IDENTIFICAÇÃO DE PADRÕES DE CLIENTES EM UMA COOPERATIVA USANDO AGRUPAMENTO POR K-MEANS

IDENTIFICATION OF CUSTOMER PATTERNS IN A COOPERATIVE USING K-MEANS CLUSTERING

Mateus Eduardo Campaner [1]
Ricardo Petri Silva [2]
Sergio Akio Tanaka [3]

RESUMO

Com o crescente volume de dados gerados pelas organizações, a identificação de padrões torna-se essencial para a tomada de decisões estratégicas. Este artigo explora o uso do algoritmo de agrupamento *K-Means* em uma base de dados real de um *Customer Relationship Management* (CRM) de uma cooperativa, avaliando seu desempenho em um cenário real. Além disso, este estudo investiga a eficácia de métodos como o *Elbow Method* e o *Silhouette Score* para o *K-means*, auxiliando a determinar um número ideal de grupos, bem como diferentes abordagens para a escolha dos valores iniciais de cada agrupamento. Dessa forma, oferece uma visão ampla sobre a configuração do algoritmo *K-Means*, visando a obtenção de resultados mais precisos e uma melhor identificação de padrões. O estudo demonstra a aplicabilidade do algoritmo em diferentes cenários, auxiliando na identificação de padrões de clientes em CRM e sugerindo práticas que podem agregar valor para cooperativas e empresas.

Palavras-chave: K-Means; customer relationship management; agrupamento; cluster; grupos.

ABSTRACT

With the increasing volume of data generated by organizations, identifying patterns has become essential for strategic decision-making. This article explores the use of the K-Means clustering algorithm on a real Customer Relationship Management (CRM) database from a cooperative, evaluating its performance in a real-world scenario. Additionally, this study investigates the effectiveness of methods such as the Elbow Method and Silhouette Score for K-Means, aiding in the determination of the optimal number of clusters, as well as different approaches to selecting the initial values for each cluster. Thus, it provides a comprehensive view of the K-Means algorithm configuration, aiming to achieve more accurate results and better pattern

[1]Discente do Centro Universitário Filadélfia - UniFil
[2]Docente do Centro Universitário Filadélfia - UniFil
[3] Docente do Centro Universitário Filadélfia - UniFil

identification. The study demonstrates the applicability of the algorithm in different scenarios, helping to identify customer patterns in CRM systems and suggesting practices that can add value to cooperatives and businesses.

Keywords: K-Means; customer relationship management; grouping; cluster; groups.

1 INTRODUÇÃO

O volume crescente de dados gerados por empresas e organizações tem ampliado a necessidade de métodos de análise e busca de padrões, visando entender melhor seus clientes (Bannayan; Hoogenboom, 2009). No contexto das cooperativas, a relação com os usuários se torna ainda mais relevante, pois está diretamente conectada às suas atividades econômicas. Conhecer o perfil e o comportamento dos clientes é essencial para a criação de estratégias e para melhorar a oferta de produtos e serviços. Nesse sentido, algoritmos de agrupamento destacam-se como uma ferramenta importante para identificar padrões de comportamento dos usuários, contribuindo para a tomada de decisões mais informadas e eficazes.

O algoritmo de agrupamento *K-Means* é utilizado, principalmente pela sua facilidade de implementação, sendo bastante útil para organizar dados com características em comum (Ahmed *et al*., 2020). Este trabalho aplica o *K-Means* a uma base de dados de um sistema de *Customer Relationship Management* (CRM) de uma cooperativa, para analisar e identificar padrões relevantes sobre os clientes, fornecendo informações que possam melhorar o relacionamento deles com a cooperativa ou aprimorar a abordagem da cooperativa em relação a esses clientes.

Além da base de CRM, foi realizado um estudo para explorar o uso e aplicabilidade do *K-Means* em uma base de dados comum como forma de validação para o uso na base de CRM.

Este estudo foi estruturado para apresentar os principais fundamentos e referências da pesquisa na seção "Fundamentação Teórica". Em "Trabalhos Relacionados", são exploradas informações recentes que complementam o contexto da pesquisa. A seção "Metodologia", detalha o processo, os métodos e os experimentos conduzidos, sendo eles explicados na seção "Experimentação". Na sequência, em "Resultados", é discutido as descobertas obtidas nos experimentos, concluindo com uma análise final na seção "Conclusão".

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção visa apresentar os fundamentos necessários para o entendimento da proposta e da problemática apresentada, com foco na aplicação de algoritmos de agrupamento, como o *K-Means*, em um sistema de CRM. O objetivo é identificar padrões de clientes de uma cooperativa agroindustrial, permitindo uma análise do comportamento dos clientes. A segmentação eficaz da base de clientes proporcionará um melhor entendimento, otimizando as estratégias de marketing e personalizando o atendimento. Assim, a implementação do *K-Means* se mostra essencial para melhorar a relação da cooperativa com seus associados e aumentar a satisfação deles.

2.1 Algoritmos de agrupamento

Algoritmos de agrupamento têm como principal objetivo organizar conjuntos de dados em grupos, sendo conhecidos também como *clusters*, com base em características semelhantes. Esses algoritmos são amplamente utilizados em diversas áreas, como marketing, vendas, saúde e sistemas de CRM, onde agrupam dados com base em comportamentos e ações semelhantes (Sardojno *et al*., 2023). Em seu uso mais comum, atuam em bases de dados, classificando e agrupando informações em diferentes padrões. Essas técnicas de agrupamento têm sido utilizadas por empresas, facilitando campanhas de marketing e a identificação de grupos de clientes (Palnati *et al.* 2024).

