Análise de K-Means e KNN: Fundamentos e Aplicações Práticas

Explorando Geometria, Escalabilidade e Impacto Prático em Machine Learning

Bruno Martins Mendes Vieira

Programa de Pós-Graduação em Física Ambiental

27 de Agosto de 2025

Roteiro da Apresentação

Roteiro da Apresentação

Introdução ao Contexto Histórico do KNN

- O algoritmo K-Nearest Neighbors (KNN) é um dos métodos mais intuitivos e fundamentais em aprendizado de máquina.

Introdução ao Contexto Histórico do KNN

- O algoritmo K-Nearest Neighbors (KNN) é um dos métodos mais intuitivos e fundamentais em aprendizado de máquina.
- Sua história pode ser dividida em três fases principais:
  - Origem (1951)
  - Formalização e Popularização (1967)
  - Evolução e Otimização (1970 em diante)

A Origem do KNN: Um Projeto Militar Secreto (1951)

- **Contexto:** Surgiu em um contexto militar, logo após a Segunda Guerra Mundial, diferente de algoritmos acadêmicos tradicionais.

A Origem do KNN: Um Projeto Militar Secreto (1951)

- **Contexto:** Surgiu em um contexto militar, logo após a Segunda Guerra Mundial, diferente de algoritmos acadêmicos tradicionais.
- **Quem:** Formulado por **Evelyn Fix** e **Joseph Hodges**, estatísticos da Força Aérea dos EUA.
- **Onde:** Desenvolvido na **Escola de Medicina da Aviação da Força Aérea dos EUA**.
- **O que:** Relatório técnico de 1951, *"Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties"*.
  - Propôs um método de classificação não-paramétrico para classificar objetos (ex.: aeronaves amigas ou inimigas) sem suposições sobre a distribuição dos dados.

A Origem do KNN: Um Projeto Militar Secreto (1951)

- **Contexto:** Surgiu em um contexto militar, logo após a Segunda Guerra Mundial, diferente de algoritmos acadêmicos tradicionais.
- **Quem:** Formulado por **Evelyn Fix** e **Joseph Hodges**, estatísticos da Força Aérea dos EUA.
- **Onde:** Desenvolvido na **Escola de Medicina da Aviação da Força Aérea dos EUA**.
- **O que:** Relatório técnico de 1951, *"Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties"*.
  - Propôs um método de classificação não-paramétrico para classificar objetos (ex.: aeronaves amigas ou inimigas) sem suposições sobre a distribuição dos dados.
- **Curiosidade:** O trabalho não foi publicado abertamente na época, provavelmente por ser confidencial, permanecendo restrito por mais de uma década.

Formalização e Popularização (1967)

- **Contexto:** O KNN ganhou notoriedade nos anos 60, entrando no campo acadêmico da ciência da computação e teoria da informação.

Formalização e Popularização (1967)

- **Contexto:** O KNN ganhou notoriedade nos anos 60, entrando no campo acadêmico da ciência da computação e teoria da informação.
- **Quem: Thomas Cover** e **Peter Hart**, da Universidade de Stanford.

Formalização e Popularização (1967)

- **Contexto:** O KNN ganhou notoriedade nos anos 60, entrando no campo acadêmico da ciência da computação e teoria da informação.
- **Quem: Thomas Cover** e **Peter Hart**, da Universidade de Stanford.
- **O que:** Publicaram o artigo *"Nearest Neighbor Pattern Classification"* em 1967, que foi crucial por:

Formalização e Popularização (1967)

- **Contexto:** O KNN ganhou notoriedade nos anos 60, entrando no campo acadêmico da ciência da computação e teoria da informação.
- **Quem: Thomas Cover** e **Peter Hart**, da Universidade de Stanford.
- **O que:** Publicaram o artigo *"Nearest Neighbor Pattern Classification"* em 1967, que foi crucial por:
  - Formalizar a regra do vizinho mais próximo ($k = 1$).

