UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Segmentação comportamental de utilizadores de cartão de crédito utilizando o algoritmo de máquina não supervisionado *K-means*

Rodrigo Pilatti

USP - São Carlos

2023

Rodrigo Pilatti

Segmentação comportamental de utilizadores de cartão de crédito utilizando o algoritmo de máquina não supervisionado *k-means*

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Aprendizado de Máquina

Orientadora: Profa. Dra. Kalinka Regina Lucas Jaquie Castelo Branco

USP - São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Pilatti, Rodrigo
P637s Segmentação comportamental de utilizadores de cartão de crédito utilizando o algoritmo de máquina não supervisionado k-means / Rodrigo Pilatti; orientadora Kalinka Regina Lucas Jaquie Castelo Branco. -- São Carlos, 2023.
39 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. . I. Lucas Jaquie Castelo Branco, Kalinka Regina, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

RESUMO

Pilatti, Rodrigo. **Segmentação comportamental de utilizadores de cartão de crédito utilizando o algoritmo de máquina não supervisionado *k-means.*** 2023. 39 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

As empresas estão enfrentando cada vez mais desafios para manter-se sólidas e atuantes no mercado. Com os avanços tecnológicos, os quais propiciaram maior concorrência no meio digital, é preciso fazer uso de ferramentas adequadas para oferecer aos clientes a melhor experiência possível. Nesse sentido, o agrupamento de dados torna-se um importante instrumento, especialmente porque facilita a compreensão de dados complexos e em grandes quantidades. Este estudo tenciona, portanto, implementar o algoritmo de aprendizado de máquina não supervisionado, *K-means*, para realizar o agrupamento de clientes com base nas informações comportamentais dos respectivos utilizadores. Os experimentos foram realizados em um conjunto de dados obtidos da plataforma Kaggle. Kaggle, um espaço virtual onde é possível compartilhar conjuntos de dados para análise. Para tanto, foram utilizados dados de 8.950 usuários, dos quais foi possível utilizar 18 atributos. Ao proceder à segmentação por meio do agrupamento utilizando o algoritmo K-means, foi possível obter 3 grupos. Esses grupos ofereceram informações cruciais sobre padrões de gastos distintos, uso frequente de dinheiro, comportamento de pagamento e hábitos de pagamento.

ABSTRACT

Pilatti, Rodrigo. **Behavioral segmentation of credit card users using the unsupervised machine learning algorithm k-means**. 2023. 40 p. Final course work (MBA in Artificial Intelligence and Big Data) – Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, 2023.

Companies are increasingly facing challenges to remain robust and active in the market. With technological advancements leading to greater competition in the digital realm, it is necessary to employ suitable tools to provide customers with the best possible experience. In this sense, data clustering becomes an important instrument, particularly as it facilitates the understanding of complex and large quantities of data. This study aims to implement the unsupervised machine learning algorithm, K-means, to perform the clustering of customers based on the behavioral information of respective users. Experiments were conducted on a dataset obtained from the Kaggle platform. Kaggle, a virtual space where datasets can be shared for analysis. For this purpose, data from 8,950 users were used, with 18 attributes available for use. By conducting segmentation through clustering using the K-means algorithm, it was possible to obtain 3 groups. These groups provided crucial information about distinct spending patterns, frequent cash usage, payment behavior, and payment habits.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Esta monografia investiga a aplicação de algoritmos de aprendizado de máquina não supervisionado na segmentação de clientes. Para alcançar esse objetivo, são abordadas etapas cruciais de pré-processamento, que incluem a limpeza e transformação dos dados. Posteriormente, são empregadas técnicas de agrupamento de dados. Nesta Seção 1, é apresentada a motivação a qual impulsionou o estudo.

1.1 Motivação e contextualização

Segundo Ashok *et al*. (2021) com a expansão do mercado, as empresas mais antigas se viram na necessidade de utilizar estratégias de *marketing* para se manterem competitivas. Devido ao aumento da concorrência o número de consumidores está em constante aumento, tornando desafiador para as empresas atender às necessidades de cada cliente.

O *marketing* direcionado e a divisão de clientes estão muito conectados e são usados de forma similar, o *marketing* direcionado refere-se ao agrupamento de clientes com base em certas características que as empresas pretendem servir, os clientes do mercado selecionado são segmentados em diferentes grupos com base em suas características (MONIL, 2020)

De acordo com Regmi *et al.* (2022), a técnica de segmentação de clientes, que consiste no procedimento de categorizar os clientes em grupos com base em características compartilhadas, proporciona às empresas uma maneira eficiente de direcionar suas estratégias. Essa abordagem permite que as empresas compreendam as preferências de compra dos consumidores, o que, por sua vez, pode resultar em um atendimento mais adequado e, consequentemente, em maior satisfação dos clientes.

A segmentação de clientes pode usar vários algoritmos alternativos, como algoritmos de associação, de agrupamento, de classificação ou de regressão. Dentre eles, os algoritmos de agrupamento são amplamente utilizados devido à sua precisão e eficácia na segmentação de clientes (CHORIANOPOULOS, 2016).

Neste contexto, é notável que para garantir uma experiência personalizada para cada cliente, é fundamental empregar ferramentas como a segmentação e o agrupamento de clientes. Essas

estratégias são excelentes maneiras de obter um melhor entendimento dos clientes, especialmente quando se trata com um grande volume de dados complexos.

1.2 Objetivos

Utilizar técnicas de aprendizado de máquina em um conjunto de dados para identificar grupos com características semelhantes baseado no comportamento de utilizadores de cartão de crédito, usando um conjunto de dados de clientes não rotulados de uma operadora de cartão de crédito.

1.3 Organização

A Seção 2 apresenta uma compreensão sobre segmentação de clientes, aprendizado de máquina, pré-processamento, em seguida abordamos agrupamento. Por fim é apresentado os trabalhos relacionados a este estudo. Já a Seção 3 trata do conjunto de dados que foi utilizado. Logo após, são detalhadas as etapas de pré-processamento, técnicas de agrupamento empregadas e os resultados obtidos. Por fim, são apresentadas as conclusões tiradas a partir desses resultados.

2 REVISÃO BIBLIOGRÁFICA

2.1 Considerações iniciais

Nesta Seção, é apresentado uma introdução sobre o conceito e os tipos de segmentação de clientes. Em seguida é apresentado os três principais tipos de aprendizado de máquina, passando pelas etapas de pré-processamento dos dados e, por fim, são descritos alguns métodos de agrupamento.

2.2 Segmentação de clientes

A Segmentação de clientes é uma área de pesquisa contínua e com a disponibilidade e acessibilidade cada vez maior de dados. Entender os clientes e prever seus comportamentos e padrões tornou-se necessário para as organizações empresariais (COOIL; AKSOY; KEININGHAM, 2007). A Segmentação de clientes pode ser definida como individualizar os clientes, colocar clientes individuais em determinado grupo com base em características e atributos comuns. Esta prática leva em consideração questões demográficas, psicográficas, geográficas e comportamentais (DAWANE; WAGHODEKAR; PAGARE, 2021).

Tynan e Drayton (1987) descrevem que os principais tipos de segmentação são compostos por quatro categorias fundamentais:

- Segmentação demográfica: Envolve a utilização de características como idade, gênero, etnia, religião, rendimento financeiro, estado civil para categorizar os consumidores;

- Segmentação geográfica: Baseia-se em características espaciais e de localização para dividir os clientes em grupos;

- Segmentação psicográfica: Utiliza dados relacionados à personalidade, estilo de vida e atitude dos consumidores;

- Segmentação comportamental: Utiliza características relacionadas ao comportamento de compra (ex: dinheiro gasto ou produtos adquiridos) e ao canal utilizado (ex: *online*, loja física).

Há diversas estratégias para segmentar clientes, sendo que as metodologias de aprendizado de máquina se destacam como ferramentas precisas e eficazes na obtenção de insights e padrões complexos nos dados dos clientes, algo desafiador de alcançar por meio de métodos convencionais (SINGH *et al*., 2023).

Conforme Wani *et al*. (2023), a abordagem de aprendizado de máquina não supervisionado é especialmente utilizada na segmentação de clientes, permitindo às empresas identificar padrões e agrupamentos nos dados dos clientes sem a necessidade de categorias ou rótulos predefinidos. Esses algoritmos analisam os dados dos clientes e agrupam indivíduos similares com base em suas características compartilhadas, revelando segmentos que poderiam passar despercebidos por métodos tradicionais

De maneira geral, a utilização de um modelo de agrupamento torna-se essencial para identificar segmentos, principalmente quando há a necessidade de combinar um grande volume de atributos de segmentação. Em contraste com as regras de negócio, um modelo de agrupamento tem a capacidade de processar múltiplos atributos e revelar segmentos baseados em dados que não são previamente conhecidos (ZIAFAT, 2014).

