Análise do Programa de Extensão Tecnológica de Pernambuco usando Técnicas de Aglomeração de Dados

Analysis of the Technology Extension Program from the Government of the state of Pernambuco Using Data Clustering Techniques.

Victor Hugo Wanderley Freire 1 orcid.org/0000-0003-2329-020X
Carmelo José Albanez Bastos Filho 2 orcid.org/0000-0002-0924-5341
Emilia Rahnemay Kohlman Rabbani 3 orcid.org/0000-0002-4016-5198

1 Graduação em Engenharia Eletrônica, Escola Politécnica de Pernambuco, Pernambuco, Brasil,
2 Coordenação de Engenharia da Computação, Escola Politécnica de Pernambuco, Pernambuco, Brasil.
3 Coordenação de Engenharia Civil, Escola Politécnica de Pernambuco, Pernambuco, Brasil.

E-mail vhwf@poli.br; carmelo.filho@upe.br; emilia.rabbani@upe.br

Resumo

A demanda por profissionais relacionados à área de STEM é crescente e foi ainda mais acelerado pela pandemia do COVID-19. No Brasil, o baixo investimento e incentivo à área faz com que se forme menos profissionais em STEM do que a demanda. Em Pernambuco o cenário é ainda mais agravado e, pensando nisso, o governo do estado lançou um programa de extensão tecnológica (PET) com o intuito de incentivar a formação de profissionais na área. Tal programa gerou uma quantidade de dados e, com isso, uma demanda de processamento e estudo destes para que seja possível tanto um entendimento de como o programa está funcionando, como para gerar um melhoramento deste para as próximas rodadas. Assim, foram utilizados algoritmos de aglomeração de dados, como Algoritmo Aglomerativo, k-Modes e Mapas SOM para analisar e gerar resultados a partir dos dados coletados. As métricas utilizadas para avaliar os agrupamentos gerados foram coeficientes de silhueta, pureza e método do cotovelo. Os agrupamentos gerados por estas técnicas mostraram características importantes do programa, além de evidenciar que este está sendo bem avaliado pelos seus beneficiários e, então, cumprindo com seu propósito.

Palavras-Chave: Aglomeração de dados, STEM+C, competências técnicas, extensão universitária, coeficiente de silhueta.

Abstract

The growing demand for STEM professionals was accelerated by the COVID-19 pandemic. In Brazil, the low investment and incentive for this particular area generates a deficit of professionals, meaning that there are not as many professionals as the market demands. In the state of Pernambuco, this scenario is even worse, so the government of the state has launched a technological extension program (PET) aiming to decrease this deficit. This program has generated a great amount of data that needed to be studied, analyzed and processed in order to lead to a better understanding of the program and also make it better in the next rounds. Therefore, some clustering techniques were used, like Agglomerative Algorithm, k-Modes, SOM maps with the objective of analyzing and absorbing some insights from the data. The metrics utilized to evaluate those clusters were the silhouette coefficient, purity and elbow method. The clusters resulting from those techniques showed important features of the project, in addition it showed that the project is being well evaluated by the beneficiaries and, so, achieving its goal.

Key-words: *Clustering, STEM+C, harc skills*

REVISTA DE Engenharia e Pesquisa Aplicada

1 Introdução

O acrônimo STEM (Science, Technology, Engineering and Mathematics) é usualmente utilizado para se referir ao ensino e aprendizagem nas áreas da ciência, tecnologia, engenharia e matemática. Intrínseco à definição supracitada, inclui-se atividades educacionais de todos os níveis de forma formal e informal. [1]

O crescimento da demanda por profissionais relacionados à área de STEM é crescente e foi acelerado ainda mais com o surgimento da pandemia do COVID-19. Segundo reportagem da CNN Brasil [3], a procura por profissionais de tecnologia cresceu 671% durante a pandemia global.

No Brasil, devido ao baixo investimento e incentivo a área, é formado menos profissionais quando comparado a demanda e ao potencial do país [4]. No estado de Pernambuco esta realidade de desenvolvimento das áreas relacionadas ao STEM é ainda mais precária. Como evidência se tem o fato de que entre 2005 e 2015, ainda que o IDEB tenha aumentado 44%, as notas do SAEB em matemática e português aumentaram apenas 8% e 11% respectivamente [2].

Devido ao cenário de desenvolvimento tecnológico em Pernambuco, houve, por parte do Governo do Estado de Pernambuco, através da Secretaria de Ciência, Tecnologia e Inovação de Pernambuco, um grande programa de formação em tecnologias habilitadoras de futuro no estado, programa este que atende 5490 pessoas com 90 turmas, diversas empresas, de portes variados, e várias instituições de ensino, sendo financiado pela FACEPE (Fundação de Amparo à Ciência e Tecnologia de Pernambuco) chamado Projeto de Extensão Tecnológica (PET).

Tal iniciativa gerou uma grande quantidade de dados sobre os cursos que são desenvolvidos dentro desta jornada, sobre as atividades, segmentadas pelo setor de aplicação, que são realizadas pelos estudantes junto com as empresas, de forma que os dados necessitam ser processados e minerados para obter um melhor entendimento do projeto.

2 Objetivos

O objetivo do trabalho é, a partir dos dados coletados, encontrar, através de técnicas de inteligência artificial e mineração de dados, relações capazes de mostrar o que têm sido mais frequente em cada um dos territórios e setores, ou seja, entender os padrões gerados pelos projetos tendo em vista em qual região estão inseridos e quais setores estão envolvidos.

Tem-se também como objetivo entender, através dos dados, quais as características dos projetos que geraram resultados positivos, as características dos projetos que obtiveram um resultado não tão bom e o que não aconteceu de forma satisfatória para que futuramente, em próximas rodadas, isto possa ser induzido.