Formalização e Popularização (1967)

- **Contexto:** O KNN ganhou notoriedade nos anos 60, entrando no campo acadêmico da ciência da computação e teoria da informação.
- **Quem: Thomas Cover** e **Peter Hart**, da Universidade de Stanford.
- **O que:** Publicaram o artigo *"Nearest Neighbor Pattern Classification"* em 1967, que foi crucial por:
  - Formalizar a regra do vizinho mais próximo ($k = 1$).
  - Provar que a taxa de erro do KNN não é pior que o dobro da taxa de erro do classificador de Bayes.

Formalização e Popularização (1967)

- **Contexto:** O KNN ganhou notoriedade nos anos 60, entrando no campo acadêmico da ciência da computação e teoria da informação.
- **Quem: Thomas Cover** e **Peter Hart**, da Universidade de Stanford.
- **O que:** Publicaram o artigo *"Nearest Neighbor Pattern Classification"* em 1967, que foi crucial por:
  - Formalizar a regra do vizinho mais próximo ($k = 1$).
  - Provar que a taxa de erro do KNN não é pior que o dobro da taxa de erro do classificador de Bayes.
  - Popularizar o termo **"Nearest Neighbor"** na comunidade de reconhecimento de padrões e aprendizado de máquina.

Evolução e Otimização (1970 em diante)

- **Contexto:** Após 1967, o KNN tornou-se um algoritmo fundamental, ensinado em cursos de aprendizado de máquina.

Evolução e Otimização (1970 em diante)

- **Contexto:** Após 1967, o KNN tornou-se um algoritmo fundamental, ensinado em cursos de aprendizado de máquina.
- **Avanços:**

Evolução e Otimização (1970 em diante)

- **Contexto:** Após 1967, o KNN tornou-se um algoritmo fundamental, ensinado em cursos de aprendizado de máquina.
- **Avanços:**
  - **KNN Ponderado:** Introdução de pesos para vizinhos mais próximos.

Evolução e Otimização (1970 em diante)

- **Contexto:** Após 1967, o KNN tornou-se um algoritmo fundamental, ensinado em cursos de aprendizado de máquina.
- **Avanços:**
  - **KNN Ponderado:** Introdução de pesos para vizinhos mais próximos.
  - **Estruturas de dados:** Desenvolvimento de *k-d trees* para buscas mais rápidas em grandes conjuntos de dados.

Evolução e Otimização (1970 em diante)

- **Contexto:** Após 1967, o KNN tornou-se um algoritmo fundamental, ensinado em cursos de aprendizado de máquina.
- **Avanços:**
  - **KNN Ponderado:** Introdução de pesos para vizinhos mais próximos.
  - **Estruturas de dados:** Desenvolvimento de *k-d trees* para buscas mais rápidas em grandes conjuntos de dados.
  - **Fuzzy KNN:** Em 1985, James Keller introduziu uma versão que lida com incertezas na classificação.

Conclusão

- O KNN nasceu como uma solução prática para um problema militar.
- Após um período de obscuridade, foi formalizado e popularizado pela academia.
- Tornou-se um dos algoritmos mais intuitivos e duradouros do aprendizado de máquina.
- Sua simplicidade e flexibilidade continuam a inspirar avanços e aplicações.

O Que é o K-Nearest Neighbors (KNN)?

- **Algoritmo Supervisionado:** Usado para classificação e regressão.
- **Baseado em Instâncias:** Não "aprende" um modelo explícito a partir dos dados de treino. Ele memoriza todo o conjunto de treinamento.
- **Aprendizagem Preguiçosa (Lazy Learning):** Todo o cômputo ocorre no momento da predição, não durante o "treinamento".
- **Premissa Central:** Pontos de dados semelhantes existem em proximidade uns dos outros. A "semelhança" é medida por uma métrica de distância.