2.3 Aprendizado de máquina

Conforme Sen (2021), o aprendizado de máquina) refere-se à capacidade de um sistema adquirir, integrar e desenvolver conhecimento automaticamente a partir de dados em grande escala. Posteriormente, esse sistema é capaz de expandir o conhecimento adquirido de forma autônoma, descobrindo novas informações, sem ser especificamente programado para fazê-lo.

Segundo Alpaydyn (2014) o aprendizado de máquina envolve a execução de um programa de computador com o propósito de otimizar parâmetros, utilizando dados de exemplo ou experiências anteriores como critérios. Esse processo pode ser empregado tanto para realizar previsões futuras, proporcionando capacidades preditivas, quanto para extrair conhecimento a partir de dados, fornecendo informações descritivas

Há uma diversidade de abordagens de aprendizado de máquina, que variam de acordo com o algoritmo em uso e seus objetivos. Esses algoritmos de aprendizado de máquina podem ser categorizados em três grupos principais de acordo com as suas finalidades, que incluem aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço, como descritos a seguir (MUELLER; MASSARON, 2019).

- **Aprendizado supervisionado:** O aprendizado supervisionado utiliza algoritmos que aprendem a partir de exemplos contendo respostas-alvo, como números ou rótulos, visando predizer corretamente respostas em novos casos. Essa abordagem pode ser empregada em situações de regressão (para prever valores numéricos) e classificação (para atribuir rótulos qualitativos).

- **Aprendizado por reforço:** O aprendizado por reforço envolve exemplos sem rótulos, mas com feedback positivo ou negativo baseado nas ações do algoritmo. É aplicado em situações em que o algoritmo precisa tomar decisões com consequências, tornando-se prescritivo.

- **Aprendizado não supervisionado:** O aprendizado não supervisionado envolve algoritmos que aprendem com exemplos sem respostas predefinidas, permitindo que o algoritmo descubra padrões nos dados por conta própria.

Segundo Dutt, Chandramouli e Dos (2019), no aprendizado não supervisionado não há dados de treinamento rotulados para aprender e nenhuma previsão a ser feita, o objetivo é tomar um conjunto de dados como entrada e tentar encontrar agrupamentos ou padrões naturais dentro dos conjuntos de dados. Portanto, a aprendizagem não supervisionada é frequentemente denominada modelo descritivo e, o processo de aprendizagem não supervisionada é referido como descoberta de padrões ou descoberta de conhecimento. Uma aplicação crítica da aprendizagem não supervisionada é a segmentação de clientes.

Atualmente, muitas empresas do setor do varejo e outros segmentos estão adotando o aprendizado de máquina para alcançar seus objetivos. Eles utilizam os registros de transações dos clientes para desenvolver modelos de aprendizado de máquina direcionado para identificar o

público apropriado, isso não resulta apenas no aumento de receita e na atração de mais clientes, mas também na otimização das operações comerciais (RAO; RAO; SARADHI, 2016).

Suas áreas de aplicação são abundantes: além do varejo, as instituições financeiras utilizam seus dados anteriores para construir modelos usados na análise e aplicação de crédito, detecção de fraudes e no mercado de ações (ALPAYDIN, 2014).

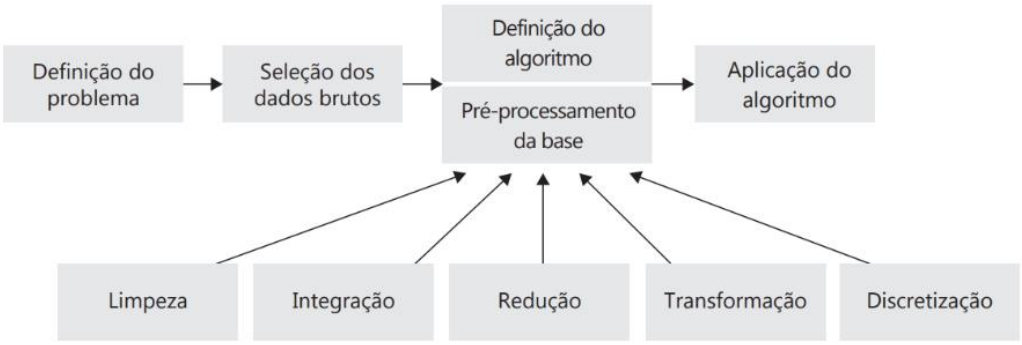
2.4 Pré-processamento

O pré-processamento de dados e a análise exploratória de dados (AED) são etapas fundamentais no aprendizado de máquina. O desempenho dos algoritmos de aprendizado de máquina frequentemente está vinculado a qualidade dos dados com que se trabalha. A etapa de pré-processamento é necessária para resolver vários tipos de problemas, incluindo dados ruidosos, redundância de dados, irrelevância de dados e valores ausentes (SAYAN *et al*., 2022).

O passo inicial envolve a identificação do problema a ser solucionado, seguido pela seleção dos dados a serem utilizados na análise. Em seguida, duas fases são parcialmente realizadas simultaneamente: a definição de um ou mais algoritmos a serem aplicados e a execução de algumas etapas de pré-processamento para preparar os dados. Importante observar que nem todas as etapas de pré-processamento são diretamente influenciadas pelo algoritmo a ser empregado; por exemplo, a redução da base de dados pode ser realizada antes ou após a aplicação do algoritmo (CASTRO, 2016).

A representação ilustrada na Figura 1 oferece uma visão abrangente do processo de preparação da base de dados para análise.

Figura 1 - Etapa de preparação da base de dado



Fonte: (CASTRO; NUNES, 2016)

As principais tarefas de pré-processamento apresentadas na Figura 1 são descritas a seguir:

- **Limpeza:** para imputação de valores ausentes, remoção de ruídos e correção de inconsistências;
- **Integração:** para unir dados de múltiplas fontes em um único local, como um armazém de dados (*data warehouse*);
- **Redução:** para reduzir a dimensão da base de dados, por exemplo, agrupando ou eliminando atributos redundantes, ou para reduzir a quantidade de objetos da base, sumarizando os dados;
- **Transformação:** para padronizar e deixar os dados em um formato passível de aplicação das diferentes técnicas de mineração;
- **Discretização:** para permitir que métodos que trabalham apenas com atributos nominais possam ser empregados a um conjunto maior de problemas. Também faz com que a quantidade de valores para um dado atributo (contínuo) seja reduzida.

Aprimorar o desempenho de algoritmos de aprendizado de máquina é viável ao empregar técnicas de pré-processamento nos dados, contribuindo para a eliminação ou redução de problemas que possam surgir (FACELI *et al*., 2021).

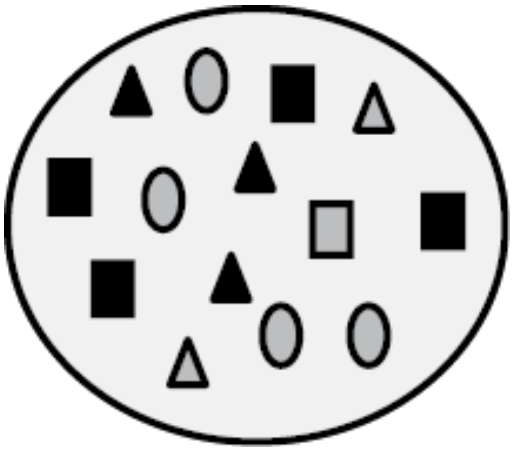
2.5 Agrupamento

Segundo Tan, Steinbach e Kumar (2013), a técnica de agrupamento consiste em agrupar objetos de dados com base apenas nas informações encontradas nos dados que descrevem os objetos e seus relacionamentos. O objetivo é que os objetos dentro de um grupo sejam semelhantes (ou relacionados) entre si e diferentes (ou não relacionados) dos objetos de outros grupos. Quanto maior a semelhança (ou homogeneidade) dentro de um grupo e quanto maior a diferença entre os grupos, melhor ou mais distinto será o agrupamento

Um desafio nas análises de agrupamento é definir o conceito de grupos homogêneos, pois diversas interpretações de homogeneidade podem resultar em agrupamentos bastante distintos.

(SICSÚ; SAMARTINI; BARTH, 2023). Na Figura 2 é ilustrado um conjunto de objetos diferenciados entre si pela forma e pela cor.