Pretende-se, então, gerar agrupamento de dados, ou seja dos projetos, de forma que estes possam ser entendidos como conjunto, facilitando, assim, seu entendimento e possíveis correções que venham a ser feitas.

3 Fundamentação Teórica

3.1 Projeto de Extensão Tecnológica (PET)

O projeto de extensão tecnológica (PET) é um projeto composto por jornadas que tem como objetivo aumentar o número de parcerias público-privadas em que estudantes, com interesse em formação nas áreas de STEM, estão sendo capacitados e realizando atividades de extensão em empresas privadas. A iniciativa é realizada com estudantes que advém de Instituições de Ensino do Estado de Pernambuco (instituições de Ensino Superior, Escolas Técnicas Estaduais de Educação ou Escolas de Referência de Ensino Médio) e empresas com o objetivo de qualificar pessoas de forma a resolver problemas de forma inovadora, preferencialmente em segmentos de maior intensidade tecnológica, contribuindo, assim, para o desenvolvimento na estrutura produtiva e social do estado [5].

Desta forma, espera-se que o PET gere de forma satisfatória capacitação profissional, melhoria da formação para empregabilidade para os estudantes envolvidos, além de criar parcerias com instituições de ensino, capacitações em conjunto com o setor produtivo e estimular carreiras nas áreas STEM [5].

3.2 Inteligência Artificial

Segundo John McCarthy, inteligência artificial é "a ciência e engenharia capaz de produzir máquinas inteligentes, especialmente programas de computadores inteligentes. É relacionada com a tarefa de usar computadores para entender a inteligência humana, porém IA não está limitada a métodos que são observáveis biologicamente." [6].

De forma sucinta, inteligência artificial é a área que combina ciência da computação e grandes conjuntos de dados com o intuito de resolver problemas [7]. A IA se torna necessária em problemas com conjunto de dados que possuem uma grande dimensionalidade, de forma que os algoritmos aplicados a estes possam vir a fugir da abstração humana quando processados.

DOI: 10.25286/repa.v7i2.2224

As técnicas de inteligência artificial podem ser utilizadas em diversas aplicações, tais como reconhecimento de fala, sistemas de recomendação, visão computacional, problemas de classificação e regressão de dados e vários outros tipos de aplicações.

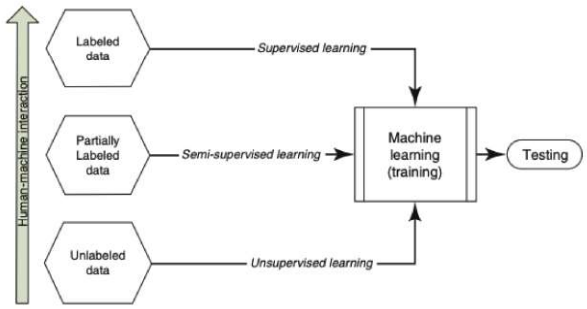
3.3 Machine Learning

Machine Learning ou Aprendizado de Máquina é uma sub-área da inteligência artificial que estuda o desenvolvimento de modelos computacionais de aprendizagem. Do ponto de vista computacional, aprendizado de máquina se refere ao fato da máquina aprender e melhorar seus resultados baseada em resultados anteriores **[8]**.

Um algoritmo de *machine learning* é um processo que usa dados de entrada para realizar uma determinada tarefa, de forma que estes algoritmos são adaptativos e podem alterar sua estrutura através da repetição com o intuito de melhorar seu aprendizado e retornar melhores resultados **[9]**.

De forma generalista, pode-se dividir os problemas de *machine learning*, de acordo com os dados de entrada, em três classes: Aprendizado supervisionado, Aprendizado não supervisionado e Aprendizado semi-supervisionado.

Figura 1 – Tipos de Aprendizado.



Fonte: [9].

3.3.1 Aprendizado não supervisionado

No aprendizado não supervisionado o objetivo da máquina é obter representações dos dados de entrada que possam ser utilizados para auxiliar na tomada de decisão, prever outros dados de entrada, além de ser capaz de comunicar dados entre duas máquinas **[10]**.

Estes tipos de algoritmos podem ser interpretados de forma que sua tarefa é encontrar padrões nos dados que lhes são fornecidos. Tem-se como exemplo de aprendizado não supervisionado os algoritmos de clusterização e de redução de dimensionalidade. **[10]**

3.3.1.1 Técnicas de Clusterização

Dividir um conjunto de dados em grupos com características similares possui muitas aplicações. Particionar elementos em grupos similares pode ter como objetivo gerar ideias a respeito de algumas estruturas inerentes dessa população ou, então, para criar uma estratégia de negócios **[11]**.

Um grupo de clusters pode ser definido como um subconjunto da população total. Os métodos de agrupamento usam estratégias de busca heurística para obter as soluções com a solução mais ótima **[11]**.

3.3.1.1.1 Algoritmo Aglomerativo Hierárquico

O algoritmo aglomerativo hierárquico é uma técnica de machine learning de aprendizado não supervisionado que é bastante conhecida e bem estabelecida. O esquema de partição de forma aglomerativa começa de forma que este algoritmo divide o conjunto de dados em pontos únicos (*nodes*) e os unem, passo a passo, em pares com características mais parecidas, formando um novo ponto e assim por diante **[13]**.

Vários algoritmos de clusterização (agrupamento) utilizam esquemas parecidos com o citado acima, porém eles diferem na forma em que a medida interna de similaridade do agrupamento é computada a cada etapa.