2.2 *K-Means*

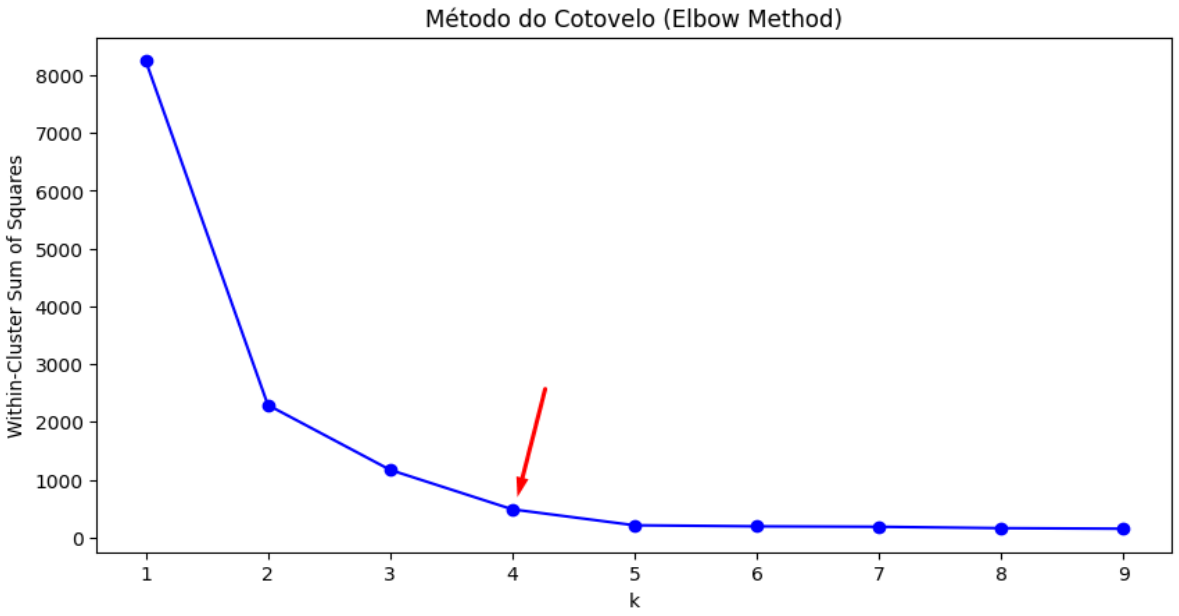
O *K-Means* é uma técnica de agrupamento bastante popular, especialmente pela sua facilidade de implementação e baixo custo computacional (Steinley, 2006). Seu funcionamento começa com a definição do número de *K*, que representa quantos *clusters* serão identificados dentro do conjunto de dados. Cada um desses clusters reflete um padrão específico. Além disso, é necessário definir os centroides iniciais, que são os pontos centrais de cada cluster. Esses centroides podem ser escolhidos de maneira aleatória ou selecionados previamente, dependendo da abordagem desejada (Franti; Sieranoja, 2019).

Após essa etapa, calcula-se a distância entre o centroide e cada ponto próximo, o que permite a definição de um novo centroide. Esse processo é repetido de forma iterativa até que os centros se estabilizem e não se alterem significativamente. Em relação à definição de *K*, se você já souber quantos grupos existem, não será necessário fazer um estudo ou estimativa para determiná-lo. No entanto, se essa informação não estiver disponível, existem alguns métodos que podem auxiliar na escolha do número ideal de grupos.

O método mais comum é o *Elbow Method*, ou método do cotovelo, que utiliza uma métrica denominada WSS (*Within-Cluster Sum of Squares*), a qual calcula a soma das distâncias quadráticas em relação ao ponto central do agrupamento (Ashari, 2022). Com essa métrica, é possível observar a variância dos dados em relação ao número de clusters. Quando essa variância apresenta estabilização após uma queda brusca, define-se o "cotovelo", que corresponde ao valor ideal para o número de *K*.

Na Figura 1, exemplifica-se como é feita a definição de *K*, utilizando a métrica do WSS. Nos pontos em que há uma convergência brusca, sugere-se o valor de *K*, que se assemelha a uma dobra. Na Figura 1, isso é mostrado com o valor 4; após esse valor, o restante segue uma linha contínua, sem alterações aparentes.

Figura 1 – Método Cotovelo para K-Means.



Fonte: Os Autores.

Outro método bastante utilizado para definir o valor de *K* é o *Silhouette Score*, que calcula o coeficiente médio da silhueta para todas as amostras. Nesse cálculo, a partir de um cluster selecionado, considera-se a média das distâncias dos pontos que estão dentro e fora desse *cluster* (Shahapure *et al*., 2020).

Na Fórmula 1, é mostrado o cálculo do *Silhouette Score* em relação a uma amostra i:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i)},$$

onde:

- $a(i)$: é referente a distância média entre a amostra i e todas as outras amostras dentro do mesmo *cluster*.
- $b(i)$: é referente a distância média entre a amostra i e todas as amostras do cluster mais próximo.

O valor do *Silhouette Score* varia de -1 a 1, onde:

- $s(i) \approx 1$: a amostra está bem agrupada com o cluster correto e distante de outros *clusters*.
- $s(i) \approx 0$: a amostra está próxima do encontro de dois clusters.
- $s(i) \approx -1$: a amostra foi agrupada no cluster incorreto.

Após a definição de *K*, são estabelecidos os centroides para os clusters, que podem ser definidos inicialmente de forma aleatória ou por escolha prévia. Com os centroides iniciais definidos, o *K-Means* calcula quais pontos estão mais próximos de cada centroide para formar os grupos. Esse processo é repetido várias vezes até que a posição dos centroides se estabilize. É importante realizar várias execuções com diferentes valores iniciais para os centroides, pois o *K-Means* é sensível à posição inicial deles.

Uma forma de contornar a aleatoriedade na escolha dos centroides iniciais é utilizar o *K-Means*++. Esse método reduz a aleatoriedade ao selecionar os centroides, escolhendo um ponto inicial e, em seguida, selecionando sucessivamente os pontos mais distantes dos já escolhidos até atingir o número de *K* desejado. Esse processo promove uma melhor distribuição dos centroides iniciais, aumentando a eficácia do agrupamento.

2.3 *Customer Relationship Management*

O CRM é um tipo de sistema utilizado principalmente por empresas para gerenciar o relacionamento e as interações com seus clientes, centralizando informações sobre comportamentos, preferências e histórico dos usuários. Com essas informações, é possível personalizar atendimentos e direcionar serviços, o que resulta em um melhor engajamento e retenção dos usuários. O uso de CRM é bastante comum em vendas, marketing e suporte ao cliente, devido às informações que fornece sobre os usuários (Sardojno *et al*., 2023).

Nesses sistemas, o uso de algoritmos de agrupamento, como o *K-Means*, pode ser útil, pois permite a identificação e segmentação de clientes em grupos com base em características comuns, ajudando a revelar padrões de comportamento. Com os resultados gerados pelo *K-Means*, as empresas podem otimizar suas estratégias de marketing e vendas, direcionando abordagens específicas para cada grupo de clientes e, assim, otimizando seus recursos. Além disso, essa segmentação permite que as empresas identifiquem clientes em risco de evasão ou perda, possibilitando ações proativas para evitar esses problemas (Yum *et al*., 2022).

