Aplicando Mineração de Dados Educacionais para a Redistribuição dos Distritos de Educação de Fortaleza

Marcos Vinicius de Andrade Lima - UECE - *marcosvinicius.lima@aluno.uece.br*
Thales Mesquita Sousa - UECE - *thales.mesquita@aluno.uece.br*
João Batista Carvalho Nunes - UECE - *joao.nunes@uece.br*

Resumo. *É importante que órgãos governamentais no setor educacional tenham embasamento de estudos que envolvam políticas públicas educacionais para a tomada de decisão. A cidade de Fortaleza abriga a quarta maior rede municipal de ensino do País e possui apenas seis Distritos de Educação para dar suporte às escolas. Esta pesquisa busca, por meio da Mineração de Dados Educacionais, mostrar o número e a localização geográfica ideais dos Distritos de Educação, de modo que eles melhor atendam o parque escolar instalado na cidade. Recorreu-se, então, a algoritmos de agrupamento não supervisionado (K-Means, Bisecting K-Means e Gaussian Mixture Model). Os achados da pesquisa estão em sintonia com a nova divisão de Fortaleza em 12 regiões e auxiliam no planejamento de futura redistribuição dos Distritos.*

Palavras-chave: Mineração de Dados Educacionais, Políticas Educacionais, Distrito de Educação.

Abstract. *Government agencies in the educational sector must have scientific bases that involve public educational policies for better decision making. The city of Fortaleza has the fourth largest municipal education network in the country and it has six Education Districts to support schools. This research aims, through Educational Data Mining, to present the ideal number and geographic location of Education Districts, so that they can better serve the municipal education system. For this, we used unsupervised clustering algorithms (K-Means, Bisecting K-Means and Gaussian Mixture Model). The results are in line with the new division of regions in Fortaleza (12 regions) and may help in planning a future redistribution of Districts.*

Keywords: Educational Data Mining, Educational Policies, Education District.

1. Introdução

São essenciais as políticas públicas na área da Educação para a mudança e melhoria de vida das pessoas, influenciando a sociedade como um todo por meio de um planejamento governamental (Duarte; Oliveira, 2018). Os órgãos governamentais precisam ter, portanto, o melhor entendimento do que está sendo realizado por suas instâncias subordinadas na implementação dos programas, projetos e ações.

A senda de estudos da Mineração de Dados Educacionais se expressa como muito útil nessa realidade, portanto é uma área que intenta extrair informações úteis ao processo decisório, com base na análise de dados provenientes do contexto educacional (Romero; Ventura, 2013).

O Município de Fortaleza abriga a quarta maior rede municipal de ensino do País e, em 2019, de acordo com o Censo Escolar (Brasil, 2019) atendeu 151.390 estudantes no Ensino Fundamental, distribuídos em 289 escolas. Somando-se a esta infraestrutura também é preciso considerar a Educação Infantil, responsável pelo atendimento de crianças de um a cinco anos de idade, distribuída em 220 unidades, entre Centros de

Educação Infantil (CEI) e creches. No mesmo ano, ainda segundo o Censo Escolar, foram atendidas 49.047 crianças no Município de Fortaleza (Brasil, 2019).

Como um meio de facilitar a gestão da cidade, visando melhor organização e planejamento municipal por via da descentralização, em 29 de janeiro de 1997, consoante a Lei Municipal nº 8.000/97, a Prefeitura de Fortaleza dividiu a Administração Executiva do Município em seis Secretarias Executivas Regionais (SER). Cada Regional foi formada por bairros circunvizinhos que expressavam semelhanças em termos de necessidades e problemas. Posteriormente, em 2005, foi criada a sétima Secretaria Executiva Regional, responsável pela região do Centro da Capital cearense.

Nas seis primeiras SERs, foram criados os Distritos de Educação, com a finalidade de realizar uma gestão mais próxima das escolas situadas nos respectivos bairros atendidos. A Tabela 1 informa os números da rede escolar sob a responsabilidade de cada SER/Distrito de Educação até o ano de 2013.

Tabela 1. Unidades escolares da rede pública municipal de Fortaleza antes de 2014

	ESCOLAS			TOTAL ESC.	CRECHES		TOTAL CRECHES	TOTAL GERAL
SER	MUN.	ANEXO	ESPECIAL		GESTÃO MUNICIPAL	GESTÃO COMUNITÁRIA		
I	30	17	2	49	11	3	14	63
II	19	8	5	32	1	5	6	38
III	30	16	0	46	3	8	11	57
IV	19	7	1	27	7	4	11	38
V	65	16	0	81	7	20	27	108
VI	63	28	0	91	7	8	15	106
TOTAL	226	92	8	326	36	48	82	410

Fonte: SEDAS/CDIE

Pelos dados da Tabela 1, verificando apenas a quantidade de escolas por SER, é facilmente perceptível um desequilíbrio. Algumas delas possuíam poucas escolas e creches sob sua responsabilidade, como as SERs II e IV, com 38 unidades no total. Entrementes, a SERs V e VI estavam responsáveis por 108 e 106 unidades, respectivamente (quase três vezes mais unidades escolares, comparadas às de menor número). Isso também é verificado ao se calcular o Coeficiente de Variação de Pearson (CVP), que, para essa distribuição, é de 46,23%. De acordo com Falco (2008), o CVP é utilizado quando se intenta comparar a variabilidade de duas ou mais distribuições, sendo seu cálculo definido pelo quociente entre o desvio padrão e a média aritmética da distribuição.