Figura 2 - Elementos a serem agrupados



Fonte: (SICSÚ; SAMARTINI; BARTH, 2023)

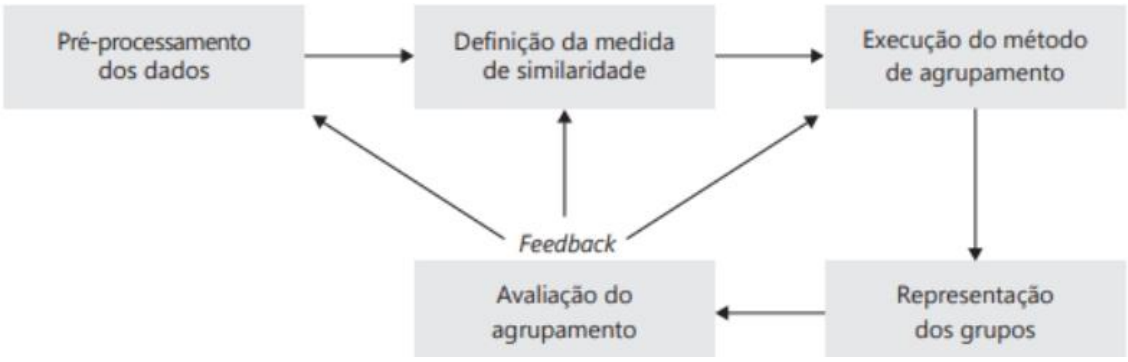
Diversas formas de agrupamento são possíveis, como por exemplo:

- Com base na geometria, surgem três clusters: retângulos, triângulos e elipses.
- A partir da cor do objeto, surgem dois clusters: cinza-claro e cinza-escuro.
- Ao considerar tanto a cor quanto a forma do objeto, são observados cinco clusters: triângulos cinza-escuros, triângulos cinza-claros, retângulos cinza-escuros, retângulos cinza-claros e elipses.
- Elipses e outras formas levam à formação de dois clusters, incluindo elipses, retângulos, triângulos, etc.

Agrupamento é um dos métodos mais comuns usados na exploração de dados para obter uma compreensão clara da estrutura de dados. Dados semelhantes são agrupados em muitos subgrupos. A técnica de agrupamento é comumente usada para segmentar clientes com base em seus comportamentos e transações (ISHANTHA, 2021).

Castro (2016), fornece as cinco principais etapas que pode ser divido o processo de agrupamento como ilustrado na Figura 3. As etapas iniciais podem ser ajustadas, visando aprimorar o resultado do agrupamento, utilizando o *feedback* gerado durante o processo de agrupamento.

Figura 3 - Processo de agrupamento de dados



Fonte: (CASTRO; NUNES, 2016)

Existem muitos métodos de agrupamento na literatura. Esses métodos podem ser amplamente categorizados em métodos de particionamento, métodos hierárquicos e métodos baseados em densidade. Os métodos de particionamento usam uma métrica baseada em distância para agrupar os pontos com base em sua similaridade (KAUFMAN; ROUSSEEUW, 2013).

2.5.1 Algoritmos de agrupamento

Os algoritmos de agrupamento funcionam de modo a analisar um conjunto de dados em busca de padrões, eles conseguem identificar semelhanças nos dados e tomar decisões com base na presença ou ausência dessas semelhanças em novos dados. Esses algoritmos são treinados com conjuntos de dados de teste que não possuem rótulos, categorias ou classificações predefinidas (NARAYANA *et al*. 2022).

Os dois algoritmos de agrupamento mais amplamente utilizados são algoritmos do tipo hierárquico e particional. Esses algoritmos têm sido muito utilizados em uma ampla gama de aplicações, principalmente devido à sua simplicidade e facilidade de implementação em relação a outros algoritmos de agrupamento (KANSAL *et al*., 2018).

Os algoritmos de agrupamento particional visam descobrir os agrupamentos presentes nos dados, otimizando uma função objetiva específica e melhorando iterativamente a qualidade das

partições. O algoritmo de agrupamento particional que será abordado é o algoritmo de clusterização *K-means*, um dos algoritmos de agrupamento mais simples e eficientes propostos na literatura de agrupamento de dados, nesta seção descrevemos esse algoritmo (NARAYANA *et al*. 2022).

2.5.2 Algoritmos K-médias

De acordo com Tan, Steinback e Kumar (2013) K-médias é uma técnica de agrupamento baseada em protótipo que busca encontrar um número de clusters (K) especificado pelo usuário, representados por seus centróides. Dentre as várias técnicas existentes, duas das mais proeminentes são o *K-means* e o *K-medoid*.

2.5.2.1 Algoritmo *K-means*

O algoritmo *K-means* é um método de análise de cluster não hierárquico que particiona objetos em um ou mais grupos com base na similaridade de suas características para que objetos que possuem características mais próximas sejam agrupados em um mesmo cluster, enquanto objetos que possuem características diferentes são agrupados em outros clusters (GARBADE, 2018).

Conforme Gan, Ma e Wu (2007), o algoritmo *K-means* pode ser dividido em duas fases: a fase de inicialização e a fase de iteração. Na fase de inicialização, o algoritmo atribui aleatoriamente os casos em k clusters. Na fase de iteração, o algoritmo calcula a distância entre cada caso e cada cluster e atribui o caso ao cluster mais próximo.

O pseudocódigo do algoritmo básico do *K-means* é descrito na Tabela1.

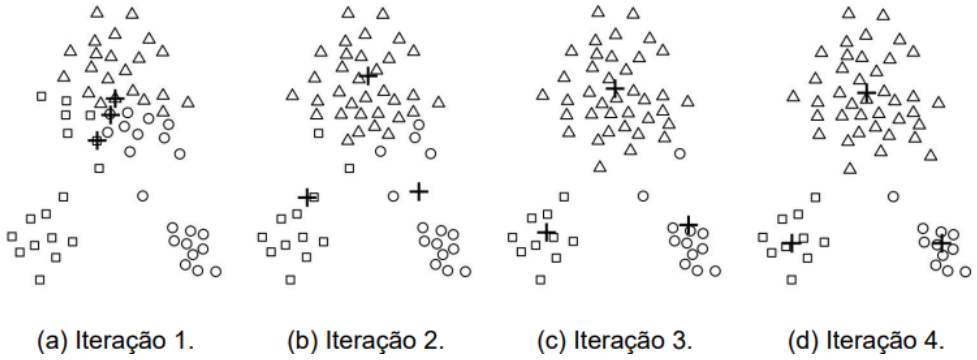
Tabela 1 – pseudocódigo do algoritmo básico do K-means

Algoritmo 1 – Algoritmo básico de *K-means*
1: Selecionar K pontos como centróides iniciais.
2: **repetir**
3: Formar K agrupamentos, atribuindo cada ponto ao seu centróide mais próximo.
4: Recalcular o centróide de cada cluster.
5: **até que** os centróides não se alterem

Fonte: Adaptada de Tan, Steinbach e Kumar (2013)

A operação de *K-means* é ilustrada na Figura 4, que mostra como, a partir de três centróides, os clusters finais são encontrados em quatro etapas de atualização de atribuição.

Figura 4 - Algoritmo K-means para encontrar três clusters em dados de amostra



Fonte: Adaptada de Tan, Steinbach e Kumar (2013)

Nas etapas 2, 3 e 4, ilustradas nas Figuras 4(b), (c) e (d), respectivamente, dois dos centróides se deslocam para os dois pequenos grupos de pontos na parte inferior das figuras. Ao concluir na Figura 4(d), o algoritmo *K-means* identifica os agrupamentos naturais dos pontos, uma vez que não há mais alterações.

Para algumas combinações de funções de proximidade e tipos de centróides, o *K-means* sempre converge para uma solução; ou seja, o *K-means* alcança um estado no qual nenhum ponto está mudando de um cluster para outro, isso resulta na estabilização dos centróides, que deixam de sofrer alterações (TAN; STEINBACH; KUMAR, 2013).

O algoritmo *K-means* se destaca por sua abordagem simplificada na identificação de clusters, exigindo menos complexidade de termos estatísticos (DUTT; CHANDRAMOULI, DOS 2019). No entando um dos principais problemas do método *K-means* é como determinar o número ideal de clusters, conforme (GUSTRIANSYAH; SUADI; ANTONY, 2023), existem vários métodos que podem ser empregados para estimar o número ideal de clusters (K), entre eles estão o Método Elbow (cotovelo), Índice Silhouette, Índice Davies-Bouldin.

2.5.3 Método do cotovelo

Determinar o número ideal de agrupamentos é crucial ao aplicar o algoritmo *K-means* um valor de K muito baixo pode levar a agrupamentos amplos e pouco distintivos, enquanto um K muito alto pode resultar em sobre ajuste, gerando agrupamentos excessivamente específicos. (Kansal *et al*., 2018). Portanto, é essencial utilizar técnicas adequadas, como o método do cotovelo ou o método da silhueta, para identificar o valor ideal de K.