Figura 2 – Definição algoritmica de agrupamento hierárquico.

procedure PRIMITIVE_CLUSTERING(S, d) ▷ S: node labels, d: pairwise dissimilarities
  $N \leftarrow |S|$ ▷ Number of input nodes
  $L \leftarrow [\,]$ ▷ Output list
  $size[x] \leftarrow 1$ for all $x \in S$
  for $i \leftarrow 0, \ldots, N-2$ **do**
    $(a, b) \leftarrow \operatorname{argmin}_{(S\times S)\setminus\Delta} d$
    Append $(a, b, d[a, b])$ to L.
    $S \leftarrow S \setminus \{a, b\}$
    Create a new node label $n \notin S$.
    Update d with the information

$$d[n, x] = d[x, n] = \text{FORMULA}(d[a, x], d[b, x], d[a, b], size[a], size[b], size[x])$$

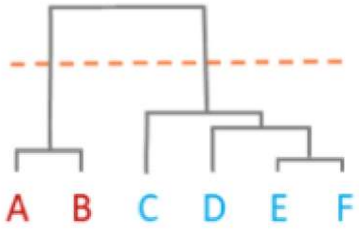
    for all $x \in S$.
    $size[n] \leftarrow size[a] + size[b]$
    $S \leftarrow S \cup \{n\}$
  end for
  return L ▷ the stepwise dendrogram, an $((N-1)\times 3)$-matrix
end procedure

Fonte: [13].

Os algoritmos aglomerativos funcionam de acordo com o mostrado na figura 2. Porém, as abordagens utilizadas podem variar de acordo com o esquema utilizado para calcular a distância entre os agrupamentos. Algumas técnicas para cálculo de distância são: distância euclidiana, distância Ward, centróide, mediana, e etc.

O resultado final de um algoritmo aglomerativo é, usualmente, um dendrograma, que é um diagrama que mostra a relação hierárquica entre objetos **[14]**.

Figura 3 – Dendrograma.



Fonte: [14].

3.3.1.1.2 *Self-Organizing Maps* (SOM)

O mapa SOM, ou mapa de Kohonen, é uma rede neural artificial (RNA) cujas células se tornam sintonizadas especificamente para agregar classes de padrões através de um aprendizado não supervisionado, que, então, produzem uma representação discreta em um espaço de baixa dimensão (normalmente duas) dos dados de entrada **[15]**. Os mapas SOM diferem de outras RNA's pois estas aplicam um processo de aprendizagem competitiva, diferentemente de redes que, por exemplo, utilizam *backpropagation*.

Figura 4 – Pseudo-Código do Mapa SOM.

Input: A set of input vectors $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
Output: A set of weight vectors $W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\}$

Set parameters α_0, σ_0, τ_1, τ_2, k, and t_{max}
Initialize all $\mathbf{w} \in W$ to random values
for $t = 1$ to t_{max} **do**
  Select random $\mathbf{x} \in D$
  Find $\mathbf{w}$ such that $d(\mathbf{x}, \mathbf{w}) = min\{d(\mathbf{x}, \mathbf{w}) \mid \mathbf{x} \in D\}$
  for all $\mathbf{w}$ in neighbourhood h **do**
    Update the weights: $\mathbf{w} = \mathbf{w} + \alpha h(\mathbf{x} - \mathbf{w})$
    Reduce learning rate α
  end for
end for

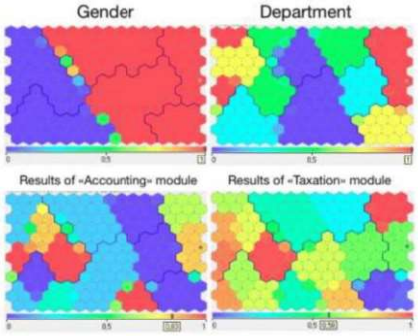
Fonte: [16].

No mapa SOM, alguns parâmetros precisam ser inicializados, conforme mostra figura 3 com o pseudo-código, são eles **[16]** :

1. a_0(*Learning Rate):* a taxa de aprendizado com que a rede aprende os padrões.
2. σ(sigma): raio, ou seja, distância em que a função da vizinhança alcança.
3. t_{max}Número de iterações máxima da rede.

O resultado do processamento do mapa SOM, como o nome diz, é um mapa que possui os agrupamentos, conforme figura 5.

Figura 5 – Exemplos de resultados do Mapa SOM



Fonte: [17].

3.3.1.1.3 K-modes

A clusterização por k-Modes é um método de aglomeração derivado do k-Means, porém é utilizado para dados categóricos, ou seja, dispostos em categorias **[18]**.

Figura 6 – Pseudo-código para o k-Modes

Input: Sequence Set *S*
Output: Clusters
1 Randomly select the initial centroids;
2 **while** *have sequences changed cluster* **do**
3 Clear all temporary centroids *tmpC*;
4 **for each** *point n in S* **do**
5 **for each** *centroid c in C* **do**
6 Calculate the distance between *c* and *n*;
7 **if** *The distance is the nearest found* **then**
8 Move *n* to cluster *c*;
9 Update the number of points in centroid *c*;
10 update the position of *tmpC*;
11 **for each** *centroid c do in C* **do**
12 calculate the new value of position c based on the *tmpC*;

Fonte: [19].

Como o k-modes não é um algoritmo hierárquico, então é necessário estabelecer um número de clusters para que seja inicializado o algoritmo. O número ótimo de agrupamentos usualmente é obtido utilizando o método do cotovelo **[18]**.