3 TRABALHOS RELACIONADOS

Esta seção trata sobre trabalhos e pesquisas relacionados recentemente em relação ao uso do *K-Means* em CRMs, sendo selecionados artigos e pesquisas referentes a esse tópico publicadas nos últimos 5 anos.

O estudo de (Montero, 2022) demonstrou o uso do *K-Means* aplicado a um sistema CRM de um supermercado, em conjunto com o algoritmo *K-Nearest Neighbors* (KNN). Nesse estudo, foram identificados grupos de clientes com diferentes níveis de fidelidade, o que ajudou o supermercado a compreender melhor o comportamento de seus clientes e a otimizar suas estratégias de marketing, aumentando os volumes de vendas.

Uma forma para melhorar a escolha inicial dos centroides e o uso do *K-Means*++, sendo utilizado para evitar a escolha de centroides muito próximos, o que pode levar a uma má convergência, garantindo que os centroides sejam bem distribuídos desde o começo da execução .

A pesquisa de (Franti; Sieranoja, 2019) aborda o problema da realocação global dos centroides, o que torna o *K-Means* sensível à inicialização, podendo levar a resultados inferiores. Para mitigar esse efeito, é proposto o uso do método *MaxMin*, que utiliza a heurística do ponto mais distante para escolher o centroide inicial em vez de selecioná-lo aleatoriamente. Outra abordagem apresentada é a *Repeat K-Means* (RKM), onde o *K-Means* é executado várias vezes com diferentes valores aleatórios, visando encontrar um agrupamento mais próximo do ideal.

Uma revisão sistemática realizada por (Ikotun *et al*., 2021), mapeou cerca de 27 algoritmos meta-heurísticos inspirados na natureza para melhorar o desempenho do algoritmo *K-means*, principalmente e relação a definição prévia do número de *K* e a definição inicial aleatória dos centroides.

Apesar de o *K-Means* ser bastante utilizado e simples, ele possui limitações importantes, como a necessidade de definir previamente o número de clusters (K) e a escolha dos centroides iniciais, que podem impactar significativamente o resultado (Ahmed *et al*., 2020). A definição do valor de *K* é importante, pois determina o número de grupos que serão formados. Quando a quantidade de clusters é conhecida previamente, os grupos obtidos refletem melhor a realidade dos dados. No entanto, se esse número não é conhecido, é importante estimá-lo utilizando métodos como o *Elbow Method* e o *Silhouette Score* para obter uma divisão mais precisa dos dados.

Outro ponto de atenção com o *K-Means* e a definição dos centroides iniciais, a escolha inicial dos centroides pode influenciar bastante nos resultados, podendo trazer *clusters* inconsistentes e com soluções menos adequadas, para mitigar isso é

necessário realizar múltiplas execuções com os valores dos centroides diferentes a cada execução, porém ao custo de aumentar o tempo de processamento.

Esse problema em relação aos centroides iniciais pode ser resolvido parcialmente com o *K-Means*++, que traz uma melhor escolha dos centros dos *clusters*, porém não elimina totalmente a chance de ter um resultado errôneo.

Vale ressaltar também que o *K-Means* não se adapta bem a dados não-globulares, ou seja, dados que não possuem uma forma esférica ou circular, para isso recomenda-se utilizar algoritmos mais especializados para essa situação, como o DBSCAN (Bushra, 2021).

4 METODOLOGIA

Nesta seção, será descrito a metodologia adotada para a identificação de padrões de clientes em uma cooperativa, para isso foi escolhido o algoritmo de agrupamento *K-Means*. O método escolhido procura agrupar os clientes de acordo com suas características e comportamentos, trazendo assim um melhor entendimento dos mesmos.

Para auxiliar na experimentação foi utilizado uma aplicação desenvolvida pelo autor, a aplicação desenvolvida com a linguagem Python versão 3.11, e a biblioteca Streamlit versão 1.40.1[3], tendo como objetivo melhorar a visualização e filtragem dos dados das bases utilizadas para este estudo.

A Figura 2 traz o passo a passo de como será aplicada a metodologia neste estudo:

Figura 2 – Passo a passo da metodologia a ser usada.

Fonte: Os Autores.

[3]Nome da Aplicação: Visualizador CSV. Disponível em: https://visualizador-csv.streamlit.app/. Desenvolvido por [Mateus Eduardo Campaner], [2024].

4.1 Estudo e tratamento da base

Nesta etapa, foi estudada a base de CRM a ser utilizada na experimentação. Foi utilizada a ferramenta desenvolvida pelo autor para verificar colunas, linhas, tipos de dados, quantidade de valores únicos e o total de linhas presentes na base. Também foi verificado e mantido apenas as colunas que tiverem maior importância, sendo definido mediante um teste de ranqueamento de características.

Para a experimentação, foi utilizado duas bases de dados: uma padrão, que serviu como teste, e outra proveniente de um sistema CRM de uma cooperativa agroindustrial, representando um estudo de caso real. Essa abordagem nos permitiu comparar os resultados obtidos em um ambiente controlado com aqueles de uma situação prática. Ao analisar as informações coletadas, foi compreendido melhor os padrões de comportamento dos clientes da cooperativa e aplicar essas informações para aprimorar as estratégias de relacionamento e atendimento.

4.2 Uso de métodos para definir o *K*

Nesta etapa, foi avaliados alguns métodos para a definição de *K*, sendo eles:

- Uso do *Elbow Method* para definir o valor de *K*
- Uso do Silhouette Method para a definição de *K*

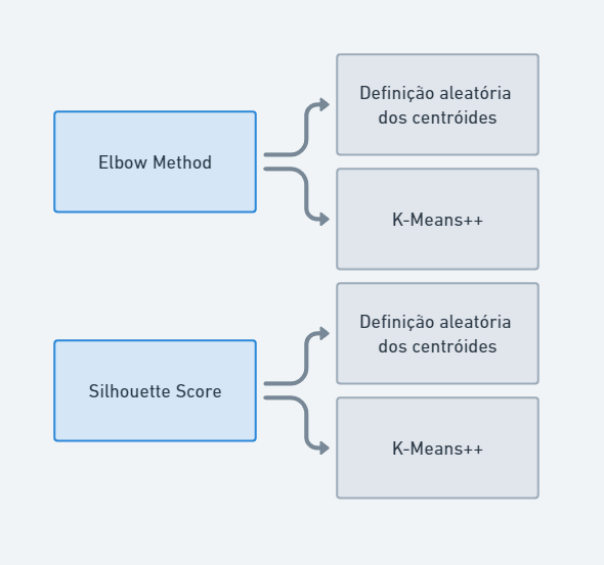
Isso ajudou a verificar qual dos métodos de definição do valor de *K* teve um melhor desempenho para avaliar os resultados propostos por ambos.

