foram apresentados três modelos de clusterização baseados em Programação Inteira Binária (PIB) e Programação Linear Inteira Mista (PLIM) aplicados aos dados do Enem por Escola, fornecidos pelo INEP. A análise dos resultados permitiu identificar padrões de desempenho das instituições de ensino e suas correlações com características socioeconômicas, evidenciando disparidades significativas entre diferentes tipos de escolas. Apesar de terem a vantagem de incorporar restrições e objetivos mais complexos por meio da Programação Matemática, a maioria dos modelos propostos não se mostrou tão eficaz quanto as técnicas de clusterização tradicionais, como o *K-means* e o *K-medians*, tanto em questão de qualidade

quanto de performance computacional. No entanto, destaca-se o desempenho do algoritmo desenvolvido a partir de Awasthi et al. (2015), cujos resultados se equiparam aos de abordagens convencionais de clusterização.

A utilização do solver *Gurobi Optimizer*, com sua capacidade de lidar com problemas de otimização complexos, foi crucial para a viabilidade computacional dos modelos propostos e a implementação prática dos algoritmos para validação dos modelos. Nesse sentido, a validação por meio da comparação com a realidade socioeconômica das escolas e a análise crítica dos resultados obtidos reforçam a importância de abordagens baseadas em Programação Matemática para a análise de dados educacionais, contribuindo para a formulação de políticas públicas mais informadas e eficazes na promoção da equidade e qualidade da educação no Brasil.

Como direções futuras, sugere-se a exploração de outras técnicas de Programação Matemática, bem como a integração de métodos híbridos que combinem a robustez dos modelos matemáticos com a flexibilidade dos algoritmos de aprendizado de máquina. Nesse sentido, o trabalho de Nueda, Gandía e Molina (2022) apresenta um novo método de *Linear Programming Discriminant Analysis*, baseado na construção de hiperplanos separadores, para classificação de dados, e pode ser uma ferramenta útil. Além disso, a ampliação do conjunto de dados para incluir edições mais recentes do Enem e a aplicação dos modelos a outras bases de dados educacionais podem proporcionar *insights* ainda mais abrangentes e atualizados sobre o cenário educacional brasileiro.

REFERÊNCIAS

ÁGOSTON, Kolos Cs.; E.-NAGY, Marianna. Mixed integer linear programming formulation for K-means clustering problem. **Central European Journal of Operations Research**, v. 32, n. 1, p. 11–27, 2023. DOI: https://doi.org/10.1007/s10100-023-00881-1.

AWASTHI, Pranjal; BANDEIRA, Afonso S.; CHARIKAR, Moses; KRISHNASWAMY, Ravishankar; VILLAR, Soledad; WARD, Rachel. Relax, no need to round: integrality of clustering formulations. In: PROCEEDINGS OF THE 2015 CONFERENCE ON INNOVATIONS IN THEORETICAL COMPUTER SCIENCE. New York, NY, USA: Association for Computing Machinery, 2015. p. 191-200. DOI: https://doi.org/10.1145/2688073.268811.

BISSCHOP, Johannes. **AIMMS**: Optimization Modelling. [S.l.]: AIMMS B.V, 2006.

BRASIL. **Microdados**. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2024. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados. Acesso em: 12 maio 2024.

FONSECA, Stella; NAMEN, Anderson. Mineração em bases de dados do INEP: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. **Educação em Revista**, v. 32, p. 133–157, jan. 2016. DOI: https://doi.org/10.1590/0102-4698140742.

FRANCELINO, Wander; MACHADO, Lucas. **Mineração de dados nos microdados Enade computação**. Tubarão: Repositório Universitário de Ânima, 2020. Disponível em: https://repositorio.animaeducacao.com.br/handle/ANIMA/8471. Acesso em: 12 maio 2024.

MAIA, Marília Magalhães; ANDRADE, Luiza Helena Felix de; FERNANDES, Silvio. K-means na análise de características socioeconômicas de candidatos ao ensino superior. In: ENCONTRO DE COMPUTAÇÃO DO OESTE POTIGUAR, 5., 2021, Pau dos Ferros, RN. **Anais** [...]. Pau dos Ferros, RN: UFERSA, 2021. Disponível em: https://periodicos.ufersa.edu.br/ecop/article/view/11168/10877. Acesso em: 13 maio 2024.

MASCHIO, Pedro de Torres; VIEIRA, Marcos Alves; COSTA, Newarney Torrezão da; MELO, Sara Luzia de; PEREIRA JUNIOR, Cleon Xavier. Um panorama acerca da mineração de dados educacionais no Brasil. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 29., 2018, Fortaleza. **Anais** [...]. [S.l.]: SBC, 2018. p. 1936–1940. Disponível em: http://milanesa.ime.usp.br/rbie/index.php/sbie/article/viewFile/8194/5873. Acesso em: 2 maio 2025.

NUEDA, María; GANDÍA, Carmen; MOLINA, Mariola. LPDA: A new classification method based on linear programming. **PLoS ONE**, v. 17, n. 7, jul. 2022. DOI: https://doi.org/10.1371/journal.pone.0270403.

SAĞLAM, Burcu; SALMAN, Sibel; SAYIN, Serpil; TÜRKAY, Metin. A mixed-integer programming approach to the slustering problem with an application in customer segmentation. **European Journal of Operational Research**, v. 173, n. 3, p. 866–879, set. 2006. DOI: https://doi.org/10.1016/j.ejor.2005.04.048.

WERNER, Hanna. **K-means Clustering as a Mixed Integer Programming Problem**. 2022. Degree Project (Technology) – Stockholm, Sweden. Disponível em: https://www.diva-portal.org/smash/get/diva2:1673547/FULLTEXT01.pdf. Acesso em: 2 maio 2025.

SOBRE OS AUTORES

Me. Victor Augusto do Carmo Duarte
https://orcid.org/0009-0005-6807-7500
http://lattes.cnpq.br/1883519126889519
Contato: vduarte@posgrad.lncc.br
Contribuição autoral: administração do projeto; curadoria de dados; escrita – primeira redação; escrita – revisão e edição; software; validação; visualização.

Dr. Erito Marques de Souza Filho
https://orcid.org/0000-0002-0381-3344
http://lattes.cnpq.br/0606341154404244
Contato: eritomarques@id.uff.br
Contribuição autoral: conceituação; escrita – revisão e edição; metodologia; supervisão.