DOI: 10.25286/repa.v7i2.2224

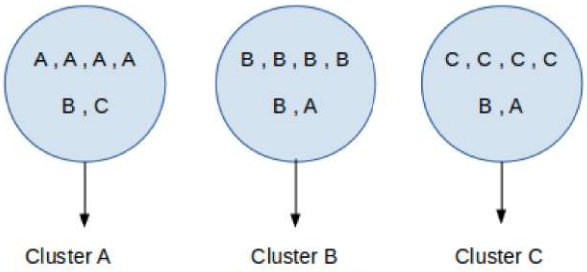
3.3.1.2 Métricas de Avaliação de Modelos de Clusterização

3.3.1.2.1 Índice de Pureza

A pureza é um critério de avaliação externo do cluster para medir a qualidade do agrupamento [20].

Para calcular a pureza, cada agrupamento é atribuído com o nome da classe que é mais frequente neste cluster e, então, a acurácia dessa atribuição é calculada conforme equação 1 [20].

Figura 7 – Exemplo nomenclatura cluster pureza.



Fonte: [22].

$$pureza\ (\Omega, \Phi) = \frac{1}{N}\sum_k \quad max|w_k \cap c_j| \qquad \textbf{(1)}$$

Os clusters com pureza próxima a 0 são agrupamentos ruins e um cluster perfeito possui pureza 1 [20].

3.3.1.2.2 Coeficiente de Silhueta

O coeficiente de silhueta é calculado levando em consideração a distância média dentro do cluster e a distância média do cluster mais próximo de cada ponto dos dados. O coeficiente pode ser calculado conforme equação 2 [21]:

$$coeficiente\ de\ silhueta = \frac{(b-a)}{max(a,b)} \qquad \textbf{(2)}$$

É sabido que um coeficiente de silhueta próximo de 1 significa que aquele determinado ponto está no cluster correto, um coeficiente perto de 0 significa que aquele ponto pode pertencer a outro agrupamento e um coeficiente de valor -1 significa que o ponto está no agrupamento errado [21].

4 Metodologia

4.1 Coleta dos dados

A coleta de dados dos projetos foi realizada através de dois formulários no Google que foram enviados pela coordenação do PET para os coordenadores de cada um dos projetos. O primeiro formulário diz respeito aos dados das empresas em que os projetos estavam sendo aplicados, então este coletou algumas informações como: setor de atividade da empresa, região de desenvolvimento do estado atendida pela empresa, porte da empresa, funcionários da empresa participantes no projeto e, também, o nível de satisfação dos coordenadores dos projetos para com o PET até o momento da resposta. O segundo formulário trata dos cursos oferecidos pelos projetos e coletou informações como: área de conhecimento do curso, tecnologia habilitadora empregada/lecionada, período de execução do curso, carga horária do curso de capacitação, modalidade do curso, quantidade de professores que colaboraram com a execução do curso, instituições de ensino participantes, número de alunos inscritos, selecionados e que finalizaram o curso, além de medir o nível de satisfação dos professores com a realização do curso e do desempenho dos alunos.

4.2 Tratamento dos dados

Inicialmente foi feito, em Python, a junção das respostas de ambos os formulários a partir do número ARC que é único. Após isso, houve a limpeza dos dados através da exclusão de algumas colunas. Houve também a transformação de colunas, como: categorização do porte das empresas e categorização das instituições de ensino. Após estas etapas, um documento em Excel foi gerado, de forma que os dados estavam limpos e organizados e então poderiam ser utilizados para sua aplicação. Ressalta-se que cada algoritmo utilizado possui uma particularidade em relação aos dados, então, para cada um, houve um tratamento diferente.

4.2.1 Algoritmo Aglomerativo e Mapa SOM

Ambos algoritmos, tanto o algoritmo aglomerativo quanto o Mapa SOM requerem dados numéricos como entrada. Através do Python se foi separado o conjunto de dados em duas tabelas, uma tabela apenas com as variáveis numéricas e

uma tabela apenas com as variáveis categóricas. Na tabela com variáveis numéricas foi utilizada a biblioteca *sklearn* e um método de pré-processamento desta que permite que haja uma normalização dos dados, o *StandardScaler* **[23]**. Já na tabela com variáveis categóricas, foi utilizada a técnica *OneHotEncoder* **[24]** para transformar as categorias em números binários.

Após os tratamentos em separado, as tabelas já tratadas foram unidas novamente, gerando um novo arquivo a ser utilizado no processamento de ambas as técnicas.

4.2.3 K-Modes

O algoritmo de clusterização k-modes requer dados categóricos, conforme citado na seção 3.3.1.1.3, ou seja, os dados de entrada do algoritmo não podem ser numéricos. Portanto, partindo do arquivo gerado, os dados numéricos como duração do curso, número de alunos inscritos, número de alunos selecionados, número de alunos que concluíram os cursos e a satisfação foram categorizados, ou seja, foram colocados em intervalos (Ex: se 47 alunos concluíram o curso, esse valor foi substituído pelo intervalo categórico "0a50", referindo-se que o número de alunos que concluíram está entre 0 e 50, conforme figura 8). Desta forma, um novo arquivo foi gerado exclusivamente para ser utilizado no processamento do k-modes.

Figura 8 – Exemplo de tratamento dos dados.

NALUINSC	NALUSEL	NALUCONCL
0a50	0a50	0a50
0a50	0a50	0a50
50a100	50a100	50a100
0a50	0a50	0a50
50a100	0a50	0a50
50a100	0a50	0a50
50a100	50a100	0a50
0a50	0a50	0a50
0a50	0a50	0a50
0a50	0a50	0a50
50a100	0a50	0a50
50a100	50a100	50a100
50a100	0a50	0a50

Fonte: Autor.