Para tentar equilibrar o número de unidades distribuídas entre os seis Distritos de Educação, com a gestão do Município iniciada em 2012, houve reestruturação da rede de ensino por meio do Decreto nº 13.165, de 27 de maio 2013. A Figura 1 exibe a nova organização do parque escolar com os seis Distritos de Educação e suas respectivas unidades de ensino. Com essa nova distribuição, o CVP baixou para 14,64%.

Embora o valor do CVP se haja reduzido com a última reestruturação, é notório, por intermédio da Figura 1, que o posicionamento dos Distritos de Educação não está situado no melhor local possível. O melhor lugar para instalação de um Distrito de Educação é aquele que consegue oferecer maior proximidade a todas as unidades que ele atende, algo que não acontece quando se observa a Figura 1, pois alguns Distritos estão posicionados nas bordas do seu agrupamento, como observado nos Distritos representados pelas tonalidades verde, cinza, vermelha e vinho. Os Distritos representados pelos tons verde e cinza, inclusive, são próximos um do outro e distantes das unidades educacionais localizadas nas respectivas bordas opostas. Sendo assim, esta pesquisa teve como objetivo: i) identificar o número ideal de Distritos de Educação do Município de Fortaleza, de maneira que as unidades de ensino sejam mais bem atendidas; ii) indicar os locais para a instalação dos Distritos de Educação. É importante destacar a ideia de que, na demanda agora sob relato, apenas a posição geográfica e a quantidade de

escolas foram consideradas para a definição dos agrupamentos. Em estudos subsequentes espera-se investigar, contudo, outros fatores.

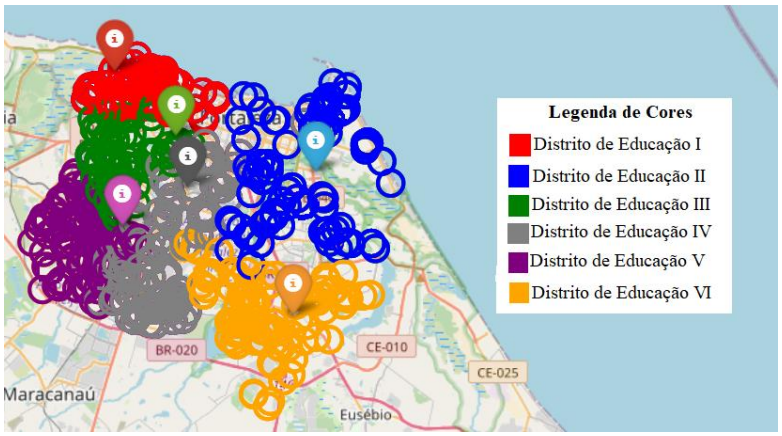

Figura 1. Mapa com a distribuição das escolas por Distritos de Educação

Como modalidade de organização, este artigo está estruturado em mais quatro segmentos, além desta seção, que explicitou a problemática da distribuição dos Distritos de Educação no Município de Fortaleza. A seguinte contém uma discussão sobre Mineração de Dados Educacionais, mostrando sua importância para as políticas educacionais. A terceira seção compreende a metodologia utilizada. Na sequência, são expressos os resultados, para, na seção de fecho – a quinta – serem indicadas as conclusões da pesquisa.

2. Mineração de Dados Educacionais

O contexto atual viabiliza maior apoio de tecnologias digitais sobre o ambiente educacional como um todo. Tal sucede, quer na sala de aula – sobretudo na modalidade a distância, na qual recursos tecnológicos são o meio para quebrar barreiras de tempo e espaço – seja no planejamento institucional, em termos de administração intraescolar; ou mesmo em um panorama mais amplo envolvendo diversas instituições de ensino.

O uso desses recursos computacionais permite grande catalogação de dados a respeito do contexto educacional. Com as técnicas e recursos de análise adequados, esses dados são utilizados para o aprimoramento do ensino, da aprendizagem e da gestão educacional. Com esse fim, surge uma área de estudos focada em analisar indicadores que, de algum modo, tenham relação com o ambiente educacional: a Mineração de Dados Educacionais – MDE (Moissa; Gasparini; Kemczinski, 2015).

De acordo com a Sociedade Internacional de Mineração de Dados Educacionais (2020), essa área se preocupa em desenvolver métodos para explorar dados advindos de fontes educacionais, de modo que se utilizem esses métodos para entender os estudantes e o contexto em que aprendem. A Mineração de Dados Educacionais louva-se, portanto, análises matemáticas, com o apoio de ferramentas computacionais, para tentar encontrar padrões no contexto educacional. Os cálculos realizados levam em conta, normalmente, relações complexas entre grandes volumes de dados, conseguindo extrair informações que seriam de impossível obtenção sem o apoio desses expedientes informáticos. Dos padrões descobertos, emergem conhecimentos úteis, com vistas a embasar tomadas de decisões significativas.

Conceituado processo para a exploração dos dados no contexto da MDE é o KDD (*Knowledge Discovery in Databases* ou, em português, Descoberta de Conhecimento em

Bases de Dados). Este processo interativo e iterativo, que tem a finalidade de extrair conhecimento útil, envolve cinco etapas: **seleção** dos dados, baseada nos objetivos definidos para a análise a ser realizada, e nos algoritmos de mineração utilizados; **pré-processamento**, que remove ruídos, valores ausentes e formata de acordo com os requisitos da ferramenta de mineração; **transformação**, a fim de assimilar identificadores úteis para representar os dados e seleção dos melhores atributos com vistas a representar os dados; **mineração de dados**, com a aplicação de técnicas (classificação, regressão, agrupamento, predição etc.) e algoritmos adequados, visando o atendimento aos objetivos estabelecidos; e, por fim, **interpretação e avaliação** dos resultados identificados (Fayyad *et al.*, 1996).