Como o KNN Funciona? (Classificação)

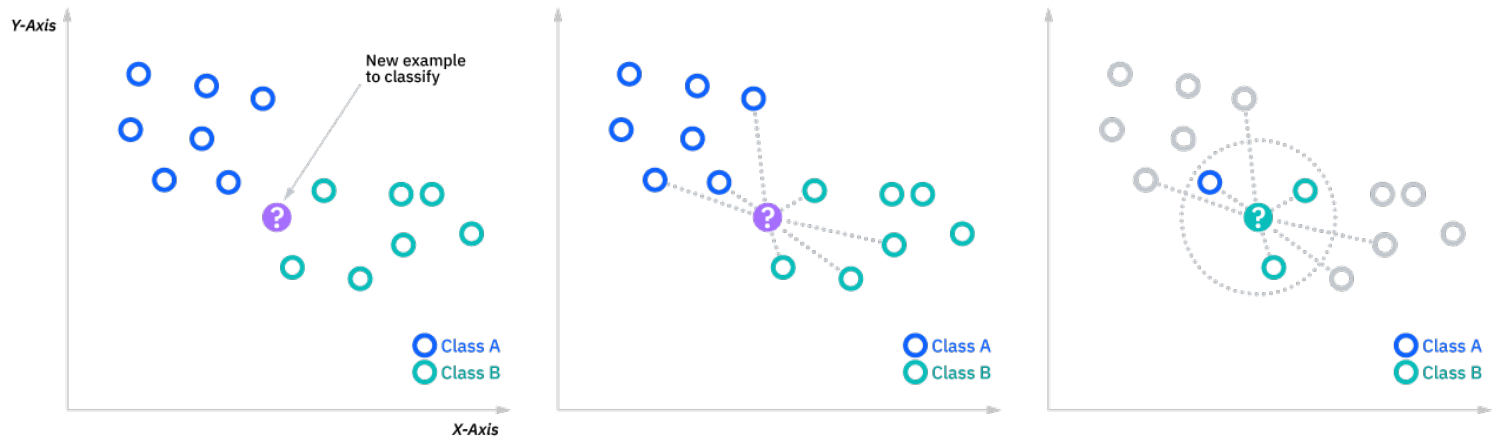


Figura: Diagrama do funcionamento do KNN [**ibm_knn_br**]

A Escolha Crítica: A Métrica de Distância

A forma como medimos a "proximidade" define a fronteira de decisão do modelo. A escolha da métrica depende da natureza dos dados.

Distância Euclideana (L_2)

É a distância mais intuitiva: o comprimento de uma linha reta entre dois pontos.

$$d(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{1}$$

Distância Euclideana (L_2)

É a distância mais intuitiva: o comprimento de uma linha reta entre dois pontos.

$$d(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{1}$$

Exemplo:

- Pontos: $p = (2, 3)$, $q = (5, 7)$

Distância Euclideana (L_2)

É a distância mais intuitiva: o comprimento de uma linha reta entre dois pontos.

$$d(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{1}$$

Exemplo:

- Pontos: $p = (2, 3),\ q = (5, 7)$

$$\begin{aligned} d(p, q) &= \sqrt{(2-5)^2 + (3-7)^2} \\ &= \sqrt{(-3)^2 + (-4)^2} \\ &= \sqrt{9 + 16} \\ &= \sqrt{25} \\ &= 5 \end{aligned}$$

Dimensionalidade

Mas, e quando aumentamos para 4 dimensões?

Clientes com quatro características

- **Idade**
- **Nº de Compras**
- **Avaliação Média**
- **Tempo como Cliente**

Perfil de Cliente:

- **Cliente A**: (Idade: 30, Compras: 15, Avaliação: 4, Tempo: 24)
- **Cliente B**: (Idade: 35, Compras: 10, Avaliação: 5, Tempo: 36)

Distância de Manhattan (L_1)

Também conhecida como "distância do táxi".

$$d(p, q) = \sum_{i=1}^{n} |p_i - q_i|$$

- **Intuição Geométrica:** Distância percorrida em uma grade (como quarteirões de uma cidade).
- **Uso Comum:** Eficaz em espaços de alta dimensão.