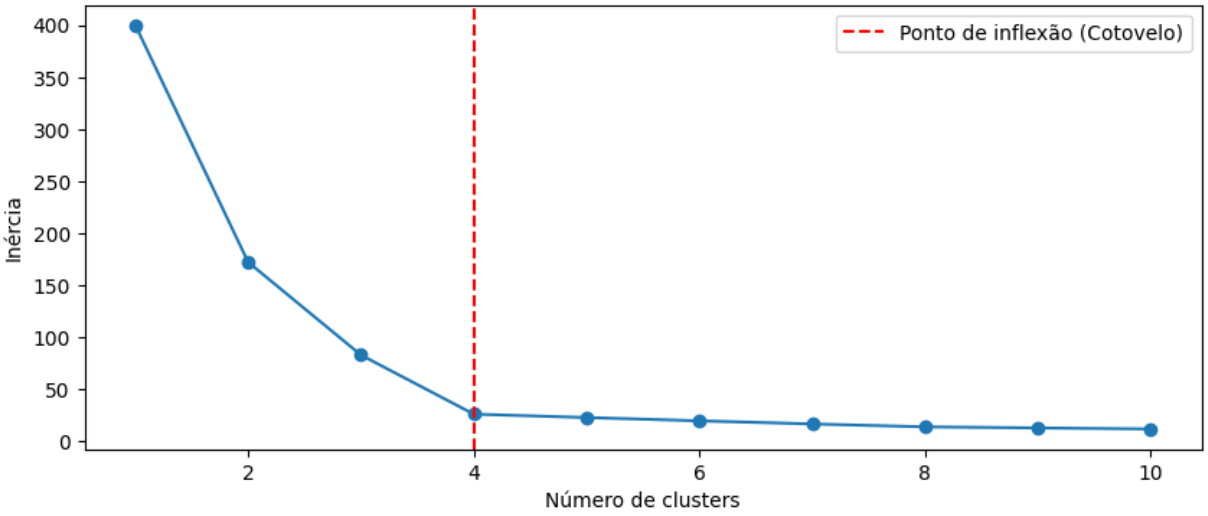
O método do cotovelo, descrito por Bholowalia e Kumar (2014), avalia a variação dos dados conforme o número de agrupamentos. Inicialmente, à medida que o número de agrupamentos aumenta, a variação dentro dos grupos diminui rapidamente. Contudo, após um certo número de agrupamentos, essa redução torna-se menos significativa, formando um ponto de inflexão no gráfico. Esse ponto, conhecido como critério do cotovelo, representa o número ideal de agrupamentos. Para um entendimento mais compreensível quanto maior o número de clusters k, o valor do WCSS será menor, ou vice-versa. A fórmula matematicamente para o cálculo do WCSS (*Within-Cluster Sum of Squares)* é representada na Equação 01.

$$\text{WCSS} = \sum_{j}^{k} = 1 \sum_{j}^{n} = 1 \|x_i^{(j)} - c_j\|^2 \quad (01)$$

onde k = o número de clusters, n = o número de objetos, $x_i = i^{th}$ elemento no cluster e c_j = o centróide do J^{th} cluster.

A abordagem é iniciar com K = 2 e aumentar gradualmente, calculando a soma dos quadrados das distâncias, em determinado valor de K, a soma dos quadrados das distâncias diminuirá significativamente e, em seguida, atingirá um patamar onde, mesmo com o aumento de K, a soma dos quadrados das distâncias permanecerá praticamente inalterada. Esse é considerado o valor ótimo de K, uma vez que incluir mais agrupamentos além desse ponto resultaria em grupos muito semelhantes entre si, o que não é desejável na análise de agrupamentos. A Figura 5 exemplifica uma curva gerada pelo método do cotovelo, mostrando claramente que K = 4 representa o ponto de "cotovelo", onde adicionar mais agrupamentos não melhora significativamente a variação dentro dos grupos.

Figura 5 - Método do cotovelo



Fonte: Adaptada de Griva *et al*. (2018).

O gráfico, ilustrado na Figura 5, gerado pelo método do cotovelo apresenta a linha tracejada em vermelho, a qual indica o número ideal de clusters. A partir desse ponto, a redução na soma dos quadrados das distâncias dentro dos clusters (inércia) diminui, sugerindo que a inclusão de mais clusters não resultará em melhorias significativas.

2.5.4 Avaliação de agrupamento

Os índices de avaliação, são responsáveis por aferir quantitativamente o agrupamento resultante de um algoritmo para uma base de dados. Existem dois tipos de índices que podem ser utilizados para avaliar os grupos formados pelos elementos de uma base de dados (CASTRO, 2016).

- **Internos:** são índices que utilizam apenas informações inerentes aos objetos do agrupamento, baseando-se em medidas de similaridade e avaliando as distâncias intragrupos e/ou intergrupos, ou, seja esses índices não recorrem a informações externas para avaliar a formação dos grupos.

- **Externos:** são índices que avaliam quão correto está um agrupamento dado um agrupamento ideal que se deseja alcançar. O cálculo dessas medidas requer o

conhecimento prévio do grupo ao qual cada objeto pertence.

Para este trabalho específico de aprendizado não supervisionado, será abordado o método de índice interno devido à ausência de rótulos prévios no conjunto de dados. Aqui são apresentados os índices de avaliação interna: o Índice de Silhueta e o Índice de Davies-Bouldin.

Índice de silhueta

Segundo Gustriansyah, Suhandi e Antony (2020). O Índice de Silhueta é uma medida utilizada para avaliar a qualidade do agrupamento em pontos específicos. Em sua abordagem, Rousseuw (1987) propõe um método que calcula o valor máximo do índice. A noção de silhueta refere-se à interpretação e validação da consistência nos agrupamentos de dados. Essas funções podem ser calculadas por meio da Equação 02.

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b\{(o)\}} \quad (02)$$

Onde:

- $s(o)$ = Representa o coeficiente de silhueta para um ponto específico o;
- $a(o)$ = Distância média entre o e todos os demais pontos do cluster do qual o faz parte;
- $b(o)$ = Indica a distância média mínima de o a todos os clusters dos quais não faz parte.

Conforme Faceli *et al* (2021), o índice de silhueta pode ser empregado na avaliação de uma partição; para isso, avalia também a adequação de cada objeto ao seu cluster e a qualidade de cada cluster individualmente. O valor de uma medida silhueta é limitado pelo intervalo [–1, 1], em que a melhor partição de acordo com a silhueta é aquela com valor igual a 1.

Índice Davies-Bouldin

O Índice Davies-Bouldin é um método usado para avaliar a validade de agrupamentos em métodos de clusterização. Dentro dessa abordagem, a coesão é definida como a soma da

proximidade dos dados em relação ao centro do cluster, enquanto a separação se baseia na distância entre os pontos centrais dos clusters. (GUSTRIANSYAH; SUHANDI; ANTONY, 2020).

Se a distância entre os clusters for máxima, isso significa que a similaridade característica entre cada cluster é pequena, permitindo que as diferenças entre os clusters sejam vistas com mais clareza. Se a distância dentro de um cluster for mínima, isso significa que cada objeto no cluster possui um alto nível de similaridade característica. Na Equação 03 é apresentado o cálculo do índice Davies-Bouldin.

$$DB(j) = \frac{1}{K}\sum_{i=1}^{K} \underset{i \neq j}{max}(Rij) \tag{03}$$

Na equação 3, K representa o número de agrupamentos e R_{ij} é a medida de similaridade entre os agrupamentos C_i e C_j. Para calcular essa medida, a Equação 04 é utilizada.

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{04}$$

Na equação 4, para cada agrupamento i e j, s_i e s_j são as distâncias médias entre cada ponto e seu respectivo centróide e, d_j é a distância entre os centróides dos agrupamentos.

O critério de avaliação do índice é: quanto mais baixo for o valor do índice Davies-Bouldin, mais ideal é o número de clusters alcançado.

2.6 Trabalhos relacionados

Nesta seção, é fornecida uma síntese dos estudos acadêmicos relevantes ao tema abordado neste trabalho.

2.6.1 Trabalhos relacionados

Rajput e Singh (2023) propuseram um estudo sobre a segmentação de clientes em dados de comércio eletrônico usando o algoritmo de agrupamento K-means. A proposta possivelmente

envolveu a aplicação do algoritmo de K-means para dividir os clientes em grupos ou segmentos com base em padrões de comportamento, preferências de compra, ou outras características identificáveis nos dados do comércio eletrônico.

Griva *et al*. 2018 (2018) propuseram um estudo sobre varejo e análise de negócios, no qual abordaram técnica de clustering para realizar segmentação e caracterização. De acordo com os autores, "A segmentação das visitas dos clientes é uma parte essencial da análise de negócios no setor de varejo" (Griva et al., 2018).