4.3 Processamento dos dados

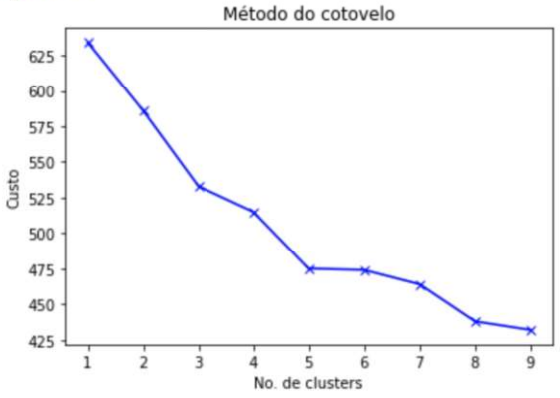
4.3.1 Algoritmo Aglomerativo

O processamento dos dados utilizando o algoritmo aglomerativo foi feito em *Python*, utilizando a biblioteca do sklearn *Agglomerative Clustering*. Para descobrir o número ideal de clusters, utilizou-se o método do coeficiente de silhueta para calcular esta métrica variando o número de agrupamentos.

4.3.2 K-modes

Para o k-Modes, também em Python, utilizou-se o algoritmo kModes da biblioteca kModes **[26]**. Como citado na seção 3.3.1.1.3, esta técnica necessita que se encontre o número ótimo de agrupamentos e, para tal, foi utilizado o método do cotovelo, calculando-se, desta forma, o custo computacional pelo número de clusters, conforme figura 9.

Figura 9 – Método do cotovelo - Kmodes.



Fonte: Autor.

4.3.3 Mapa SOM

O processamento da clusterização utilizando o Mapa SOM foi feito em Python, a partir da biblioteca *minisom* **[27]**. Partido do arquivo gerado na etapa de tratamento de dados, o arquivo foi carregado e se foi utilizado como parâmetros da rede, um sigma 50%, uma taxa de aprendizagem de 0,1 e a distância de ativação sendo a euclidiana. A inicialização do mapa se deu através de pesos utilizando o PCA (*Principal Component Analysis*), um algoritmo de redução de dimensionalidade. Além disso, foram utilizadas 100.000 iterações com o random seed igual a 1.

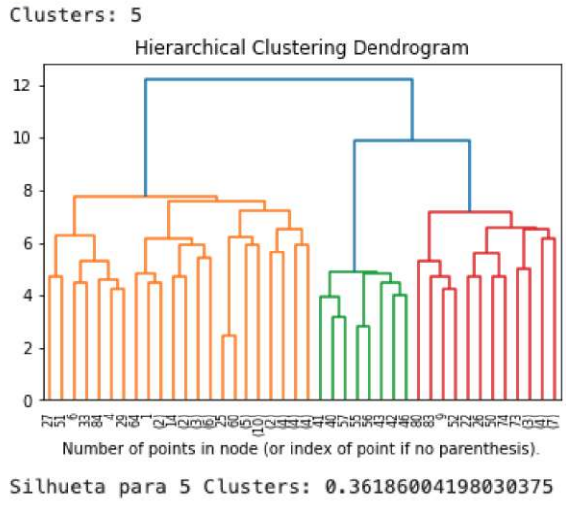
DOI: 10.25286/repa.v7i2.2224

5 Resultados

5.1 Algoritmo Aglomerativo

Para o algoritmo aglomerativo, se testou vários número de clusters de 2 até 9, o que melhor apresentou resultado, a partir do coeficiente de silhueta, foi a clusterização com 5 agrupamentos, conforme mostra a figura 10. O coeficiente de silhueta para este caso foi 0,36. A figura 10 mostra, ainda, o dendrograma para este agrupamento.

Figura 10 – Resultado Algoritmo Aglomerativo



Fonte: Autor.

As características mais frequentes de cada um dos 5 clusters obtidos, estão no apêndice A.

5.2 K-modes

Conforme mostrado na seção 4.3.2 o método do cotovelo concluiu que o número ótimo de clusters era 6. Portanto, foram gerados 6 grupos os quais as características destes estão descritas no apêndice B.

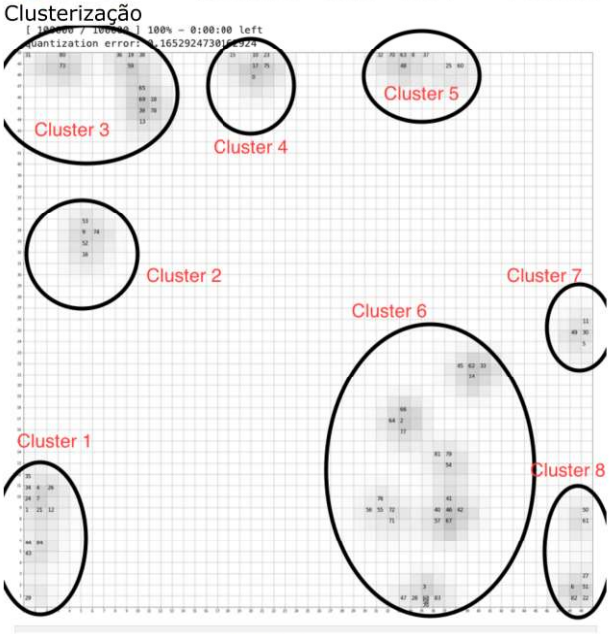
5.3 Mapa SOM

O mapa SOM gera, como diz o nome, um mapa, porém, o número de cluster pode ser arbitrado de acordo com a distribuição gerada. A partir do mapa gerado foram feitas duas distribuições de clusters, como as vistas nas figuras 12 e 13.