Distância Manhattan

- **Idade**: $|35 - 30| = 5$
- **Nº de Compras**: $|10 - 15| = |-5| = 5$
- **Avaliação Média**: $|5 - 4| = 1$
- **Tempo como Cliente**: $|36 - 24| = 12$

$$d_{\text{manhattan}} = 5 + 5 + 1 + 12 = 23 \tag{2}$$

Resultado: A distância Manhattan é **23**.

Distância Euclideana

Distância Euclideana

- **Idade**: $(35 - 30)^2 = 5^2 = 25$
- **Nº de Compras**: $(10 - 15)^2 = (-5)^2 = 25$
- **Avaliação Média**: $(5 - 4)^2 = 1^2 = 1$
- **Tempo como Cliente**: $(36 - 24)^2 = 12^2 = 144$

Distância Euclideana

- **Idade**: $(35 - 30)^2 = 5^2 = 25$
- **Nº de Compras**: $(10 - 15)^2 = (-5)^2 = 25$
- **Avaliação Média**: $(5 - 4)^2 = 1^2 = 1$
- **Tempo como Cliente**: $(36 - 24)^2 = 12^2 = 144$

$$d_{\text{euclidiana}} = \sqrt{25 + 25 + 1 + 144} = \sqrt{195} \approx 13.96 \quad (3)$$

Comparação e Análise

Distância:

- **Manhattan**: 23
- **Euclideana**: ≈ 13.96

Ponto chave: Tempo como cliente

- **Manhattan**: Contribui com 12 de 23.
- **Euclideana**: Contribui $12^2 = 144$ em 195 (mais de 73%).

A Maldição da Dimensionalidade

- **Concentração de Distância:** Em alta dimensão, a distância entre o vizinho mais próximo e o mais distante de um ponto se torna quase a mesma.

Impacto no KNN: A premissa de "vizinho próximo" perde o sentido, e a votação dos vizinhos se torna aleatória.

O Trade-off Viés-Variância na Escolha de K

K Pequeno (e.g., K=1)	**K Grande (e.g., K=N)**
Alta Variância: Sensível a ruídos e outliers.	**Baixa Variância**: Ignora nuances locais.
Resultado: *Overfitting* (superajuste).	**Resultado**: *Underfitting* (subajuste).

Tabela: Comparação entre K Pequeno e K Grande

Variações: O KNN Ponderado

A Ideia da Ponderação

Dar mais peso aos vizinhos mais próximos, usando o **inverso da distância** como peso.

Vantagem: Torna o modelo mais robusto a outliers e menos sensível à escolha exata de K.

A Vantagem do "Lazy Learning"

- **Sem Custo de Treinamento:** O "treino" é apenas armazenar os dados, o que é muito rápido.
- **Adaptação Contínua:** É fácil adicionar novos dados sem precisar retreinar o modelo do zero.
- **Interpretabilidade Local:** As predições podem ser explicadas olhando para os vizinhos que as influenciaram.

Resumo Final do KNN

- **Forças:** Simples, sem tempo de treino, naturalmente não-linear e adaptável.

Resumo Final do KNN

- **Forças:** Simples, sem tempo de treino, naturalmente não-linear e adaptável.
- **Fraquezas:** Sensível à "maldição da dimensionalidade", requer escalonamento de características, computacionalmente caro para predição.

Resumo Final do KNN

- **Forças:** Simples, sem tempo de treino, naturalmente não-linear e adaptável.
- **Fraquezas:** Sensível à "maldição da dimensionalidade", requer escalonamento de características, computacionalmente caro para predição.
- **Sucesso ou Fracasso:** Depende criticamente da escolha da **métrica de distância**, do valor de **K** e de um **pré-processamento** cuidadoso.

Roteiro da Apresentação

K-Means: Visão Geral

Definição

O K-Means é um algoritmo de aprendizado não supervisionado que agrupa um conjunto de dados em K cluster distintos e não sobrepostos.