O processo de mineração de dados envolve a utilização de algoritmos de aprendizado de máquina supervisionado e não supervisionado. O primeiro, chamado de indutor, necessita de um banco de dados de treinamento para o qual já se conheça um atributo classe (também chamado de variável independente ou desfecho). Enquadram-se nesse conceito algoritmos de classificação e regressão. Já os não supervisionados, buscam descobrir padrões de dados não rotulados. Os principais algoritmos dessa categoria são regras de associação, de agrupamento e de sumarização de dados (Borges, 2017). Neste experimento, a técnica mais adequada para o atendimento aos objetivos delimitados é a de agrupamento (ou *clustering*), haja vista que ele configura um

> [...] processo de examinar uma coleção de "pontos" e agrupar os pontos em "clusters" de acordo com alguma medida de distância. O objetivo é que pontos no mesmo cluster tenham uma pequena distância um do outro, enquanto pontos em diferentes clusters estão a uma grande distância um do outro. (Leskovec; Rajaraman; Ullman, 2014, p. 241).

Os algoritmos de agrupamento foram empregados e testados utilizando-se os dados das escolas da rede municipal de ensino de Fortaleza. O percurso metodológico desta busca estar delineado no módulo que vem.

3. Metodologia

O percurso metodológico seguiu o KDD, conceituado anteriormente, obedecendo os cinco passos (seleção, pré-processamento, transformação, mineração de dados, interpretação e avaliação), com o escopo de extrair conhecimento relacionado aos objetivos definidos.

O primeiro passo foi realizar a coleta de dados referentes às unidades de ensino atendidas pela Prefeitura de Fortaleza. Essas informações foram obtidas por meio de uma planilha eletrônica fornecida pela Secretaria Municipal de Educação de Fortaleza (SME). Os dados disponibilizados em formato CSV (*Comma Separated Values* - em português, valores separados por vírgula) continham diversas informações, como nome e endereço de 579 unidades de ensino. Como expediente de preservar a identidade das escolas, em atendimento à Resolução CNS nº 510/2016, foram mostrados neste artigo apenas dados agrupados dessas instituições, de modo que estes não permitem a identificação individual dos referidos estabelecimentos escolares.

Para realizar as operações de agrupamento, era preciso saber, contudo, a localização geográfica de todas as unidades. A solução encontrada foi o recurso a uma API (Interface de Programação de Aplicativos, do inglês *Application Programming Interface*) chamada *Geocoding*, que permite obter, entre outras informações, a latitude e a longitude de endereços fornecidos. A API é fornecida pelo serviço gratuito de mapas *online* chamado *MapQuest*. Com isso, foram pinçadas em lote as coordenadas de cada

uma das escolas. A ferramenta permite que vários endereços sejam enviados por meio de um *link*, e fornece como retorno os dados de geolocalização, obtidos em formato CSV. Então, os dados de latitude e longitude alcançados foram anexados às informações das escolas como novas colunas da planilha de dados dessas unidades de ensino, para que se realizasse a etapa seguinte do KDD.

Depois de localizar todas as unidades de ensino, o arquivo de dados foi filtrado para que apenas as informações relevantes estivessem disponíveis. Na filtragem, foram selecionados: os números da Secretaria Executiva Regional e do Distrito de Educação, o código na Prefeitura de Fortaleza, o nome da escola/creche, a latitude e a longitude.

O arquivo resultante da filtragem foi enviado para o ambiente do *JupyterLab* com a plataforma *Spark*, disponibilizado pelo Laboratório de Sistemas Digitais (LASID) da Universidade Estadual do Ceará (UECE), que possui um *cluster* formado por dez máquinas, contendo 72 TB de armazenamento de disco, 526 GB de memória RAM e com capacidade de executar 232 *threads* simultaneamente. Embora esta pesquisa não necessite de uma infraestrutura deste porte, ela facilitou o trabalho em equipe, pois o *JupyterLab* (ferramenta livre) permite que pesquisadores tenham um *notebook* na Web, possibilitando, inclusive, o uso de várias linguagens de programação num só ambiente (Perkel, 2018). Já o *Spark* é uma plataforma de computação em *cluster* de propósito geral que oferece escalabilidade, flexibilidade e velocidade para quem trabalha com *Big Data* (Shyam *et al.*, 2015).

Com os dados inseridos na plataforma, foi executado o algoritmo para cálculo do *k* ótimo por meio da técnica conhecida como Método Cotovelo (do inglês *Elbow Method*). Em seguida, foram realizadas as tarefas de agrupamento dos dados. Para isso, impôs-se a escolha de três algoritmos, a fim de se analisar a eficiência de cada um deles no agrupamento de escolas em relação aos Distritos de Educação da cidade de Fortaleza: *K-means, Bisecting K-means* e *Gaussian Mixture Model* (GMM).

O *K-means* representa a mais conhecida família de algoritmos de agrupamento de atribuição de pontos. Ele assume um espaço euclidiano entre os pontos em um plano e declara que o número de *clusters k* é conhecido antecipadamente, apesar de admitir que *k* seja atribuído também por tentativa e erro. O algoritmo percorre cada ponto e atribui ele ao *cluster* cujo centróide está mais próximo. Após agrupar todos os pontos aos *clusters*, define novos centroides para os novos grupos formados e, em outra iteração, reagrupa os pontos com base nessa nova posição. Realiza esse processo até que não haja mais mudanças significativas nos *clusters*.