4.3 Execuções iterativas do *K-Means*

Nesta etapa, foram realizadas execuções iterativas do algoritmo *K-Means* utilizando a biblioteca *Scikit-learn* da linguagem *Python*. Serão feitas várias execuções utilizando o método *K-Means++* para uma definição menos aleatória dos centroides iniciais, além de execuções com os centroides iniciais definidos de forma aleatória.

Além disso, foram realizadas execuções com as duas abordagens mencionadas acima, sendo o Elbow Method e o Silhouette Score, conforme a Figura 3.

Figura 3 – Fluxograma das execuções dos métodos



Fonte: Os Autores.

4.4 Avaliação dos resultados

Após as execuções mencionadas na Figura 3, aplicadas à base de dados de teste e à base proveniente de um sistema de CRM de uma cooperativa, foi avaliada qual combinação da escolha do método para definição de *K* e dos centroides iniciais se comportou melhor nas bases, levando em conta as definições dos valores de *K*, assim como a segmentação e o tamanho dos grupos formados como fatores.

5 EXPERIMENTAÇÃO

Nesta seção, serão apresentados os experimentos realizados para validar o uso do *K-Means* na identificação de padrões de clientes da cooperativa. Além do *K-Means*, foi avaliado os métodos para definição de *K*, como o *Elbow Method* e o *Silhouette Score*, bem como o comportamento dos centroides iniciais, com e sem o uso do *K-Means*++. Para este estudo, foi utilizado uma base de dados padrão como teste inicial, juntamente com uma base coletada e tratada de um CRM, que contém

cerca de 37.000 linhas e 6 colunas. A aplicação do algoritmo foi feita com o método *K-Means* da biblioteca *Scikit-Learn*[4].

Como base para experimentação foi utilizado a base de dados *Iris Dataset*, como forma de teste para verificar como o *K-Means* se adapta a esse conjunto de dados, e uma base de dados provinda de um sistema CRM de uma cooperativa, como estudo de um caso real, A experimentação foi conduzida seguindo os passos propostos na seção de metodologia.

Em relação à base de dados do CRM da cooperativa foi utilizado a aplicação desenvolvida pelos autores para verificar dados estatísticos sobre a massa de dados, foi identificado 31 colunas com cerca de 140000 registros. Após isso foram realizadas várias etapas de tratamento e filtragem de dados, dentre elas foram feitas remoções de colunas e registros com registros vazios e nulos, foi realizado também um *encoding* de colunas que continham dados categóricos, isso foi realizado para padronizar o tipo dos dados como dados numéricos para possibilitar o uso do algoritmo *K-Means*.

Com o *encoding* foi possível também aplicar um ranqueamento de características utilizando o desvio padrão e a variância dos dados, sendo removidas várias colunas para que a base se adequasse melhor ao *K-Means*, em sua forma final a base ficou com 6 colunas e cerca de 37000 registros. A Figura 4, mostra os resultados do teste de ranqueamento, que ajudou a definir as características mais relevantes a serem utilizadas no *K-Means*.

[4]Biblioteca *Scikit-Learn* disponível em: https://scikit-learn.org/stable/modules/clustering.html

Figura 4 – Ranqueamento de características do CRM

area_cult_tot	0.213600
vlr_potencial_fat	0.192200
vlr_potencial_rec	0.174498
vlr_potencial_geral_rec	0.105740
area_cult_prop	0.102082
perc_geral	0.052990
perc_area_prop	0.033940
vlr_realizado_rec	0.026145
perc_recebimento	0.022486
vlr_realizado_fat	0.021015
perc_faturamento	0.016755
vlr_realizado_geral_rec	0.007615
estrutura_rec	0.005453
estrutura_area	0.004888
cod_produtor	0.003283
grupo_produtor	0.003034
estrutura_geral	0.003004
area_com_mat	0.002886
safra	0.002748
estrutura_fat	0.002546
cdn_repres	0.002162
cod_regional	0.000931

Fonte: Os Autores.

Para a definição de *K*, foram utilizados o *Elbow Method* e o *Silhouette Score* para indicar o valor mais adequado para a segmentação de grupos de clientes, para ambos os métodos foram realizados testes com o valor de *K* variando de 3 a 10 clusters, observando-se o parâmetro WSS para o uso do *Elbow Method* e a pontuação para o *Silhouette Score*. Para os centroides iniciais foi avaliado o uso do *K-Means++* e sem o uso dele nas execuções, visando avaliar a influência dos centroides iniciais nos resultados.

Os experimentos foram executados usando a linguagem *Python* e a biblioteca *Scikit-Learn*. Para cada valor de *K*, o algoritmo foi executado 10 vezes, garantindo a estabilidade dos resultados. Em cada execução, os clusters gerados foram analisados quanto ao volume de dados e à segmentação. O Quadro 1 apresenta os experimentos realizados, incluindo seus nomes e descrições.

Quadro 1 – Experimentos realizados com diferentes bases de dados

Número Experimento	Base Utilizada	Valor de K	Método de Inicialização
1	Iris	3	K-Means++
2	Iris	3	Centroides Aleatórios
3	CRM	3	K-Means++
4	CRM	3	Centroides Aleatórios
5	CRM	4	K-Means++
6	CRM	4	Centroides Aleatórios