Kansal *et al*. (2018) os autores implementaram três algoritmos de clustering (k-Means, Agglomerative e Meanshift) para segmentar os clientes. Eles usaram um programa Python que foi treinado em um conjunto de dados contendo duas características: a média da quantidade de compras dos clientes e a média da visita anual do cliente à loja. O resultado da segmentação gerou cinco segmentos de clientes, que foram rotulados como "Descuidados", "Cuidadosos", "Padrão", "Alvo" e "Sensíveis".

Tavakoli *et al*. (2018) realizaram um estudo sobre segmentação de clientes e desenvolvimento de estratégias com foco na análise do comportamento do usuário, utilizando o modelo RFM (Recency, Frequency, Monetary) e técnicas de mineração de dados. Os resultados obtidos a partir da campanha demonstraram que o modelo de Segmentação proposto melhorou o número de compras realizadas pelos clientes, bem como o valor médio das compras efetuadas.

Monil (2020) realizou um estudo relacionado à segmentação de clientes utilizando técnicas de aprendizado de máquina. com o objetivo de identificar e agrupar clientes com base em características específicas para desenvolver estratégias de marketing mais direcionadas e personalizadas.

2.7 Considerações finais

Nesta seção, foram discutidos alguns conceitos fundamentais utilizados neste trabalho. Além disso, foram apresentados estudos relevantes para a prática de segmentação de clientes. Na seção seguinte será apresentada a proposta para elaboração deste trabalho.

3 METODOLOGIA

Nesta seção é descrita a metodologia adotada para conduzir as experimentações e são apresentados os resultados alcançados. Os experimentos foram realizados em um conjunto de dados obtidos da plataforma Kaggle. Kaggle é um espaço virtual público onde cientistas de dados e especialistas em aprendizado de máquina podem interagir para explorar, descobrir e compartilhar conjuntos de dados para análise.

3.1 Conjunto de dados

O conjunto de dados utilizado contém informações transacionais sobre usuários de cartão de crédito. Está é uma tentativa de agrupar clientes identificando semelhanças dos usuários por meio do algoritmo de aprendizado de máquina não supervisionado *K-means*.

O conjunto de dados contém o comportamento de 8.950 usuários de cartão de crédito durante um período de seis meses, incluindo 18 atributos que são as características comportamentais dos usuários, a descrição dos atributos é ilustrada na Tabela 2.

Tabela 2 – Descrição dos atributos

Atributos	Descrição
cust_id	Identificação do titular do cartão de crédito
balance	Saldo disponível para compras
balance_frequency	Frequência de atualização do saldo (entre 0 e 1)
purchases	Quantidade de compras realizadas
oneoff_purchases	Quantidade de compras feitas à vista
installments_purchases	Quantidade de compras em parcelas
cash_advance	Dinheiro adiantado
purchases_frequency	Frequência de compras (entre 0 e 1)
oneoff_purchases_frequency	Frequência de compras à vista (entre 0 e 1)
purchases_installments_frequency	Frequência de compras parceladas (entre 0 e 1)
cash_advance_frequency	Frequência de saques de dinheiro adiantado
cash_advance_trx	Total de transações de adiantamento em dinheiro
purchases_trx	Total de transações de compra
credit_limit	Limite do cartão de crédito
payments	Valor total pago pelo usuário
minimum_payments	Valor mínimo do pagamento
prc_full_payment	Percentual de pagamentos da fatura completa
tenure	Tempo de posse do cartão de crédito pelo usuário

Fonte: O autor

A etapa de pré-processamento é necessária para resolver vários tipos de problemas, incluindo dados ruidosos, redundância de dados, irrelevância de dados e valores ausentes (SAYAN et al., 2022). Este conjunto de dados consiste em 18 atributos e 8.950 registros. Todos os atributos são numéricos, exceto o identificador do cliente (cust_id). Após analisar a Tabela 2 é possível concluir que o atributo (cust_id) é do tipo categórico e apenas representa a identificação do cliente. Como esse atributo não contribui para o agrupamento, optou-se por remover esta coluna. No conjunto de dados não foi identificado valores duplicados; no entanto, foi identificada a presença de valores nulos em duas variáveis: *credit_limit*, com 1 valor nulo; e *minimum_payments*, com 313 valores nulos. Em vez de remover os dados ausentes, uma alternativa viável, conforme sugerido por Sicsú Samartini e Barth (2023), é preenchê-los por meio da imputação de um valor estimado. No caso de variáveis numéricas (ou quantitativas), a estratégia utiliza comum foi a de substituir os valores faltantes pela média dos valores presentes na coluna correspondente.

3.2 Normalização dos dados

A normalização dos dados consiste em escalonar os valores de cada atributo de forma que estes sejam convertidos para valores menores e específicos, como de –1 a 1, ou de 0 a 1. A normalização faz-se necessária para prevenir que alguns atributos, por apresentarem uma escala de valores maior que outros, influenciem de forma tendenciosa quando aplicado em algoritmo de aprendizado de máquina (GOLDSCHMIDT, 2015).

Para este trabalho fez-se o uso da normalização por desvio padrão, também conhecida por *Z-Score* uma técnica comumente utilizada para padronização de dados, esta técnica ajusta os valores de cada atributo do conjunto de dados para que tenha a média 0 e desvio padrão igual a 1 (GOLDSCHMIDT, 2015).

A Equação 05, de normalização, é expressa da seguinte forma:

$$A' = \frac{A - \mu}{\sigma} \tag{05}$$

Onde: *A'* representa o valor resultante após a normalização, *A* refere-se ao valor original do elemento, μ é a média aritmética simples dos valores do atributo e σ é o desvio padrão dos valores do atributo.

Para determinar a necessidade de normalizar os dados do conjunto, foi realizada uma análise por meio de estatísticas descritivas, utilizando o método (*describe*). É possível observar na Tabela 3 que vários atributos, como *balance,* variam em uma ampla faixa, enquanto outros, como o *balance_frequency*, variam entre 0 e 1. Conforme apontado por (SICSÚ; SAMARTINI; BARTH, 2023), atributos com distribuição muito assimétrica podem prejudicar ou influenciar negativamente o modelo.

Tabela 3 – Estatísticas descritivas dos atributos

Atributos	Count	Mean	Std	Min	25%	50%	75%	Max
balance	8950	1564.47	2081.53	0.00	128.28	873.39	2054.14	19043.14
balance_frequency	8950	0.877	0.237	0.000	0.889	1.000	1.000	1.000
purchases	8950	1003.20	2136.63	0.00	39.64	361.28	1110.13	49039.57
minimum_payments	8950	864.21	2372.45	0.00	169.12	312.34	825.49	76406.21
prc_full_payment	8950	0.154	0.292	0.00	0.000	0.000	0.143	1.000
tenure	8950	11.51	1.34	6	12	12	12	12

Fonte: O autor.

Para solucionar essa questão, foi implementado a técnica de normalização por desvio padrão, após essa etapa, ao aplicar o método *describe* novamente, foi obtido os resultados apresentados na Tabela 3. Agora todos os atributos possuem a média 0 e desvio padrão 1.

Tabela 4 - Estatísticas descritivas normalizadas

Atributos	Count	Mean	Std	Min	25%	50%	75%	Max
balance	8950	-0.0	1.0	-0.8	-0.7	-0.3	0.2	8.4
balance_frequency	8950	0.0	1.0	-3.7	0.0	0.5	0.5	0.5
purchases	8950	0.0	1.0	-0.5	-0.5	-0.3	0.1	22.5
minimum_payments	8950	0.0	1.0	-0.4	-0.3	-0.2	0.0	32.4
prc_full_payment	8950	-0.0	1.0	-0.5	-0.5	-0.5	-0.0	2.9
tenure	8950	0	1	-4.1	0.4	0.4	0.4	0.4

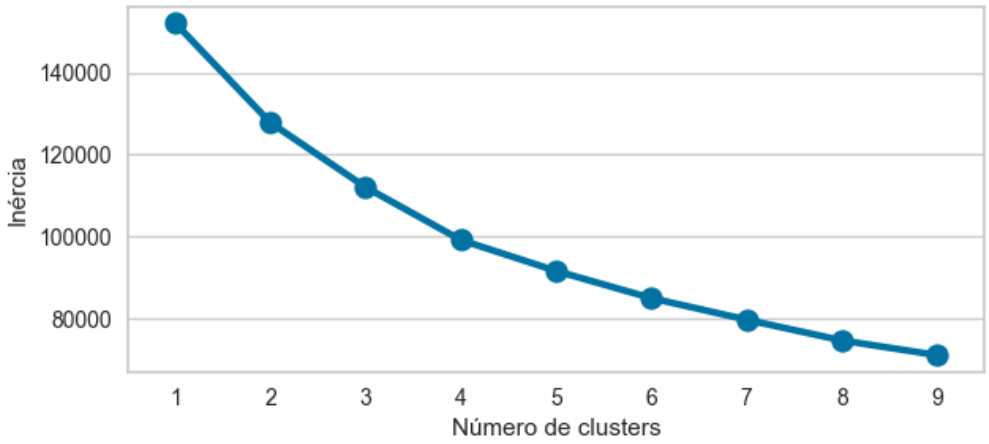
Fonte: O autor.