Existem distintas variações do algoritmo *K-means*, com estratégias e técnicas diversas (Jain; Murty; Flynn, 1999). O *Bisecting K-means* é uma importante variação desse algoritmo e é fundamentado no conceito de seleção hierárquica dos dados, para realização do agrupamento. Ele divide os pontos em um número de grupos cada vez maior, formando uma hierarquia, iniciando em um grupo, até chegar em *k* grupos (Fontana; Naldi, 2009). De modo mais detalhado: a) ele agrupa todos os pontos em um *cluster*; b) depois encontra dois *subclusters* utilizando o algoritmo *K-means* clássico; c) repete o passo b) até obter similaridade global; d) repete os passos anteriores até obter *k clusters* (Steinbach; Kumar; Karypis, 2000).

Já o *Gaussian Mixture Model* (GMM) é um algoritmo probabilístico, que representa a probabilidade de um ponto pertencer a um determinado grupo. Segundo Oliveira *et al.* (2011, p. 277),

> Os parâmetros da função de densidade de probabilidade (*pdf*) de uma Mistura Gaussiana são o número de Gaussianas, o coeficiente de ponderação da mistura, a média e a matriz de covariância de cada função de densidade Gaussiana (Resch, 2010). O objetivo é determinar

> o número de Gaussianas que melhor represente a densidade de probabilidade dos dados coletados.

Com os valores definidos para o *k*, por meio do Método Cotovelo, cada um dos algoritmos de agrupamento não supervisionado foi aplicado. Para esse problema, o custo de processamento não foi relevante, dado que o número de registros processados, referentes às escolas analisadas, é de apenas 579 unidades educacionais. O mais importante neste estudo, com base no objetivo principal, coincidiu, portanto, com a localização dos centroides e a quantidade dos elementos selecionados para cada agrupamento (unidades de ensino). Ademais, o cálculo do CVP foi utilizado para a verificação de homogeneidade entre os grupos gerados pelos algoritmos.

Por fim, após o agrupamento das escolas, as localizações dos Distritos de Educação foram plotadas no mapa do Município de Fortaleza, assim como as posições das unidades escolares coloridas da mesma tonalidade do agrupamento de que fazem parte, como meio de facilitar a visualização. Nessa tarefa de plotagem, utilizou-se a biblioteca *folium*, que tem o propósito de unir o poder de manipulação de dados da linguagem *Python* com a força de visualização de mapas da biblioteca *JavaScript Leaflet*.

4. Resultados e Discussões

Após preparação e limpeza dos dados iniciais contendo todas as escolas municipais de Fortaleza, o primeiro passo foi realizar o cálculo do *k* ótimo para ser utilizado nos três algoritmos de agrupamento não supervisionado em uso neste trabalho. O Gráfico 1 denota o resultado do custo calculado para dois até cinquenta agrupamentos. Informa que o valor ideal para o *k* encontra-se no cotovelo, formado por valores de 10 a 20.

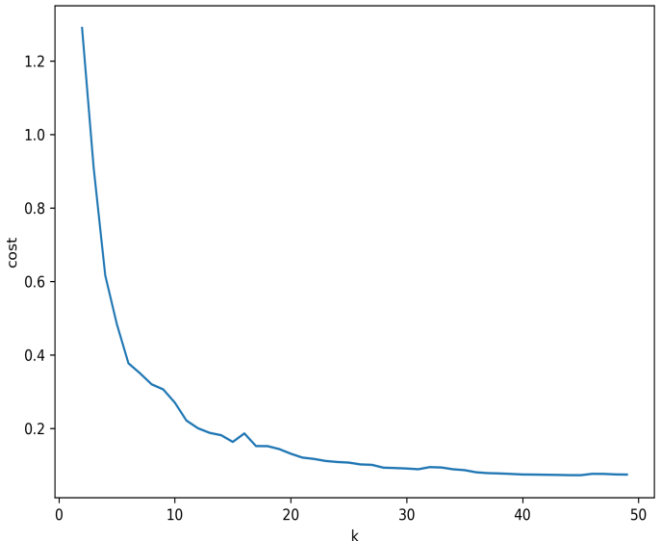

Gráfico 1 - Resultado do Método Cotovelo para *k* ótimo

Para testar os agrupamentos gerados pelos algoritmos, foram selecionados os valores para *k* variando de 10 a 20, além do valor 6, que corresponde à quantidade atual de Distritos de Educação. A Tabela 2 contém o total de unidades educacionais para cada Distrito de Educação após a execução dos três algoritmos. Esses valores são fundamentais para a realização do cálculo do Coeficiente de Variação de Pearson (CVP), que é uma medida de dispersão relativa (como visto na Tabela 3).

Martins e Domingues (2014) definem valores de referência para determinar se um conjunto de dados é mais homogêneo ou mais heterogêneo. De acordo com os autores, se CVP < 15%, há baixa dispersão (mais homogêneo); se 15% ≤ CVP < 30%, há média dispersão; se, no entanto, CVP ≥ 30%, há alta dispersão (mais heterogêneo).