Clusterização Inicial

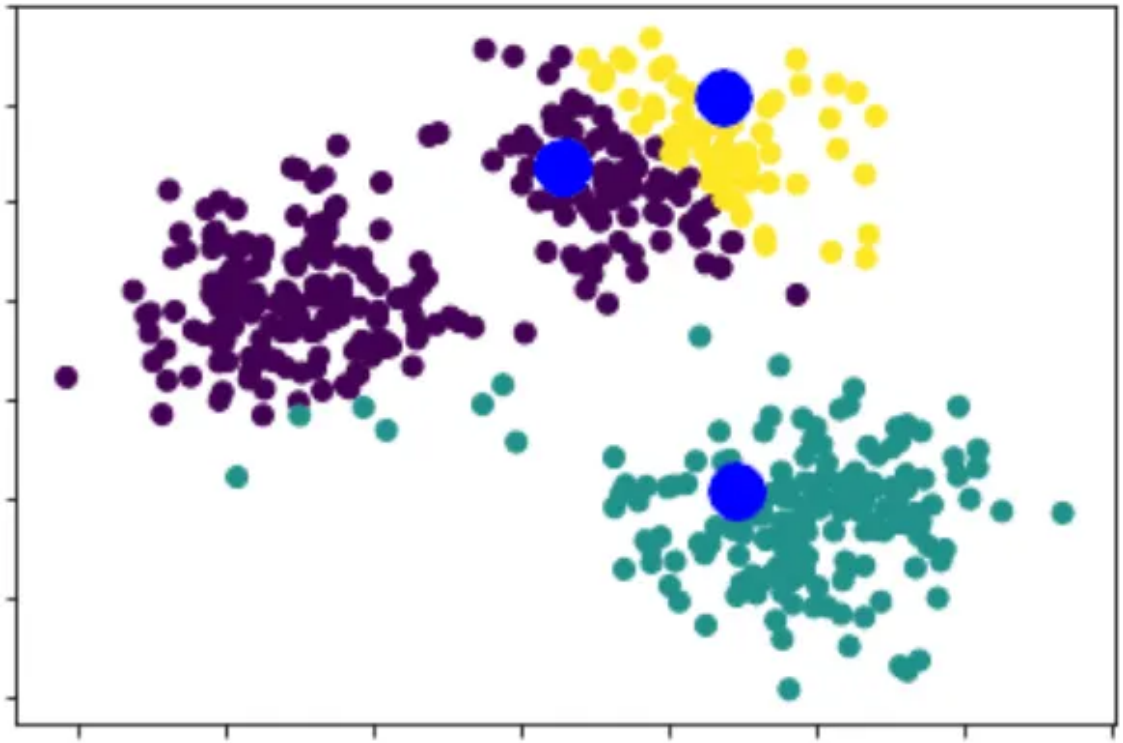


Figura: Primeira iteração, posição randômica dos centroides [**medium_cwi_kmeans**].

Processo Iterativo e Convergência

O algoritmo repete dois passos até que a atribuição dos cluster não mude:

1. **Passo de Atribuição:** Cada ponto é atribuído ao centroide mais próximo.

$$C_i^{(t)} = \{\mathbf{x} : ||\mathbf{x} - \boldsymbol{\mu}_i^{(t)}||^2 \leq ||\mathbf{x} - \boldsymbol{\mu}_j^{(t)}||^2 \quad \forall j, 1 \leq j \leq K\}$$

2. **Passo de Atualização:** Os centroides são recalculados como a média de todos os pontos atribuídos a eles.

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{\mathbf{x} \in C_i^{(t)}} \mathbf{x}$$

Problemas Inerentes e Soluções

- **Sensibilidade à Inicialização:** A escolha aleatória dos centroides iniciais pode levar a resultados ruins.
  - **Solução: K-Means++**, que escolhe os centroides iniciais de forma a estarem distantes uns dos outros.

$$P(x) = \frac{D(x)^2}{\sum_{x \in X} D(x)^2}, \tag{4}$$

$P(x)$ = Probabilidade do ponto x ser escolhido como próxima centroide

$D(x)^2$ = Distância para o centróide mais próximo

$\sum_{x \in X}$ = Soma de $D(x)^2$ para todos os pontos no dataset.