Para efeito de avaliação, foram gerados os coeficientes de silhueta para ambas as distribuições. A primeira distribuição, exposta na figura 12, obteve um coeficiente de silhueta de -0,17, enquanto a segunda, exposta na figura 13, obteve um coeficiente de -0,06. Ambos os coeficientes geraram números próximos de 0.

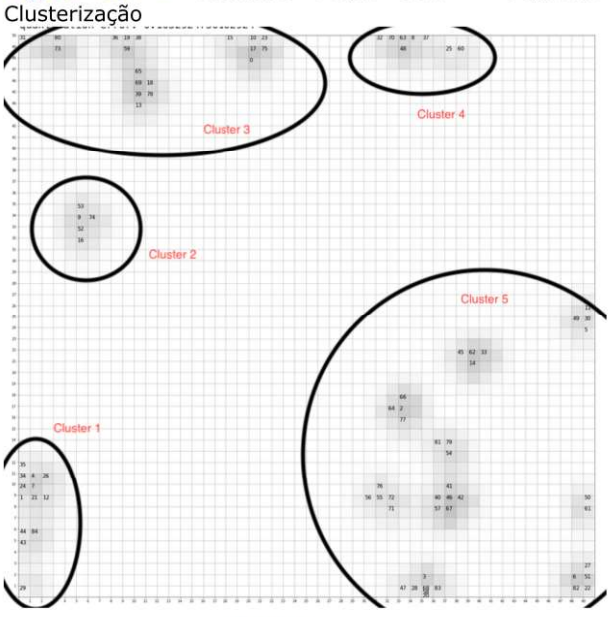
As características que mais aparecem em cada um dos clusters do melhor agrupamento gerado pela rede, ou seja, o segundo agrupamento, estão relatadas no apêndice C.

Figura 12 – Resultado Mapa SOM - Primeira Clusterização



Fonte: Autor.

Figura 13 – Resultado Mapa SOM - Segunda Clusterização



Fonte: Autor.

6 Conclusões

Primeiramente é importante ressaltar que, devido a formas diferentes de tratamento de dados entre as técnicas de k-Modes e Algoritmo Aglomerativo e Mapa SOM, o primeiro ficou impossibilitado de ser comparado com os outros dois, que, por sua vez, podem ser comparados entre si, inclusive pela métrica do coeficiente da silhueta. Porém, isso não inviabiliza uma análise individualizada da clusterização gerada pelo algoritmo do k-Modes.

Quando comparados, os agrupamentos gerados pelo Mapa SOM (os dois) e o algoritmo aglomerativo hierárquico, fica claro, a partir da métrica utilizada, de que o último apresenta uma melhor clusterização, ou seja, agrupa os dados de uma forma melhor, com um coeficiente de silhueta de, aproximadamente, 0,36. Desta forma, este agrupamento gerado, com 5 grupos, pelo algoritmo aglomerativo, pode, entre os três analisados, ser considerado o melhor.

Já analisando o resultado gerado pelo k-Modes, pode-se perceber que, mesmo não havendo uma comparação com algoritmos de outras classes, o agrupamento gerado, ainda, assim passou por métricas de validação, como o método do cotovelo, ou seja, dentro da classe dos agrupamentos possíveis a serem gerados pelo k-Modes, o resultado obtido pode ser considerado um resultado otimizado.

Quando analisada as características mais frequentes dos melhores agrupamentos gerados, os que estão expostos na seção de apêndice, pode-se perceber que os agrupamentos que possuem como origem mais frequentes as instituições privadas, são aqueles que possuem o maior número de alunos inscritos, ou seja, os projetos que estão sendo executados em instituições privadas são aqueles que possuem e ofertam o maior número de vagas. Ainda na análise das características mais frequentes, conclui-se que o projeto tem sido muito bem avaliado em termos de satisfação, visto que todos os agrupamentos tem como maior frequência no quesito satisfação a resposta "Excelente" ou "Bom". A região do estado que mais foi contemplada pelo projeto foi a RMR (Região Metropolitana do Recife), a modalidade mais utilizada para o ensino foi a virtual (síncrona e/ou assíncrona), o setor de comércio e serviços foi bastante contemplado e os projetos contaram com a presença de funcionários das empresas envolvidas.

Portanto, a partir dos dados, do tratamento e aglomeração destes, pode-se concluir que o projeto de extensão tecnológica, com o intuito de fomentar o desenvolvimento da área de STEM no estado de Pernambuco, tem sido extremamente bem avaliado pelos seus beneficiários, está atingindo vários setores tecnológicos, regiões e um número significativo de alunos.

Como sugestão de trabalho futuro e de melhoramento da atual pesquisa, sugere-se utilizar uma forma de tratamento de dados iguais para todas as técnicas utilizadas. Uma dessas formas pode ser se utilizando da matriz de distância com o método de Hamming, visto que este trata de dados categóricos.

Referências

[1] GONZALEZ, H. B., KUENZI, J. J. Science, Technology, Engineering, and Mathematics (STEM) Education: A Primer. Congressional Research Service, Agosto, 2012. Disponível em: <https://fas.org/sgp/crs/misc/R42642.pdf> Acesso em 5 dez. 2021.

[2] BASTO-FILHO, C., GOKHALE, A., & CAMPELLO, B. Relatório: Análise Preliminar e Sugestões para Criação de um Plano para Ampliação e Aperfeiçoamento do Ensino de STEM+C em Pernambuco. Relatório Técnico da Secretaria de Ciência, Tecnologia e Inovação de Pernambuco. 2016.