Tabela 2 - Distribuição dos totais de escolas, por Distrito

Algoritmos	Valores de *k*	Unidades Educacionais por Distrito de Educação
K-Means	6	[152, 62, 142, 49, 98, 64]
Bisecting K-Means		[89, 84, 80, 76, 151, 87]
GMM		[50, 127, 120, 87, 91, 92]
K-Means	10	[94, 75, 75, 20, 36, 88, 54, 34, 62, 41]
Bisecting K-Means		[75, 79, 55, 38, 43, 44, 28, 58, 80, 79]
GMM		[28, 44, 77, 59, 89, 42, 134, 42, 32, 32]
K-Means	11	[50, 58, 69, 69, 81, 20, 61, 34, 36, 40, 61]
Bisecting K-Means		[75, 79, 55, 38, 43, 44, 28, 58, 36, 44, 79]
GMM		[26, 48, 132, 28, 45, 87, 25, 22, 41, 54, 71]
K-Means	12	[38, 83, 30, 32, 64, 55, 64, 17, 33, 39, 58, 54]
Bisecting K-Means		[39, 50, 55, 29, 45, 35, 76, 73, 46, 32, 51, 36]
GMM		[26, 62, 20, 36, 31, 83, 91, 25, 23, 44, 26, 100]
K-Means	13	[46, 77, 67, 27, 32, 49, 16, 30, 48, 42, 62, 39, 44]
Bisecting K-Means		[75, 51, 28, 55, 38, 43, 44, 28, 58, 36, 44, 48, 31]
GMM		[18, 80, 94, 48, 5, 45, 56, 32, 16, 13, 138, 21, 13]
K-Means	14	[40, 53, 17, 30, 42, 31, 54, 55, 32, 18, 42, 52, 54, 47]
Bisecting K-Means		[50, 39, 55, 29, 36, 44, 29, 47, 26, 47, 46, 32, 51, 36]
GMM		[29, 26, 30, 92, 14, 32, 20, 32, 14, 63, 61, 13, 25, 116]
K-Means	15	[50, 52, 51, 32, 41, 26, 56, 8, 29, 33, 33, 48, 45, 31, 44]
Bisecting K-Means		[38, 37, 51, 28, 38, 17, 38, 43, 44, 58, 28, 36, 44, 31, 48]
GMM		[39, 7, 24, 13, 30, 18, 28, 41, 9, 24, 85, 39, 61, 145, 16]
K-Means	16	[37, 71, 43, 27, 25, 45, 16, 18, 42, 33, 47, 52, 46, 26, 43, 8]
Bisecting K-Means		[38, 37, 51, 28, 38, 17, 8, 30, 43, 44, 58, 28, 36, 44, 31, 48]
GMM		[32, 25, 41, 0, 87, 33, 29, 8, 29, 30, 61, 21, 10, 123, 18, 32]
K-Means	17	[34, 45, 34, 50, 33, 35, 25, 16, 25, 27, 50, 30, 52, 28, 46, 8, 41]
Bisecting K-Means		[38, 37, 51, 28, 38, 17, 8, 30, 43, 44, 35, 23, 28, 36, 44, 31, 48]
GMM		[50, 47, 33, 19, 22, 27, 24, 39, 21, 32, 5, 18, 71, 25, 64, 25, 57]
K-Means	18	[49, 47, 48, 26, 27, 45, 16, 14, 32, 15, 33, 32, 50, 34, 30, 8, 47, 26]
Bisecting K-Means		[38, 37, 35, 16, 28, 38, 17, 8, 30, 43, 44, 35, 23, 28, 36, 44, 31, 48]
GMM		[16, 80, 13, 6, 13, 25, 47, 39, 76, 27, 22, 9, 0, 99, 26, 32, 24, 25]
K-Means	19	[49, 47, 48, 26, 22, 45, 9, 14, 28, 15, 33, 32, 50, 34, 30, 8, 47, 26, 16]
Bisecting K-Means		[38, 37, 35, 16, 28, 38, 17, 8, 30, 43, 44, 35, 23, 28, 36, 44, 31, 37, 11]
GMM		[31, 28, 16, 85, 27, 0, 73, 9, 24, 0, 74, 32, 4, 39, 25, 38, 31, 32, 11]
K-Means	20	[35, 78, 34, 32, 33, 24, 17, 17, 28, 25, 21, 46, 25, 28, 26, 27, 27, 20, 19, 5]
Bisecting K-Means		[18, 32, 39, 29, 26, 29, 36, 44, 23, 24, 29, 26, 18, 29, 46, 32, 36, 15, 19, 17]
GMM		[13, 23, 18, 0, 18, 49, 11, 36, 14, 59, 29, 3, 19, 32, 25, 90, 28, 1, 15, 84]

Analisando a Tabela 3, observa-se que o GMM exibe os piores resultados com relação ao equilíbrio do número de unidades agrupadas em cada Distrito de Educação. A título de análise, então, foram apenas considerados os resultados obtidos pelos algoritmos *K-Means* e *Bisecting K-means*, como se vê nas Figuras 2, 3 e 4.

O grupo com menor dispersão encontrado é o produzido com base no algoritmo *Bisecting K-Means*, com valor de *k* igual a 14, que possui média dispersão, segundo os valores de referência. Por outro lado, o GMM denota os piores resultados de agrupamentos, no que concerne à dispersão, para o intervalo de *k* variando de 10 a 20. Já os *clusters* gerados pelo algoritmo *K-Means*, apesar de possuírem alta dispersão para todos os valores de *k*, possuem valores bem melhores que os do GMM, sendo o melhor agrupamento o que tem 14 grupos, assim como ocorreu com o *Bisecting K-Means*.