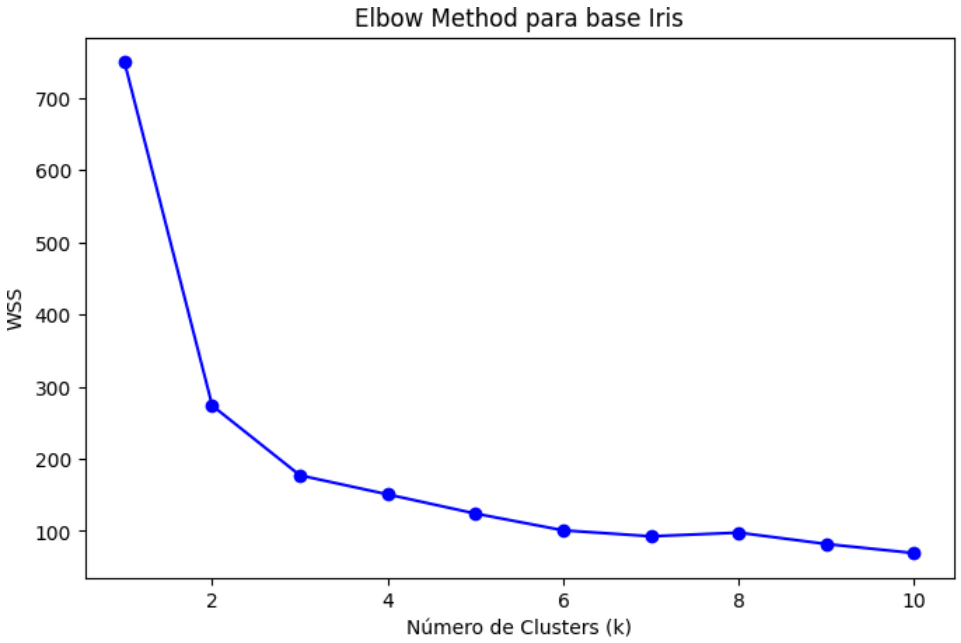
6 RESULTADOS

Nesta seção, serão discutidos os resultados obtidos com a aplicação do *K-Means* em duas bases de dados distintas: a base de dados *Iris Dataset*, usada para validar o comportamento do algoritmo, e uma base de dados de CRM de uma cooperativa. A seguir, são apresentados os principais resultados obtidos nas duas bases durante a etapa de experimentação, com ênfase nos clusters formados e nos padrões identificados.

6.1 Resultados da Iris Dataset

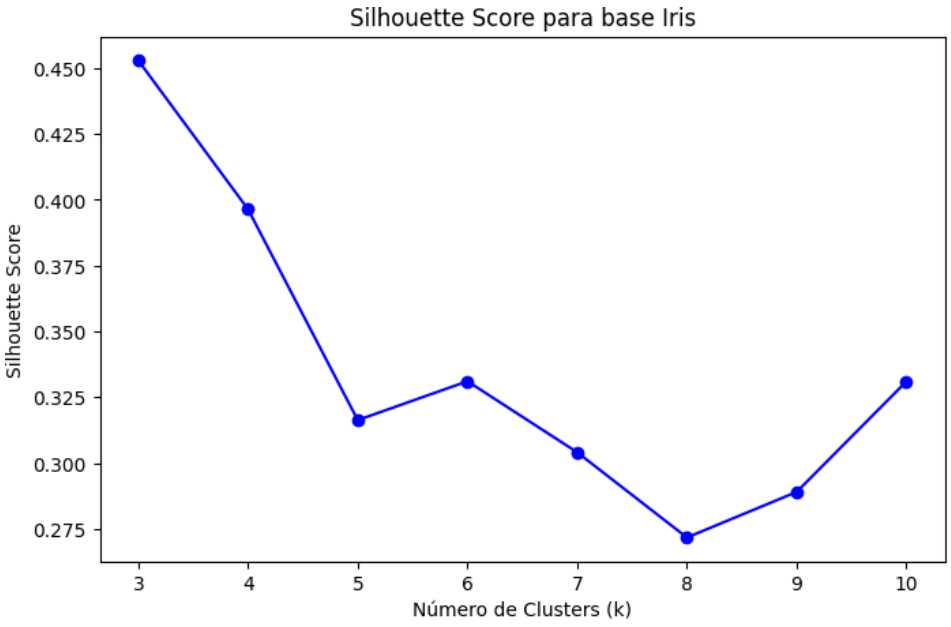
No experimento, o *Iris Dataset* foi utilizado para validar o comportamento do *K-Means* no conjunto de dados. Esta base possui três classes de flores (Setosa, Virgínica e Versicolor), com quatro atributos relacionados ao comprimento e largura das sépalas e pétalas. Conforme mostrado na Figura 5, foi aplicado o *Elbow Method*, e observou-se que o valor do "cotovelo" se situa em 3 (Figura 6). Ao aplicar o *Silhouette Method*, identificou-se que o melhor valor de *K* é 3, pois esse valor foi o mais próximo de 1, indicando o agrupamento mais adequado em comparação com os demais *clusters*.

Figura 5 – *Elbow Method* para base Iris.



Fonte: Os Autores.

Figura 6 – *Silhouette Score* para base Iris.



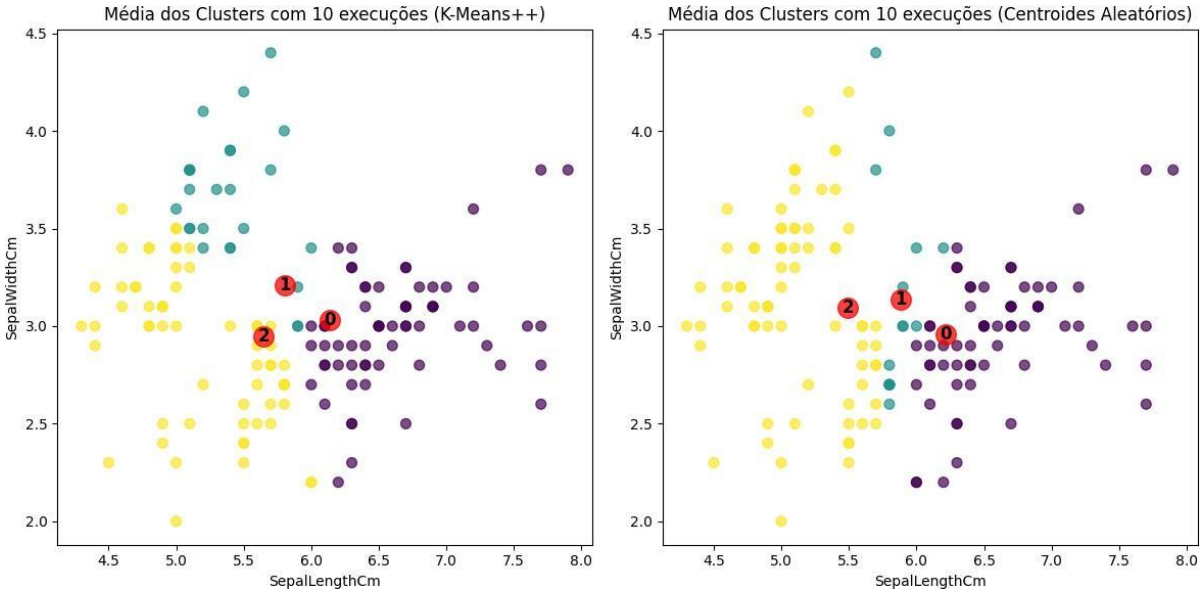
Fonte: Os Autores.