[3] CNN. Procura por profissionais de tecnologia cresce 671% durante a pandemia . CNN Brasil, São Paulo. 27 out. 2021. Disponível em: <https://www.cnnbrasil.com.br/business/procura-por-profissionais-de-tecnologia-cresce-671-durante-a-pandemia/> Acesso em 5 dez. 2021.

[4] ALMEIDA, C. Sem cientistas de ponta, Brasil fica fora de cadeias globais. O Globo, São Paulo. 08 mar. 2018. Disponível em: < https://oglobo.globo.com/economia/sem-cientistas-de-ponta-brasil-fica-fora-de-cadeias-globais-22454291> Acesso em 5 dez. 2021.

[5] Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE). Edital FACEPE Nº 12/2021. Disponível em: <http://www.facepe.br/wp-content/uploads/2021/06/Edital_FACEPE_-12-2021_PET-20211.pdf> Acesso em 5 dez. 2021.

[6] MC CARTHY, JOHN. What is Artificial Intelligence?. 2004. Disponível em: <https://borghese.di.unimi.it/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_whatisai.pdf> Acesso em: 5 dez. 2021

[7] IBM. Artificial Intelligence (IA). Junho 2020. Disponível em: <https://www.ibm.com/cloud/learn/what-is-artificial-intelligence> Acesso em: 5 dez. 2021

DOI: 10.25286/repa.v7i2.2224

[8] YAO, XIN & LIU, YONG. 2013. Machine Learning. Search Methodologies, 477-517.

[9] El Naqa ,Issam. Murphy, Martin J. (2015).Machine Learning in Radiation Oncology Theory and Applications. Disponível em: < https://link.springer.com/chapter/10.1007/978-3-319-18305-3_1 > Acesso em: 28 de setembro de 2021.

[10] Ghahramani, Z. (2004). Unsupervised Learning. Lecture Notes in Computer Science, 72–112. doi:10.1007/978-3-540-28650-9_5.

[11] Michaud, P. (1997). Clustering techniques. Future Generation Computer Systems, 13(2-3), 135–147. doi:10.1016/s0167-739x(97)00017-4

[12] Ackermann, M. R., Blömer, J., Kuntze, D., & Sohler, C. (2012). Analysis of Agglomerative Clustering. Algorithmica, 69(1), 184–215. doi:10.1007/s00453-012-9717-4

[13] Mullner, Daniel (2011). Modern hierachical, agglomerative clustering algorithms.

[14] Bock, Tim. What is a Dendrogram?. DisplayR. Disponível em: <https://www.displayr.com/what-is-dendrogram/> Acesso em: 12 dez. 2021

[15] Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464–1480. doi:10.1109/5.58325

[16] Kristensen, Terje & Jakobsen, Vemund. Three Different Paradigms for Interactive Data Clustering.

[17] Tynchenko, V. S., Tynchenko, V. V., Bukhtoyarov, V. V., Kukartsev, V. V., Kukartsev, V. A., & Eremeev, D. V. (2019). Application of Kohonen self-organizing maps to the analysis of enterprises' employees certification results. IOP Conference Series: Materials Science and Engineering, 537, 042010. doi:10.1088/1757-899x/537/4/042010

[18] Aprilliant, Audhi. The k-modes as Clustering Algorithm for Categorical Data Type. Medium. Disponível em: <https://medium.com/geekculture/the-k-modes-as-clustering-algorithm-for-categorical-data-type-bcde8f95efd7> Acesso em: 12 dez. 2021

[19] Castro, G. T., Zárate, L. E., Nobre, C. N., & Freitas, H. C. (2019). A Fast Parallel K-Modes Algorithm for Clustering Nucleotide Sequences to Predict Translation Initiation Sites. Journal of Computational Biology. doi:10.1089/cmb.2018.0245

[20] Cambridge University Press. Disponível em: <https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html> Acesso em: 12 dez. 2021

[21] Shahapure, K. R., & Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). doi:10.1109/dsaa49011.2020.00096

[22] Y., Soner. (2021). Evaluation Metrics for Clustering Models. Medium. Disponível em: <https://towardsdatascience.com/evaluation-metrics-for-clustering-models-5dde821dd6cd> Acesso em 12 dez 2021

[23] Sklearn. Standard Scaler. 2021. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html?highlight=standard%20scaler#sklearn.preprocessing.StandardScaler > Acesso em: 12 dez 2021.

[24] Sklearn. One Hot Encoder. 2021. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html?highlight=one%20hot%20encode#sklearn.preprocessing.OneHotEncoder> Acesso em: 12 dez 2021.

[25] Sklearn. Agglomerative Clustering. 2021. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html?highlight=agglomerative%20clustering#sklearn.cluster.AgglomerativeClustering> Acesso em: 12 dez 2021.

[26] KModes. Kmodes. 2021. Disponível em: <https://pypi.org/project/kmodes/> Acesso em: 13 dez 2021.

[27] MiniSom. Disponível em: <https://pypi.org/project/MiniSom/> Acesso em: 13 dez 2021.