Tabela 3 - Cálculo do Coeficiente de Variação de Pearson (CVP) dos totais de escolas, por Distrito (*k*)

	Algoritmos de Agrupamento não Supervisionado		
Valores de *k*	K-Means	Bisecting K-Means	GMM
6	46,44%	29,71%	29,02%
10	43,08%	33,49%	57,74%
11	34,85%	34,33%	63,16%
12	39,66%	**31,83%**	61,61%
13	37,93%	29,85%	87,84%
14	32,42%	**23,00%**	77,17%
15	33,07%	26,42%	93,05%
16	43,61%	34,75%	85,68%
17	36,07%	32,64%	52,93%
18	41,56%	33,47%	84,41%
19	46,50%	35,75%	79,29%
20	50,87%	31,16%	87,88%

É importante observar que o valor de *k*=6 foi aplicado aos algoritmos a fim de comparar com o agrupamento real utilizado pela Prefeitura de Fortaleza para os seus atuais Distritos de Educação. Comparando-se a Figura 1 com a Figura 2, é fácil perceber como a primeira estava com a localização inadequada de seus Distritos de Educação, em virtude de seus centros de gravidade estarem desequilibrados. Embora o valor do CVP do agrupamento atual das escolas seja de 14,64%, as distâncias de cada escola para o seu respectivo Distrito estão em completo desequilíbrio, trazendo grandes transtornos de deslocamento de equipes do Distrito para a escola ou da escola para o Distrito.

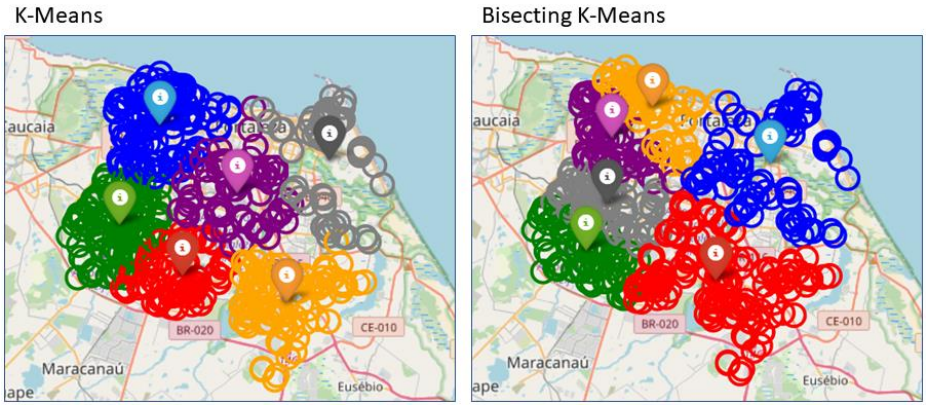

Figura 2 - Mapa com os Distritos de Educação reposicionados (k = 6)

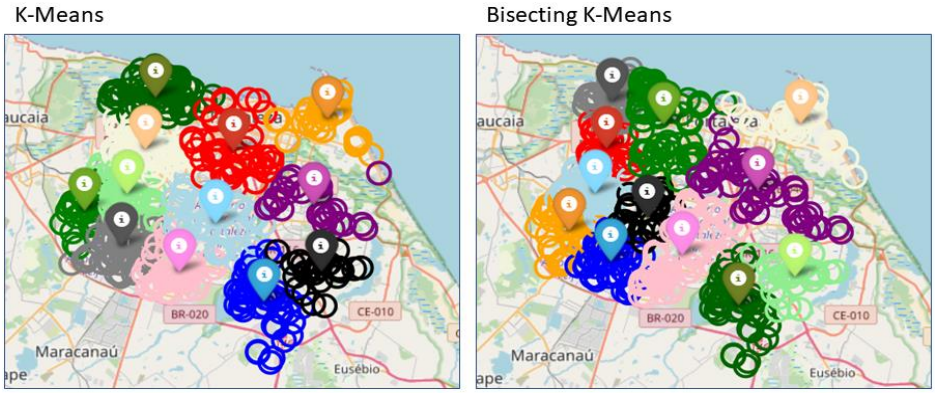

Figura 3 - Mapa com os Distritos de Educação reposicionados (k = 12)

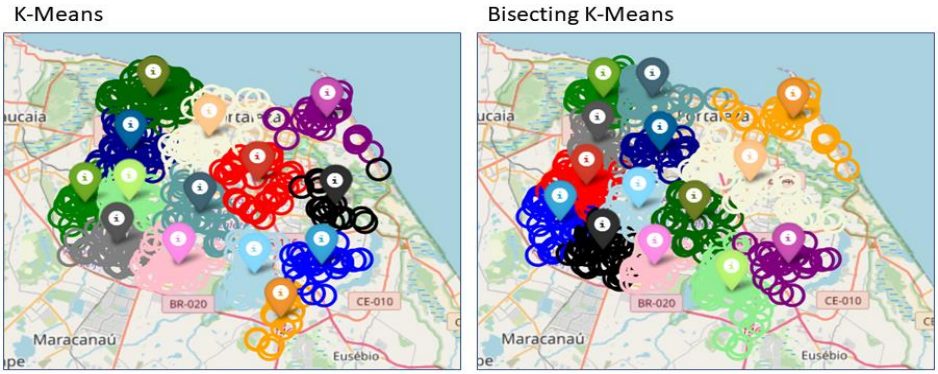

Figura 4 - Mapa com os Distritos de Educação reposicionados (k = 14)

O valor para *k*=12 foi utilizado porque, além de estar dentro da faixa de valores ótimos apontados no Gráfico 1, é o número atual das Secretarias Executivas Regionais de Fortaleza, aprovadas no ano de 2020, de acordo com o Projeto de Lei Complementar nº 0037/2019 (embora ainda não implantadas, no momento em que este artigo foi escrito), que altera a organização e estrutura administrativa do Poder Executivo Municipal. Portanto, espera-se que os Distritos de Educação também sejam ampliados para 12. A Figura 3 encerra o resultado do agrupamento para *k*=12, cujo melhor resultado foi obtido pelo *Bisecting K-Means*.