Verificou-se que ambos os métodos conseguiram prever o melhor valor de *K*, já que é conhecido que a *Iris Dataset* possui 3 conjuntos de dados. Com o valor de

K=3 foi aplicado o *K-Means*, sendo realizadas 10 execuções com os centroides iniciais escolhidos de forma aleatória e com o *K-Means*++ minimizando essa aleatoriedade inicial.

A Figura 7 apresenta um gráfico mostrando os clusters formados conforme as configurações dos experimentos 1 e 2 respectivamente, seguindo o Quadro 1.

Figura 7 – Média de execuções do *K-Means* com 3 *clusters* para *Iris Dataset.*



Fonte: Os Autores

A partir desse dados observa-se que o *K-Means*++ como método de inicialização performou melhor em relação à escolha a partir de centroides aleatórios durante a média de 10 execuções, trazendo uma melhor separação dos grupos, servindo como validação inicial para o uso do *K-Means* na base de CRM da cooperativa.

O Quadro 2 apresenta a quantidade de amostras presentes em cada cluster utilizando o *K-means*++, já o Quadro 3 mostra a quantidade de valores presentes em cada cluster utilizando centroides aleatórios.

Quadro 2 – Quantidade de valores em cada *cluster* (*K-Means*++).

Número Cluster	Quantidade de dados	Porcentagem dos dados
0	64	42,38%
1	25	16,56%
2	61	40,26%

Quadro 3 – Quantidade de valores em cada *cluster* (*K-Means* com Centroides Aleatórios).

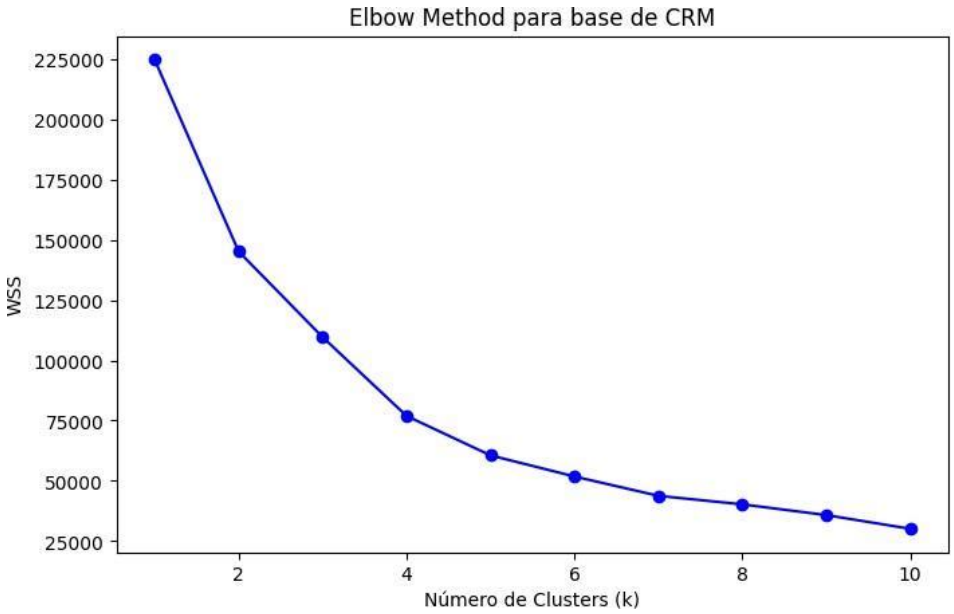
Número Cluster	Quantidade de dados	Porcentagem dos dados
0	64	42,67%
1	15	10,00%
2	71	47,33%

6.2 Resultados da base de CRM da cooperativa

A base de dados do CRM da cooperativa, apresenta informações de produtores como área da propriedade, e valores relativos a faturamento e recebimento, para o uso do *K-Means* foi utilizado as características do valor potencial de faturamento (vlr-potencial-fat) e área de cultura total (area-cul-tot), que foram as que se obtiveram uma maior importância no ranqueamento de características feito.

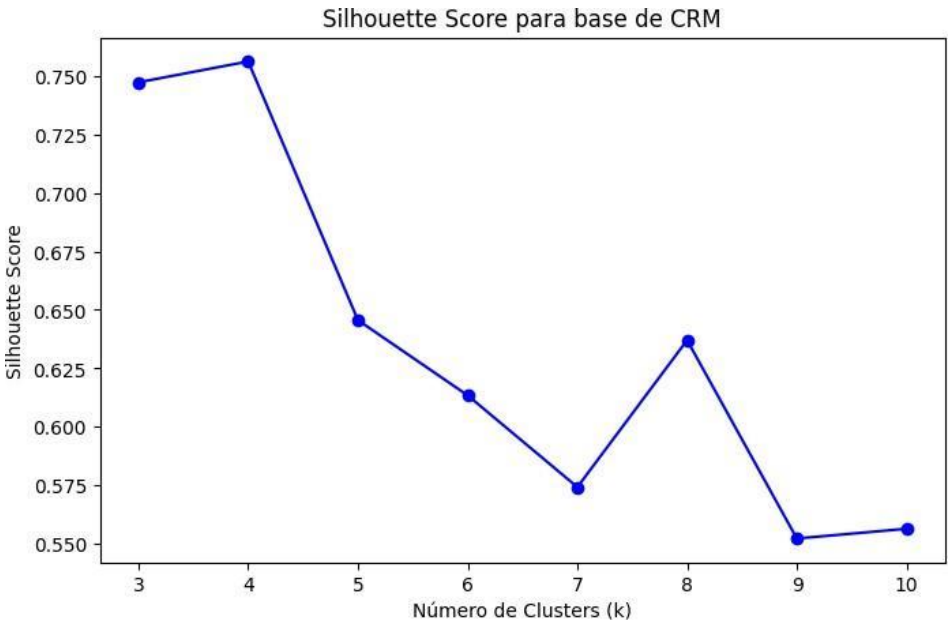
Conforme a Figura 8, foi aplicado o *Elbow Method*, observando-se que a dobra ocorre em *K*=4. Na Figura 9, mostra-se que *K*=4 é o melhor valor; no entanto, identificou-se que o valor *K*=3 está muito próximo de *K*=4.

Figura 8 – *Elbow Method* para base de CRM.



Fonte: Os Autores.

Figura 9 – *Silhouette Score* para base de CRM.