Primeiro Centróide

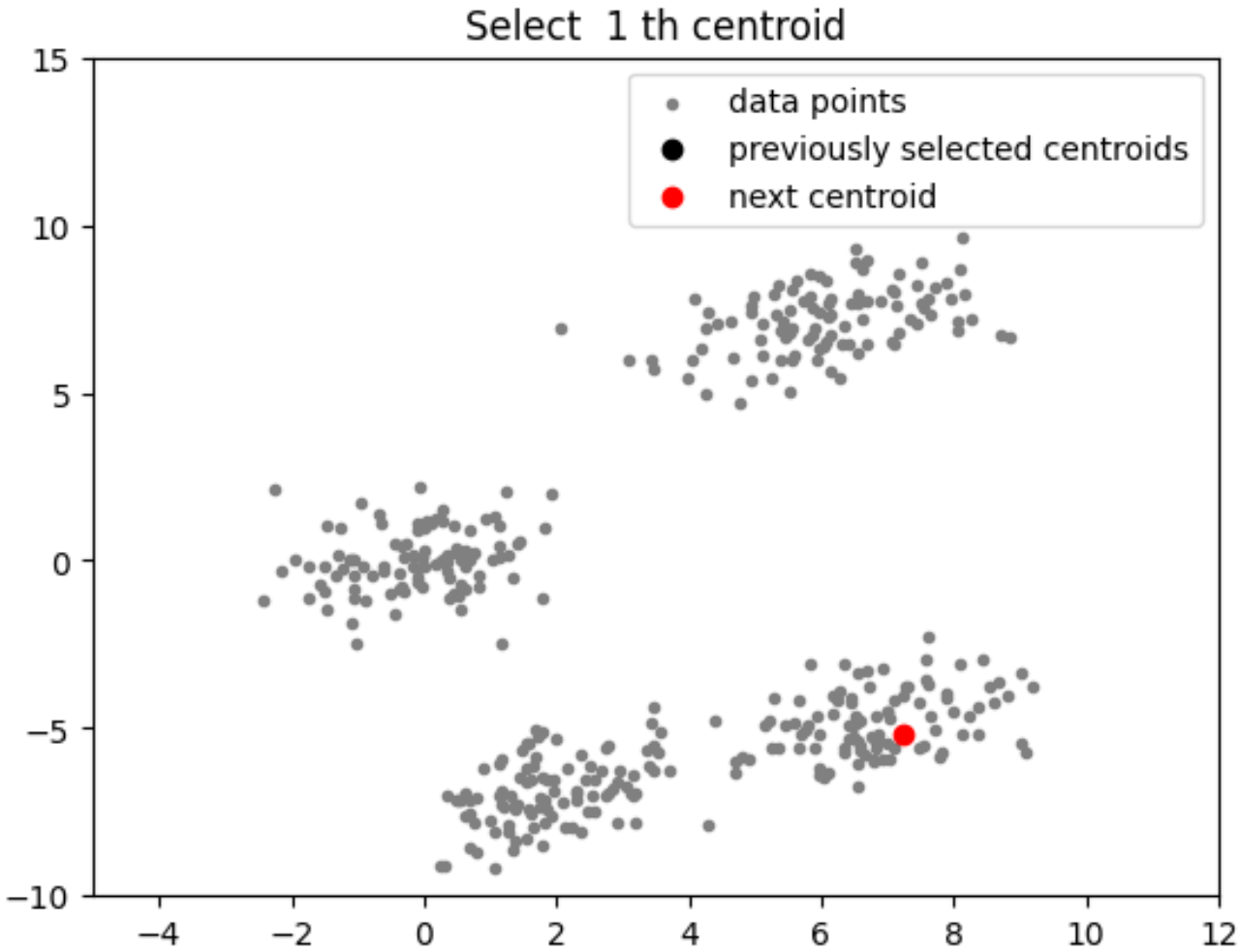


Figura: Primeiro centroide escolhido[**geeksforgeeks_kmeans**].