Quadro 1 - Seleção de bairros para a instalação dos Distritos de Educação

Nº de Distritos de Educação	Geoposicionamento	Bairro	IDH	Classificação do IDH
12	-3.71675446 -38.57693181	Álvaro Weyne	0,364625068	Baixo
	-3.78489533 -38.480862	Edson Queiroz	0,350300888	Baixo
	-3.75057333 -38.52953569	Fátima	0,694795867	Médio
	-3.77778207 -38.59420029	Granja Portugal	0,190184768	Baixo
	-3.82734096 -38.47687489	Guajeru	0,288810144	Baixo
	-3.84594373 -38.51148667	Jangurussu	0,172086984	Baixo
	-3.78815839 -38.62005645	Jardim das Oliveiras	0,198287378	Baixo
	-3.82714208 -38.56303358	Mondubim	0,232790791	Baixo
	-3.79616392 -38.54045765	Parque Dois Irmãos	0,251057366	Baixo
	-3.81101025 -38.5979732	Parque São José	0,284064862	Baixo
	-3.74984054 -38.58317536	Pici	0,218649272	Baixo
	-3.73096143 -38.47259257	Vicente Pinzon	0,331471934	Baixo
14	-3.74713611 -38.59204583	Antônio Bezerra	0,348284739	Baixo
	-3.71762458 -38.55921854	Carlito Pamplona	0,299736489	Baixo
	-3.81852236 -38.58933923	Conjunto Esperança	0,287965762	Baixo
	-3.77326945 -38.50009109	Eng. Luciano Cavalcante	0,522377372	Médio
	-3.77544128 -38.59925079	Granja Portugal	0,190184768	Baixo
	-3.82734096 -38.47687489	Guajeru	0,288810144	Baixo
	-3.84594373 -38.51148667	Jangurussu	0,172086984	Baixo
	-3.79101464 -38.56730929	Maraponga	0,390382558	Baixo
	-3.73147474 -38.47386605	Mucuripe	0,793081592	Médio
	-3.79901054 -38.53359757	Passaré	0,224672553	Baixo
	-3.82867789 -38.55794184	Pref. José Walter	0,395269872	Baixo
	-3.75357161 -38.55457677	Rodolfo Teófilo	0,481883008	Baixo
	-3.7990814 -38.61468837	São Bento	0,198287378	Baixo
	-3.71747091 -38.58870273	Vila Velha	0,271651977	Baixo

Mesmo sabendo que o total de SERs aprovadas para o Município de Fortaleza é igual a 12, também foi utilizado nos agrupamentos um *k* variando de 10 a 20 para verificar o comportamento dos agrupamentos. Com o valor de *k*=14, foi obtida a menor dispersão relativa, medida pelo CVP, de 23,00% para o algoritmo *Bisecting K-Means*.

Para o valor de *k*=20, os resultados dos agrupamentos não foram bons no quesito dispersão da quantidade de escolas, por Distrito. Isso mostra que realmente esse valor já está fora da faixa considerada ótima para *k*.

Analisando os dados da Tabela 2, infere-se que os melhores resultados foram alcançados com o algoritmo *Bisecting K-Means*, para *k*=12 e para *k*=14. Esse último obteve um bom valor para a dispersão relativa, representando um conjunto de valores mais homogêneo, em que a quantidade das unidades de ensino em cada um dos Distritos ficou mais equilibrada.

Ainda se for analisada a localização geográfica de cada um dos centroides para a instalação dos Distritos de Educação sugeridos para *k*=12 e para *k*=14, observa-se a predominância de bairros com baixo Índice de Desenvolvimento Humano (IDH), conforme expresso no Quadro 1.

A predominância de bairros com baixo IDH (91,67% e 85,71%, para uma distribuição de 12 e 14 Distritos, respectivamente) é importante, haja vista o fato de que a Prefeitura de Fortaleza precisa levar infraestrutura e melhores condições de vida para as comunidades que necessitam de maior apoio do Poder Público. A título de comparação, na atual distribuição de seis Distritos de Educação, tem-se que 83,33% das localizações estão posicionadas em bairros com baixo IDH. O posicionamento sugerido por este artigo consegue ampliar, portanto, o acesso ao Poder Público para aqueles que mais necessitam, principalmente se for adotado o agrupamento com 12 Distritos, sendo este o valor esperado por causa do número de SERs aprovado atualmente.

5. Conclusões

Para um Município como Fortaleza, que possui uma das maiores redes municipais de ensino, estar mais próximo das escolas, realizando ações de acompanhamento e apoio às atividades pedagógicas e da gestão, é de fundamental importância para o Poder Público. Nesta perspectiva, os Distritos de Educação desempenham esse papel, quer seja dando assistência à direção, à coordenação, aos professores, aos alunos e até mesmo aos pais. Decidir, entretanto, sobre a quantidade de Distritos e sua distribuição pelo Município não é uma tarefa simples para qualquer gestor. Existem, contudo, procedimentos facilitadores da tomada de decisão, como o uso da Mineração de Dados Educacionais.