Após a normalização dos valores, o algoritmo de agrupamento de dados foi aplicado para categorizar os clientes de acordo com suas similaridades.

3.3 Agrupamento com *K-means*

Para implementar o algoritmo *K-means*, o primeiro passo é determinar o número ideal de clusters (representado por K) que o algoritmo utilizará para agrupar os dados. Uma abordagem comumente empregada para encontrar esse valor ótimo de K é o método do cotovelo. Esse método analisa a variação dos dados em relação ao número de clusters. Inicialmente, à medida que aumentamos o número de clusters, a variação interna diminui consideravelmente. Contudo, chega um ponto em que essa redução se torna menos acentuada, formando um ponto de inflexão no gráfico ilustrado na Figura 6, o qual se assemelha a um "cotovelo". Esse ponto de inflexão, denominado "método do cotovelo", indica o número ideal de clusters. O gráfico ilustrado na Figura 6 foi gerado usando o método do cotovelo, neste gráfico o ponto onde ocorre a mudança da curvatura indica o número ideal de clusters.

Figura 6 - Método do cotovelo aplicado a conjunto de dados dos usuários de cartão de crédito



Fonte: O autor

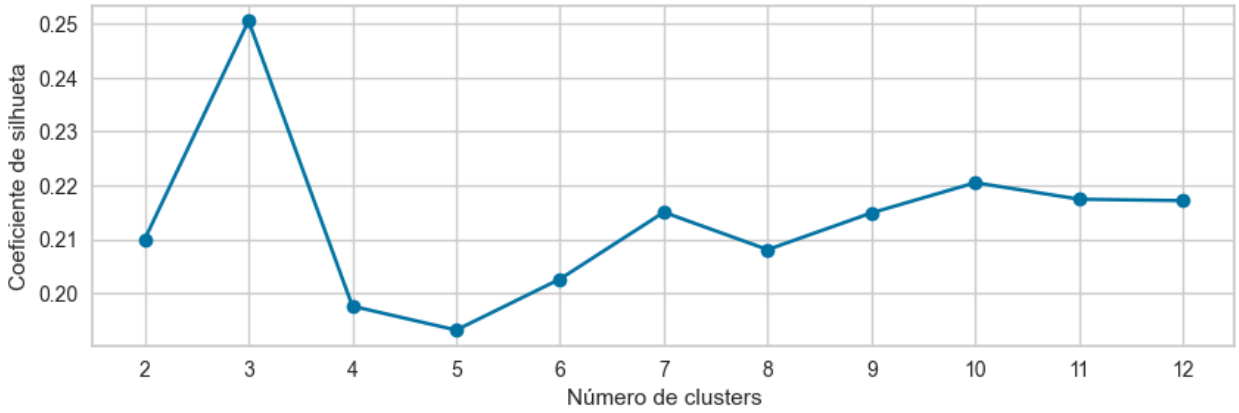
Com base na análise do gráfico ilustrado na Figura 6, observa-se que o número ideal de clusters está entre 3 e 4. Esse entendimento foi alcançado ao analisar o ponto em que a inércia começa a se estabilizar, resultando em uma leve inclinação na curva. Isso indica que a adição de mais clusters não geraria uma mudança significativa na inércia. Essa observação sugere que o

algoritmo *K-means*, operando com 3 ou 4 clusters, é capaz de oferecer uma boa separação, dos dados.

Outros métodos podem ser utilizados para encontrar o número ideal de clusters. Foram analisados os índices de silhueta e Davies-Bouldin. Os gráficos ilustrados nas Figuras 7 e 8 foram gerados por meio das funções *silhouette_score* e *davies_bouldin_score* presentes na biblioteca sklearn.

Certamente, ao determinar o número ideal de agrupamentos, é fundamental destacar as métricas utilizadas e os valores obtidos associados a essas métricas. No método da silhueta, o valor é limitado ao intervalo [–1, 1], onde uma pontuação mais próxima de 1 indica uma melhor estruturação dos clusters, representando uma boa separação entre eles.

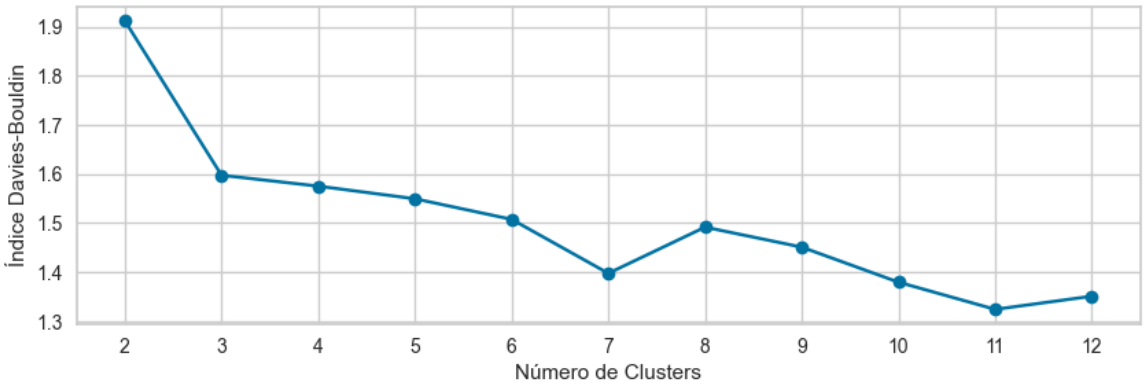
Figura 7 - Coeficiente de silhueta aplicado ao conjunto de dados dos usuários de cartão de crédito



Fonte: o Autor

Por outro lado, o método de Davies-Bouldin, é um método empregado para avaliar a qualidade dos agrupamentos em técnicas de clusterização. Essa métrica considera a coesão dentro dos clusters e a separação entre eles para determinar a eficácia do agrupamento. Ao seguir o critério de avaliação desse índice, é importante observar que quanto menor for o valor obtido, melhor será a configuração de clusters alcançada, indicando uma estrutura mais definida e distintiva entre os grupos formados.

Figura 8 - Método de Davies-Bouldin aplicado conjunto de dados dos usuários de cartão de crédito



Fonte: O autor

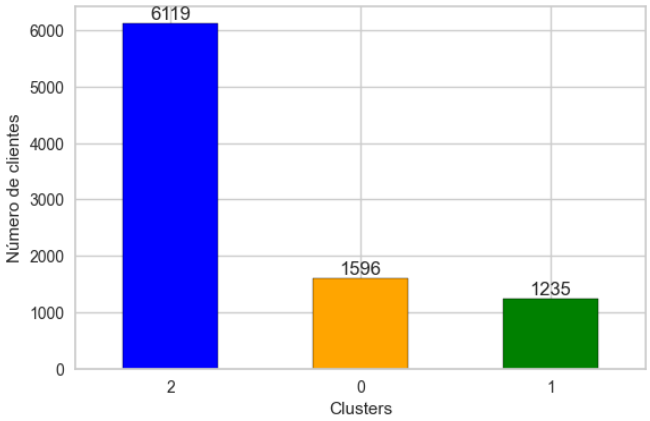
Após realizar a análise dos resultados obtidos, concluiu-se que o valor ideal de *K* é 3.

3.4 Resultados e Discussões do agrupamento *K-means*

Para realizar a segmentação dos clientes foi implementa o algoritmo de agrupamento *K-means* nos dados, e foram observadas métricas de avaliação essenciais para entender a estrutura dos clusters. O índice de silhueta foi utilizado para escolher o melhor número de clusters a serem gerados. De modo análogo, o índice de Davies-Bouldin diminuiu gradualmente à medida que o número de clusters aumentou, sugerindo uma boa separação e coesão entre os clusters nessa configuração.

Com a implementação do algoritmo *K-means* foi possível agrupar os clientes em três grupos. Na Figura 9 é ilustrado o gráfico que representa a quantidade de clientes atribuídos a cada grupo após a execução do algoritmo de agrupamento *K-means*.

Figura 9 - Quantidade de clientes atribuídos a cada grupo



Fonte: O autor

Na Tabela 4 é apresentado os valores médios dos atributos para cada um dos três clusters resultantes do processo de agrupamento. Cada linha representa um atributo específico e cada coluna representa um cluster.