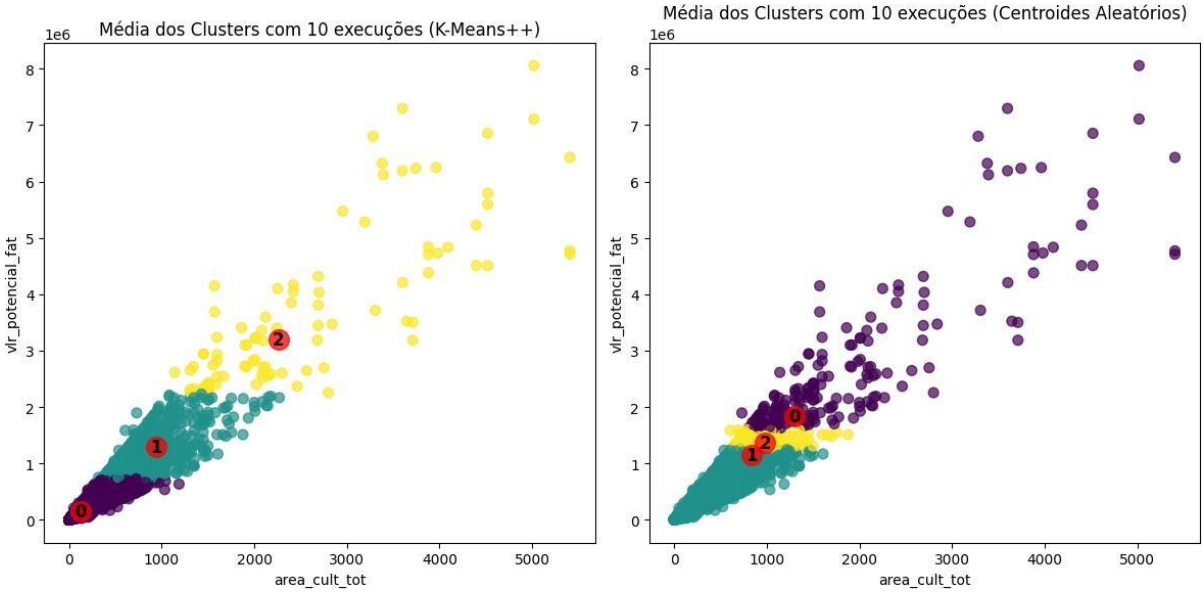


Fonte: Os Autores.

Com isso os dois métodos apontaram o melhor valor de *K*=4, porém como o *Silhouette Score* apontou *K*=3 como um valor próximo, foram realizadas duas execuções com o *K-Means* sendo uma com *K*=3 e outra com *K*=4. A Figura 10

apresenta um gráfico mostrando os clusters formados conforme a configuração dos experimentos 3 e 4, seguindo o Quadro 1.

Figura 10 – Média de execuções do *K-Means* com 3 *clusters* para base de CRM.



Fonte: Os Autores.

O Quadro 4 apresenta a quantidade de amostras presentes em cada cluster utilizando o *K-means*++, já o Quadro 5 mostra a quantidade de valores presentes em cada *cluster* utilizando centroides aleatórios.

Quadro 4 – Quantidade de valores em cada *cluster* (K-Means++).

Número Cluster	Quantidade de dados	Porcentagem dos dados
0	36286	96,89%
1	1071	2,86%
2	89	0,24%

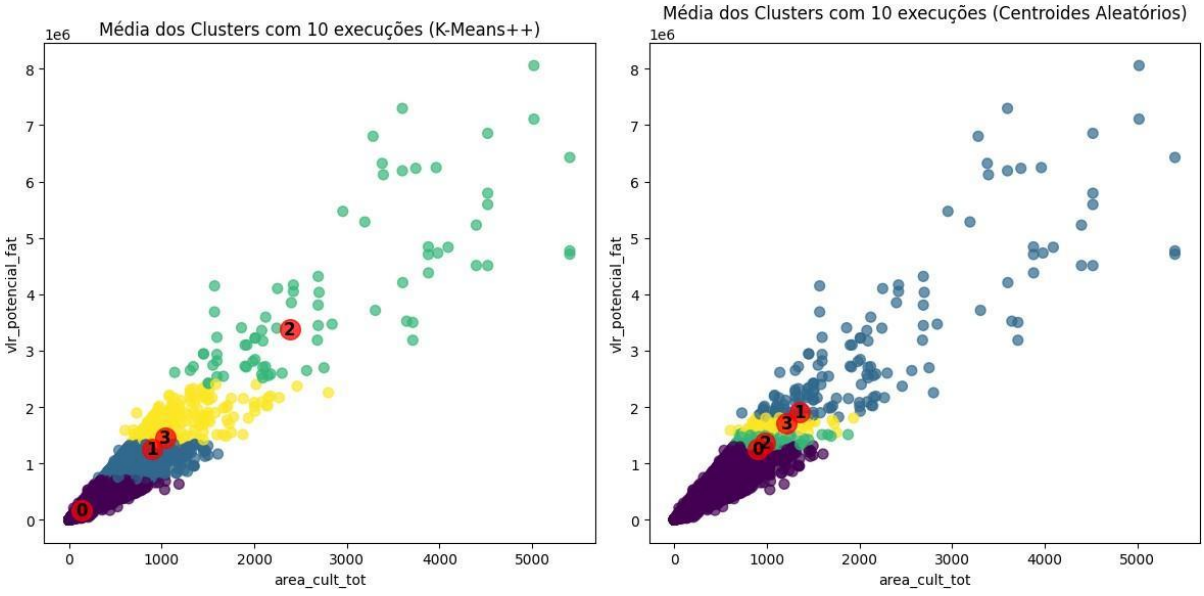
Quadro 5 – Quantidade de valores em cada *cluster* (K-Means com Centroides Aleatórios).

Número Cluster	Quantidade de dados	Porcentagem dos dados
0	194	0,52%
1	37095	99,06%
2	157	0,42%

Com esses resultados, observa-se que o *K-Means*++ teve um desempenho superior como método de inicialização em comparação à escolha de centroides aleatórios, resultando em uma melhor separação dos grupos em média, considerando 10 execuções. No entanto, nota-se que os grupos apresentam tamanhos distintos, com um *cluster* contendo a grande maioria dos dados, um cluster intermediário com cerca de 1.000 registros e um *cluster* menor com apenas 89 amostras.

A execução como o valor de *K*=4 se encontra na Figura 11 apresentando um gráfico mostrando os clusters formados conforme os experimentos 5 e 6, seguindo o Quadro 1.

Figura 11 – Media de execuções do *K-Means* com 4 *clusters* para base de CRM.



Fonte: Os Autores.

O Quadro 6 apresenta a quantidade de amostras presentes em cada *cluster* utilizando o *K-means*++, já o Quadro 7 mostra a quantidade de valores presentes em cada *cluster* utilizando centroides aleatórios.