Esta pesquisa utilizou uma base de dados das escolas de Fortaleza e aplicou algoritmos de agrupamento não supervisionado, obtendo, além do número ideal de Distritos de Educação para a Cidade, seu posicionamento geográfico; assim como quais agrupamentos demonstraram a menor dispersão relativa, ou seja, quais agrupamentos conseguiram melhor equilibrar a quantidade de unidades escolares por grupo (Distrito).

Foram utilizados três algoritmos: *K-means*, *Bisecting K-Means* e GMM. Destes, apenas o GMM foi descartado para realização de uma análise detalhada, pois demonstrou alto índice de dispersão relativa, chegando até 98,31% (para *k*=20). Já para 12 agrupamentos, os outros dois algoritmos exprimiram desempenho aproximado, embora os resultados obtidos ainda indiquem dispersão relativa alta (acima de 30%). O melhor resultado da dispersão relativa foi encontrado com 14 agrupamentos, quando o índice ficou em 23% (média dispersão), utilizando o algoritmo *Bisecting K-Means*. O agrupamento com *k*=6, por outro lado, foi usado apenas para comparação com os outros

grupos, porquanto esse é o valor atual de Distritos de Educação, e espera-se que a quantidade aumente, acompanhando a ampliação de SERs.

Para a tomada de decisão sobre políticas públicas, também há que se levar em consideração outros fatores, como, num exemplo, as áreas mais vulneráveis e que necessitam de maior atuação do Poder Público. Nesse quesito, analisando-se o IDH de cada bairro sugerido pelas localizações geográficas para os novos Distritos, o maior percentual (91,67%) de bairros com IDH baixo foi atingido com 12 agrupamentos.

Esta demanda de cariz acadêmico atingiu seus objetivos iniciais, pois indicou o número ideal de Distritos de Educação, de acordo com a localização e a quantidade de escolas; assim, também, foram propostas localizações geográficas mais bem distribuídas para os Distritos. Se for levado em consideração apenas o valor da dispersão relativa obtida nos agrupamentos, o número ideal de Distritos é de catorze; mas, se também for observado o percentual de bairros com baixo IDH e notada a existência de estrutura governamental instalada, o número ideal é de doze. Mais estudos devem ser realizados para verificar quais outros fatores também influenciariam na formação dos agrupamentos e seleção dos locais dos futuros Distritos, como, *e.g.*, a quantidade de alunos atendidos em cada Distrito, além de uma análise minuciosa sobre o influxo financeiro no orçamento da Prefeitura de Fortaleza.

Referências

BORGES, V. A. **Definição de um modelo de referência de dados educacionais para a descoberta de conhecimento**. São Carlos: Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional/Universidade de São Paulo, 2017. 184p. Tese de Doutorado.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Censo Escolar: resultados e resumos, 2019**. Disponível em: <http://inep.gov.br/web/guest/resultados-e-resumos>. Acesso em: 15 out. 2020.

DUARTE, M. R. T.; OLIVEIRA, R. de F. Análise de políticas públicas de educação: a importância das narrativas. **Revista de Estudios Teóricos y Epistemológicos en Política Educativa**, v. 3, p. 1-20, 2018.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, p. 37–54, 1996.

FALCO, J. G. **Estatística aplicada**. Cuiabá: EdUFMT; Curitiba: UFPR, 2008.

FONTANA, A.; NALDI, M. C. **Estudo e comparação de métodos para estimação de números de grupos em problemas de agrupamento de dados**. São Paulo: Universidade de São Paulo, 2009. 54p. Relatório.

INTERNATIONAL EDUCATIONAL DATA MINING SOCIETY, 2020. Disponível em: <http://www.educationaldatamining.org/>. Acesso em: 21 set 2020.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM computing surveys (CSUR)**, v. 31, n. 3, p. 264-323, 1999.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. **Mining of massive datasets**. Cambridge: Cambridge University Press, 2014.

MARTINS, G. de A.; DOMINGUES, O. **Estatística geral e aplicada: revisada e ampliada**. São Paulo: Atlas Editora, 2014.

MOISSA, B.; GASPARINI, I.; KEMCZINSKI, A. Educational Data Mining versus Learning Analytics: estamos reinventando a roda? Um mapeamento sistemático. In: **SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO**, 26., 2015, Maceió. Anais. Porto Alegre: Sociedade Brasileira de Computação, 2015, p. 1167-1176.

OLIVEIRA, C. M.; VIEIRA, F. H. T.; SOUSA, M. A.; BORGES, M. A. D. F. Aplicação de misturas gaussianas na análise e modelagem de tráfego VoIP. In: **CONGRESSO DE MATEMÁTICA APLICADA E COMPUTACIONAL (CMAC-SE 2011)**, 33., 2011, Uberlândia. Anais. São Carlos: SBMAC, 2011, p. 276-279.

PERKEL, J. M. Why Jupyter is data scientists' computational notebook of choice. **Nature**, v. 563, p. 145-146, 2018.

RESCH, B. Mixtures of gaussians: **A tutorial for the course computational intelligence.** 2010. Disponível em: <https://www2.spsc.tugraz.at/www-archive/downloads/MixtGaussian.pdf>. Acesso em: 12 dez. 2020.

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 3, n. 1, p. 12-27, 2013.

SHYAM, R.; GANESH, H. B.; KUMAR, S.; POORNACHANDRAN, P.; SOMAN, K. P. Apache spark a big data analytics platform for smart grid. **Procedia Technology**, v. 21, p. 171-178, 2015.

STEINBACH, M; KARYPIS, M. S. G.; KUMAR, V.A comparison of document clustering techniques. In: **TEXTMINING WORKSHOP AT KDD2000**. 2000.