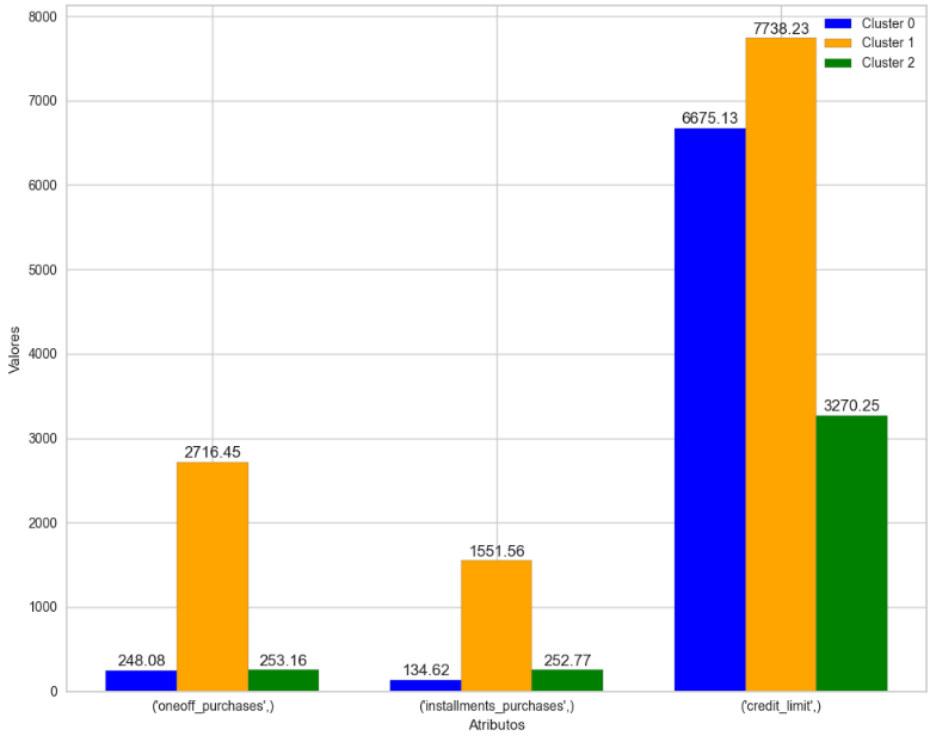
Tabela 4 – Valores médios dos atributos

Atributos	Clusters		
	0	**1**	**2**
balance	3981.96	2228.03	799.77
balance_frequency	0.96	0.98	0.84
purchases	382.62	4267.53	505.62
oneoff_purchases	248.08	2716.45	253.16
installments_purchases	134.62	1551.56	252.77
cash_advance	3868.30	460.63	329.81
purchases_frequency	0.23	0.95	0.47
oneoff_purchases_frequency	0.11	0.67	0.13
purchases_installments_frequency	0.14	0.74	0.35
cash_advance_frequency	0.45	0.06	0.07
cash_advance_trx	12.39	1.54	1.21
purchases_trx	5.57	56.48	8.66
credit_limit	6675.13	7738.23	3270.25
payments	3016.46	4151.34	909.83
minimum_payments	1793.70	1226.21	548.59
prc_full_payment	0.03	0.30	0.16
tenure	11.35	11.92	11.48

Fonte: O Autor

O Cluster 1 parece ter os maiores valores médios em várias categorias, como *oneoff_purchases* (quantidade de compras feitas à vista), *installments_purchases* (quantidade de compras em parcelas) e *credit_limit* (limite do cartão de crédito). Isso pode indicar um grupo de clientes que faz compras frequentes, tanto parceladas quanto à vista, e possui limites de crédito mais elevados em comparação aos clientes que se encontram nos outros dois clusters, conforme ilustrado na Figura 10.

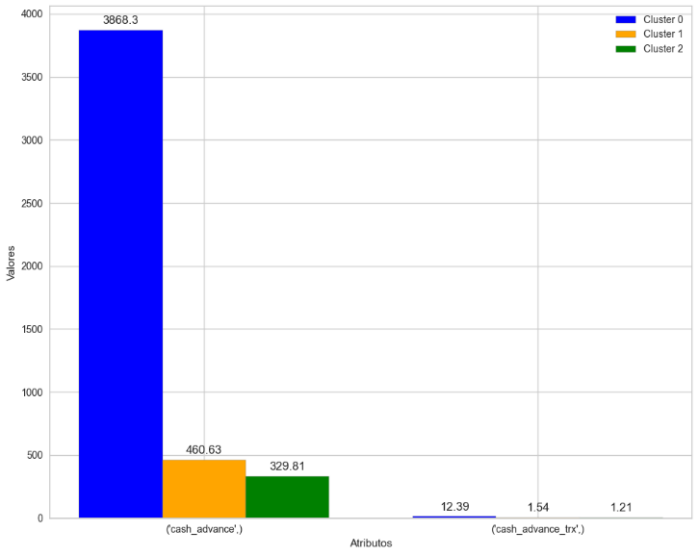
Figura 10 - Divisão dos Clientes em Clusters por Padrões de Gastos



Fonte: O Autor

O Cluster 0 (uso frequente de dinheiro) exibe um valor muito mais alto em *cash_advance* (dinheiro adiantado) e *cash_advance_trx* (total de transações de adiantamento em dinheiro), em comparação com os outros clusters. Isso sugere que os clientes deste grupo utilizam com mais frequência o recurso de adiantamento de dinheiro, o que pode indicar maior dependência ou necessidade de crédito emergencial, como ilustrado na Figura 11.

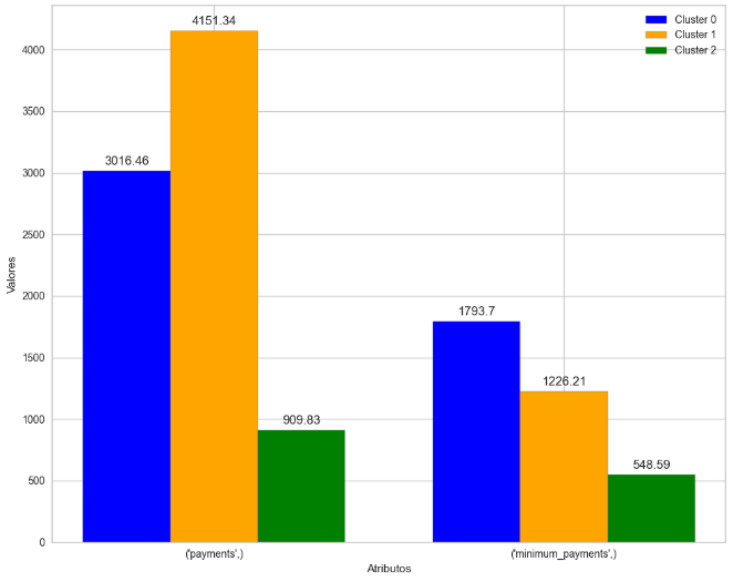
Figura 11 - Divisão dos Clientes em Clusters por Uso Frequente de Dinheiro



Fonte: O Autor

O Cluster 2 possui valores mais baixos em *payments* (valor total pago pelo usuário) e e*m* (Valor mínimo do pagamento) *minimum_payments.* Isso sugere que esses clientes, em média, pagam menos do que os outros grupos ou possuem saldos menores em relação aos valores mínimos de pagamento, conforme Figura 12.

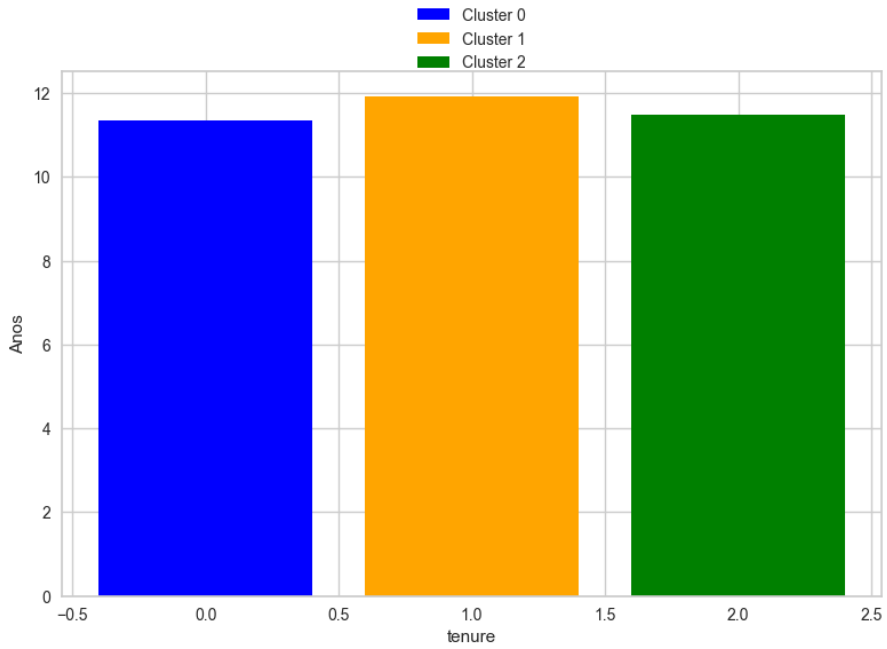
Figura 12 - Divisão dos Clientes em Clusters por Comportamento de Pagamento



Fonte: O Autor

Todos os clusters parecem ter uma média semelhante em *tenure* (tempo de posse do cartão de crédito pelo usuário, sugerindo que, em média, os clientes mantêm um relacionamento estável e de longo prazo com a instituição financeira, independentemente do cluster a que pertençam, conforme ilustrado na Figura 14, onde é levado em consideração a estabilidade do relacionamento com a instituição financeira.