Quadro 6 – Quantidade de valores em cada *cluster* (*K-Means*++).

Número Cluster	Quantidade de dados	Porcentagem dos dados
0	36255	96,81%
1	904	2,41%
2	77	0,21%
3	210	0,56%

Quadro 7 – Quantidade de valores em cada *cluster* (*K-Means* com Centroides Aleatórios).

Número Cluster	Quantidade de dados	Porcentagem dos dados
0	37130	99,18%
1	145	0,39%
2	106	0,28%
3	55	0,15%

Nos experimentos 5 e 6, observou-se novamente um melhor desempenho do método de inicialização *K-Means*++, sendo o mesmo com desempenho superior em 3 testes realizados, trazendo novamente uma melhor separação e representação dos clusters em uma média de 10 execuções. Como também no experimento com a base de CRM da cooperativa como *K*=3, notou-se novamente a diferença de tamanho entre os *clusters*, tendo um grupo representando a grande maioria dos dados, e os demais *clusters* representando uma minoria.

Com base no Quadro 7, nota-se, também, que os *clusters* 1, 3, 2, possuem tamanhos diferentes e com uma certa progressão de tamanho, sendo o *cluster* 1 e o *cluster* 2 menor.

7 CONCLUSÃO

Neste estudo, foi aplicado o algoritmo de agrupamento *K-Means* em duas bases distintas: uma para testes e outra proveniente de um sistema de CRM de uma cooperativa, que contém dados como área da propriedade e valores relacionados a faturamento e recebimento.

Nos experimentos com o *Iris Dataset* observou se tanto que o *Elbow Method* e o *Silhouette Score*, indicaram o valor de *K*=3, estando correto com base em conhecimento prévio sobre o conjunto de dados, já nos experimentos 1 e 2 do Quadro 1, o método *K-Means*++ se desempenhou melhor na formação dos *clusters*, reduzindo a aleatoriedade dos centros dos grupos formados.

Na base de dados do CRM da cooperativa, o *Elbow Method* apontou o *K*=4 como um valor ideal, o *Silhouette Score* mostrou também *K*=4 como um valor ideal, apesar de sugerir *K*=3 como um valor adequado também possuindo pouca diferença do *Silhouette Score* em comparação com *K*=4.

Observou-se novamente nos experimentos 3, 4, 5 e 6 do Quadro 1 que o *K-Means*++ se mostrou uma opção superior ao uso de centroides aleatórios, especialmente com os valores de *K*=3 e *K*=4 para a base. Essa abordagem proporcionou uma melhor separação entre os clusters. No entanto, nota-se a presença de um *cluster* que representa a maioria dos dados, enquanto 2 a 3 *clusters* representam grupos menores. Para compreender o motivo desse comportamento, é necessária uma análise mais aprofundada.

Uma possível hipótese para esse fenômeno é que as características escolhidas e consideradas mais importantes, (valor potencial de faturamento e a área de cultura total) podem refletir a realidade do mercado. Essas características indicam que há um número significativo de clientes com uma área de cultura menor e um potencial de faturamento igualmente baixo. Em contrapartida, existem poucos clientes que possuem uma área de cultura muito maior e um potencial de faturamento elevado. Entre esses dois grupos, há uma quantidade intermediária de clientes que apresenta valores médios em ambas as métricas.

Essa hipótese apresentada, juntamente com os experimentos realizados, ajuda a trazer uma melhor visão para as cooperativas e sistemas de CRM, podendo os

mesmos tomarem decisões com base em grupos maiores com um faturamento menor ou grupos menores com mais faturamento.

Como direções para trabalhos futuros, este estudo poderia explorar outras configurações para o *K-Means,* conforme descrito no trabalho de (Ikotun, 2021). Além disso, seria interessante investigar o uso de outros algoritmos de agrupamento, como DBSCAN ou *MeanShift*, para avaliar o desempenho desses métodos em comparação ao *K-Means*.

REFERÊNCIAS

AHMED, M.; SERAJ, R.; ISLAM, S. M. S. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. **Electronics**, v. 9, n. 8, p. 1295, 12 ago. 2020.

ASHARI, I. F. et al. Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies. **Journal of Applied Informatics and Computing**, v. 6, n. 1, p. 07-15, 14 jul. 2022.

BANNAYAN, M.; HOOGENBOOM, G. Using pattern recognition for estimating cultivar coefficients of a crop simulation model. **Field Crops Research**, v. 111, n. 3, p. 290–302, abr. 2009.

BUSHRA, A. A.; YI, G. Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. **IEEE Access**, v. 9, p. 87918–87935, 2021.

FRÄNTI, P.; SIERANOJA, S. How much can k-means be improved by using better initialization and repeats? **Pattern Recognition**, v. 93, p. 95–112, 1 set. 2019.

IKOTUN, A. M.; ALMUTARI, M. S.; EZUGWU, A. E. K-Means-Based Nature-Inspired Metaheuristic Algorithms for Automatic Data Clustering Problems: Recent Advances and Future Directions. **Applied Sciences**, v. 11, n. 23, p. 11246, 26 nov. 2021.

MONTERO, Z. D. Customer Grouping for Customer Relationship Management Optimization with the K-Means Algorithm. **Journal of Computer Scine and Information Technology**, p. 98–105, 31 out. 2022.

NEHA REDDY PALNATI; VIJAY; NIKHIL BAYYAVARAPU. Leveraging Machine Learning For Enhanced Database Integration. **Procedia computer science**, v. 235, p. 1623–1633, jan. 2024.

SHAHAPURE, K. R.; NICHOLAS, C. Cluster Quality Analysis Using Silhouette Score. *In*: INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ADVANCED ANALYTICS (DSAA), 7., 2020, Sydney. **Proceedings** […]. Sydney: IEEE, 2020.

STEINLEY, DOUGLAS. K-means clustering: A half-century synthesis. **British Journal of Mathematical and Statistical Psychology**, v. 59, n. 1, p. 1–34, maio 2006.

WAHYU SARDJONO; ACHMAD CHOLIDIN; JOHAN, N. Implementation of Artificial Intelligence-Based Customer Relationship Management for Telecommunication Companies. **E3S Web of Conferences**, Indonésia, v. 388, p. 03015–03015, 1 jan. 2023.

YUM, K.; YOO, B.; LEE, J. Application of AI-based Customer Segmentation in the Insurance Industry. **Asia Pacific Journal of Information Systems**, v. 32, n. 3, p. 496–513, 30 set. 2022.