Aêndice A: Características Mais Frequentes dos Clusters obtidos com o Algoritmo Aglomerativo

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
AREA	Ciências Exatas e da Terra	Ciências Exatas e da Terra	Ciências Exatas e da Terra	Ciências Exatas e da Terra	Ciências Exatas e da Terra
DURAÇÃO	21 dias	21 dias	13 dias	82 dias	23 dias
CARGA HORARIA	30 a 44h	30 a 44h	Mais de 60h	30 a 44h	Mais de 60h
MODALIDADE	Virtual (Sinc. e Assinc.)	Virtual (Sinc. e Assinc.)	Virtual (Apenas Assíncrono)	Híbrido (presencial e virt.)	Virtual (Sinc. e Assinc.)
NPROF	1	1	3	1	3
ORIGEM	Diversos	Instituição Federal	Instituição Privada	Instituição Estadual	Instituição Privada
NALUINSC	45	17	300	43	411
NALUSEL	50	17	82	50	33
NALUCONC	10	21	72	48	5
SIST_AVALIACAO	Presença, Tarefas ou trabalho	Provas ou Teste	Presença, Tarefas ou trabalho	Presença, Projeto e Desafios	Presença, Tarefas ou trabalho
SATISF_PROF	5	4	5	5	5
SATISF_DESEM	5	4	3	5	4
SATISF_ALUN	5	4	4	5	4
SETOR	Comércio, Serviços e Turismo	Tec. e Info e Comunic.	Comércio	Tec. e Info e Comunic.	Energia
REGIÃO	RMR	RMR	RMR	Mata Norte	RMR
PORTE	Pequenas Empresas	Médio Porte	Médio Porte	Grande Porte	Micro Empresa
FUNC_EMPRESA	Sim	Sim	Sim	Sim	Sim
NFUNCEMPRESA	1 a 2	1 a 2	1 a 2	3 a 5	1 a 2
SATISF	5	5	4	5	4

Apêndice B: Características Mais Frequentes dos Clusters obtidos com o kModes

kModes						
	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
AREA	Ciências Exatas e da Terra	Engenharias	Ciências Sociais Aplicadas	Ciências Exatas e da Terra	Ciências Exatas e da Terra	Ciências Exatas e da Terra
DURAÇÃO	15 a 30 dias	10 a 15 dias	0 a 10 dias	15 a 30 dias	15 a 30 dias	15 a 30 dias
CARGA HORARIA	30 a 44h	30 a 44h	45 a 60h	30 a 44h	Mais de 60h	30 a 44h
MODALIDADE	Virtual (Sinc.)	Virtual (Sinc. e Assinc.)	Híbrido (presencial e virt.)	Virtual (Sinc. e Assinc.)	Virtual (Sinc. e Assinc.)	Virtual (Sinc. e Assinc.)
NPROF	1 a 2	Mais de 5	3 a 5	1 a 2	3 a 5	1 a 2
ORIGEM	Instituição Federal	Diversos	Diversos	Instituição Estadual	Instituições Privadas	Instituição Federal
NALUINSC	0 a 50 alunos	50 a 100 alunos	0 a 50 alunos	0 a 50 alunos	400 a 500 alunos	0 a 50 alunos
NALUSEL	0 a 50 alunos	0 a 50 alunos	0 a 50 alunos	0 a 50 alunos	0 a 50 alunos	0 a 50 alunos
NALUCONC	0 a 50 alunos	0 a 50 alunos	0 a 50 alunos	0 a 50 alunos	0 a 50 alunos	0 a 50 alunos
SATISF_PROF	Excelente	Excelente	Excelente	Excelente	Excelente	Bom
SATISF_DESEM	Excelente	Excelente	Bom	Excelente	Bom	Bom
SATISF_ALUN	Excelente	Excelente	Bom	Excelente	Excelente	Bom
SETOR	Indústria	Indústria	Comércio/Serviço	Comércio/Serviço	Comércio/Serviço	TICeEletrônica
REGIÃO	RMR	RMR	Agreste Central	RMR	RMR	RMR
PORTE	Médio Porte	Pequena Empresa	MEI	Pequena Empresa	Micro Empresa	Pequena Empresa
FUNC_EMPRESA	Sim	Sim	Não	Sim	Sim	Sim
NFUNCEMPRESA	3 a 5	1 a 2	1 a 2	1 a 2	1 a 2	1 a 2
SATISF	Excelente	Excelente	Bom	Excelente	Excelente	Excelente

DOI: 10.25286/repa.v7i2.2224

Apêndice C: Características Mais Frequentes dos Clusters obtidos com o Mapa SOM

Mapa SOM					
	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
AREA	Ciências Exatas e da Terra	Ciências Exatas e da Terra	Ciências Exatas e da Terra	Ciências Exatas e da Terra	Ciências Sociais Aplicadas
DURAÇÃO	10 dias	21 dias	23 dias	21 dias	14 dias
CARGA HORARIA	30 a 44h	30 a 44h	30 a 44h	30 a 44h	45 a 60h
MODALIDADE	Híbrido (presencial e virt.)	Virtual (Sinc.)	Virtual (Sinc. e Assinc.)	Virtual (Sinc. e Assinc.)	Virtual (Sinc. e Assinc.)
NPROF	1	1	3	2	5 ou mais
ORIGEM	Instituição Federal	Diversos	Diversos	Instituição Estadual	Instituição Federal
NALUINSC	45	10	411	44	24
NALUSEL	45	50	50	44	23
NALUCONC	10	8	6	48	12
SIST_AVALIACAO	Presença, Tarefas ou trabalho	Projeto	Presença, Tarefas ou trabalho	Presença, Projeto e Desafios	Provas, tarefas e Trabalhos
SATISF_PROF	5	5	5	5	4
SATISF_DESEM	5	5	5	5	4
SATISF_ALUN	5	5	5	5	4
SETOR	Energia	Tec. e Info e Comunic.	Comércio	Comércio	Comércio
REGIÃO	RMR	RMR	RMR	Mata Norte	RMR
PORTE	Pequenas Empresas	Não Respondeu	Médio Porte	Pequenas Empresas	MEI
FUNC_EMPRESA	Sim	Sim	Sim	Sim	Sim
NFUNCEMPRESA	3 a 5	1 a 2	1 a 2	1 a 2	1 a 2
SATISF	5	5	5	5	5