Figura 13 - Figura 14 - Tempo de relacionamento do cliente



Fonte: O Autor

3.5 Considerações finais

Esta seção apresentou os resultados e a discussão da aplicação do *K-means* para a segmentação de clientes frente ao tipo de gastos e de compras realizadas. No próximo capítulo são apresentadas as conclusões.

4 CONCLUSÃO

Por meio deste estudo foi possível verificar diversos tipos de comportamentos de clientes quanto ao uso do cartão de crédito, valendo-se do algoritmo *K-means*.

Para tanto, foram utilizados 18 atributos disponíveis em uma base de dados. Na prática, verificou-se que é possível desenvolver segmentações de mercado de forma eficaz, as quais possibilitam uma melhor forma de analisar e compreender o modo de agir dos clientes que fazem uso do cartão de crédito.

Nesse sentido, a presente pesquisa demonstrou ser possível o agrupamento de clientes em 3 principais grupos, cluster 0, cluster 1 e cluster 2. O agrupamento tornou viável descobrir características dos 3 principais grupos. O número de clusters foi obtido por meio do método do cotovelo, índice de silhueta e índice de Davies-Bouldin. Isso foi realizado com o propósito de encontrar o número ideal de clusters para obter um resultado mais preciso na tarefa de agrupamento e descoberta do número ideal de clusters a fim de se alcançar um resultado mais preciso.

5 REFERÊNCIAS

ALPAYDIN, E. **Introduction to machine learning**. Cambridge, Massachusetts: The Mit Press, 2014.

ASHOK, V.; KAMATH, R.; RK, A.; SINGH, S.; BHATI, A. **Customer Segmentation in E-Commerce** (2021).

BOYD, STEPHEN.; VANDENBERGHE, LIEVEN. **Convex Optimization**. Cambridge: Cambridge University Press, 2004.

CASTRO, D.G.F.; NUNES, L. **Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações**. Editora Saraiva, 2016.

CHORIANOPOULOS, A. **Effective CRM using predictive analytics**. John Wiley & Sons, 2016.
COOIL, B.; AKSOY, L.; KEININGHAM, T. **Approaches to Customer Segmentation.** Journal of Relationship Marketing, v. 6, p. 9-39, 2007. DOI: 10.1300/J366v06n03_02.

DAWANE, V.; WAGHODEKAR, P.; PAGARE, J. **RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention**. SSRN Electronic Journal, 2021.

DUTT, S; CHANDRAMOULI S..; DOS, A. **Machine Learning, 1ª ed., Pearson**, 2019.

FACELI, K.; LORENA, A. C.; GAMA, J.; ALMEIDA, T. A.; CARVA, A. C. P. L. F. **Inteligência Artificial: Uma abordagem de aprendizado de máquina**. [S.l.]: Grupo Gen-LTC, 2021.

GAN, G.; MA.; C.; WU, J. **Data Clustering: Theory, Algorithms, and Applications**. Philadelphia: Society for Industrial and Applied Mathematics, 2007. DOI: 10.1137/1.9780898718348

GARBADE, M. J. **Understanding K-means Clustering in Machine Learning**. Towards Data Science, 2018. Disponível em: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

GOLDSCHMIDT, R. **Data Mining**. Grupo GEN, 2015. E-book. ISBN 9788595156395.

GRIVA, A. *et al*. **Retail business analytics: Customer visit segmentation using market basket data**. Expert Systems with Applications, v. 100, p. 1-16, 2018. ISSN 0957-4174. Disponível em: https://doi.org/10.1016/j.eswa.2018.01.029.

GUSTRIANSYAH, R.; SUHANDI, N.; ANTONY, F. (2020). **Clustering optimization in RFM analysis Based on k-Means**. Indonesian Journal of Electrical Engineering and Computer Science, 18, 470-477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

GUSTRIANSYAH, R.; SUHANDI, N.; ANTONY, F. **Clustering optimization in RFM analysis Based on k-Means**. Indonesian Journal of Electrical Engineering and Computer Science, v. 18, n. 1, p. 470, 1 abr. 2020.

ISHANTHA, Asith. **Mall customer segmentation using clustering algorithm**. 2021.

KAUFMAN, L.; ROUSSEEUW, P. J. **Data Clustering: Algorithms and Applications**. 1ª edição. [S.I]: Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2013.

MONIL, P. **Customer Segmentation using Machine Learning**. International Journal for Research in Applied Science and Engineering Technology, v. 8, n. 6, p. 2104–2108, 30 jun. 2020.
MUELLER, JOHN P.; MASSARON, LUCA. **Aprendizado de Máquina Para Leigos**. [S.I]: Editora Alta Books, 2019. E-book. ISBN 9788550809250.

NARAYANA, V. L. et al. **Mall Customer Segmentation Using Machine Learning**. International Conference on Electronics and Renewable Systems (ICEARS), 2022, Tuticorin, India. Proceedings... Tuticorin: [S.I], 2022. p. 1280-1288. DOI: 10.1109/ICEARS53579.2022.9752447.
RAJPUT, L.; SINGH, S. N. Customer Segmentation of E-commerce data using K-means Clustering Algorithm. **2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)**, 19 jan. 2023.

RAO, L. JAGAJEEVAN; RAO, M. VENKATA; SARADHI, T. VIJAYA. **How The Smartcard Makes the Certification Verification Easy**. Journal of Theoretical and Applied Information Technology, Vol. 83, No. 2, pp. 180-186, 2016.

REGMI, S. R. *et al*. **Customer Market Segmentation using Machine Learning Algorithm**. 6th International Conference on Trends in Electronics and Informatics (ICOEI), 2022, Tirunelveli, India. 2022. p. 1348-1354. DOI: 10.1109/ICOEI53556.2022.9777146.

ROUSSEEUW, P. J. **Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis**. J. Comput. Appl. Math., v. 20, p. 53–65, nov. 1987.

SAYAN, I. U.; DEMIRDAG, M.; YUCETURK, G.; YALCINKAYA, S. M. **A Review of Customer Segmentation Methods: The Case of Investment Sector**. Em: 5th IEEE International Conference on Big Data and Artificial Intelligence (BDAI), 2022, Fuzhou, China. Pages 200-204. DOI: 10.1109/BDAI56143.2022.9862801.

SEN, J. **Machine Learning**. Rijeka: IntechOpen, 2021. ISBN 978-1-83969-485-1. Disponível em: https://doi.org/10.5772/intechopen.94615.

SICSÚ, A.L.; SAMARTIN, A.; BARTH, NELSON. L**. Técnicas de machine learning.** [S.I.] Editora Blucher, 2023. E-book. ISBN 9786555063974.

SINGH, J.; JAISWAL, K.; SINGH, M.; SAMA, M.; SINGHAL, S. **Market segmentation using ML**. International Conference on Disruptive Technologies (ICDT), 2023, Greater Noida, India. Páginas 703-707. DOI: 10.1109/ICDT57929.2023.10150639.

T. KANSAL, S.; BAHUGUNA, V.; SINGH AND T.; CHOUDHURY: **Customer Segmentation using K-means Clustering**. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171.9

TAN, P.-N.; STEINBACH, M.; KUMAR, V**. Introduction to Data Mining**. [S.l.]: Pearson, 2013 TAVAKOLI, M. et al. **Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques**. 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), out. 2018.

TYNAN, A. C.; DRAYTON, J. **Market segmentation.** Journal of Marketing Management, v. 2, n. 3, p. 301-335, 1987. DOI: 10.1080/0267257X.1987.9964020.

WANI, A.; PRIYANKA, M.; PRASATH, R. **Unleashing Customer Insights: Segmentation Through Machine Learning**. World Conference on Communication & Computing (WCONF), 2023, RAIPUR, Índia. Páginas 1-5. DOI: 10.1109/WCONF58270.2023.10235136.

ZIAFAT, H. **International Journal of Engineering Research and Applications**, ISSN: 2248-9622, Vol. 4, Issue 9 (Version 3), September 2014, páginas 70-79.