Segundo Centróide

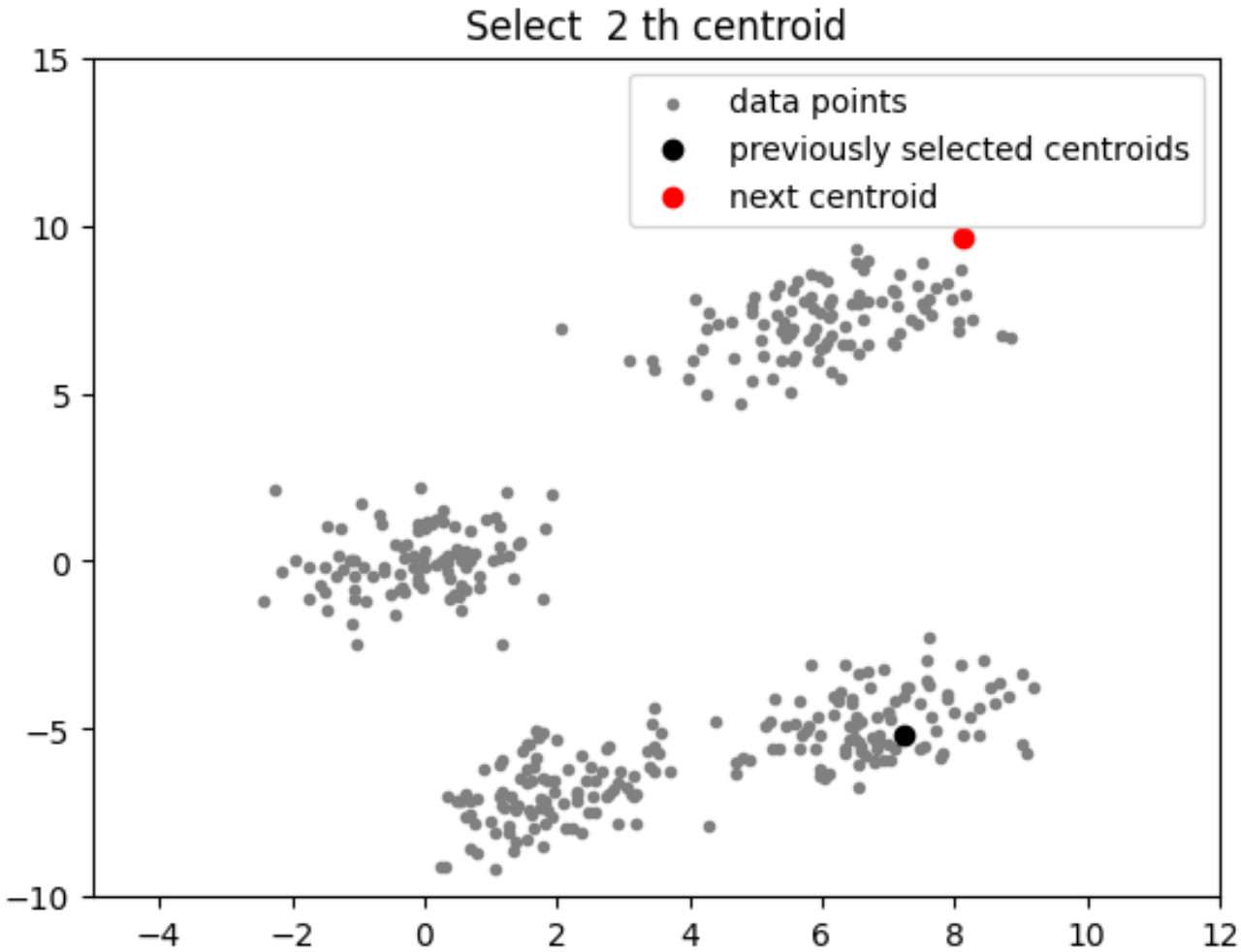


Figura: Segundo centroide escolhido [**geeksforgeeks_kmeans**].

Terceiro Centróide

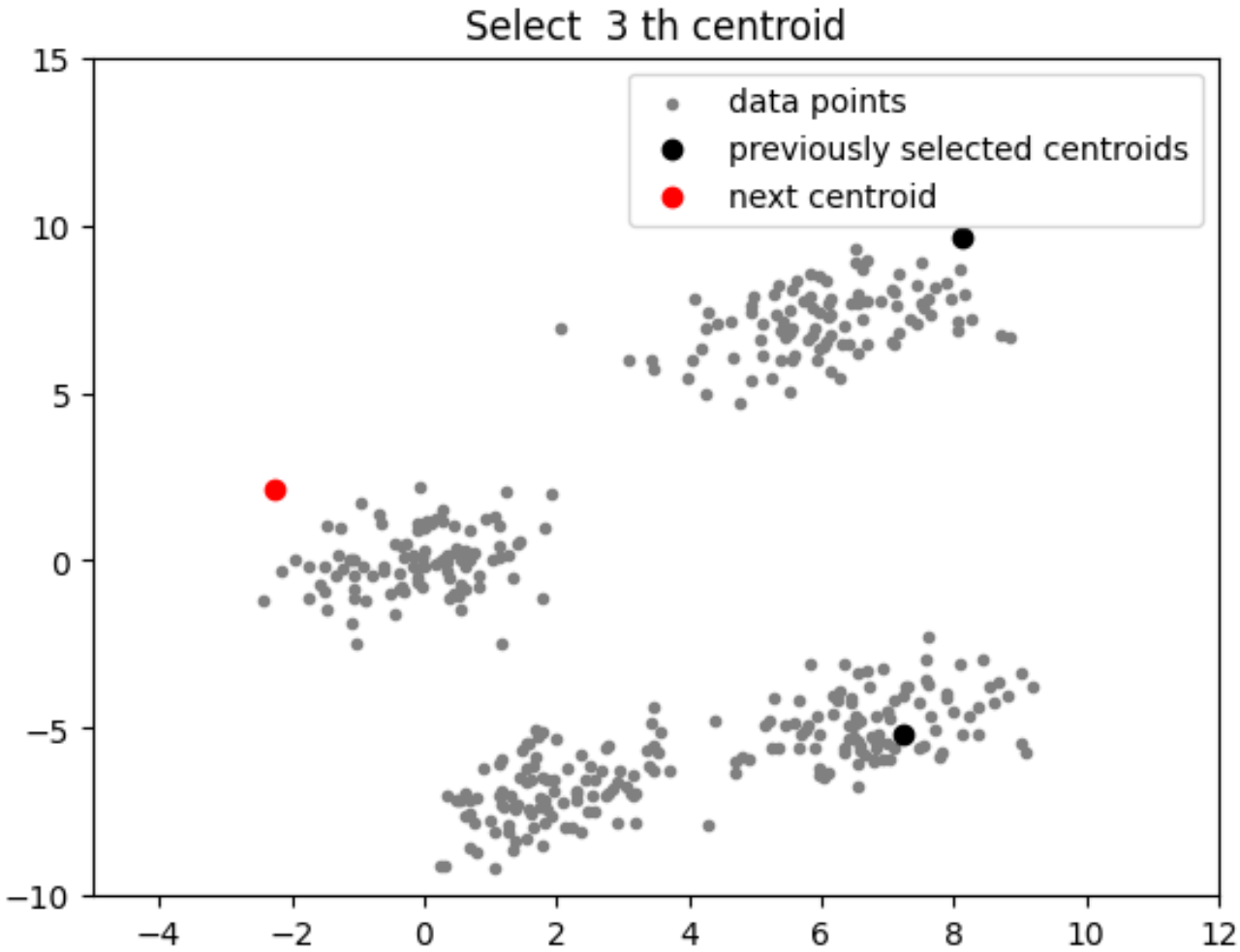


Figura: Terceiro centroide escolhido[**geeksforgeeks_kmeans**].

Quarto Centróide

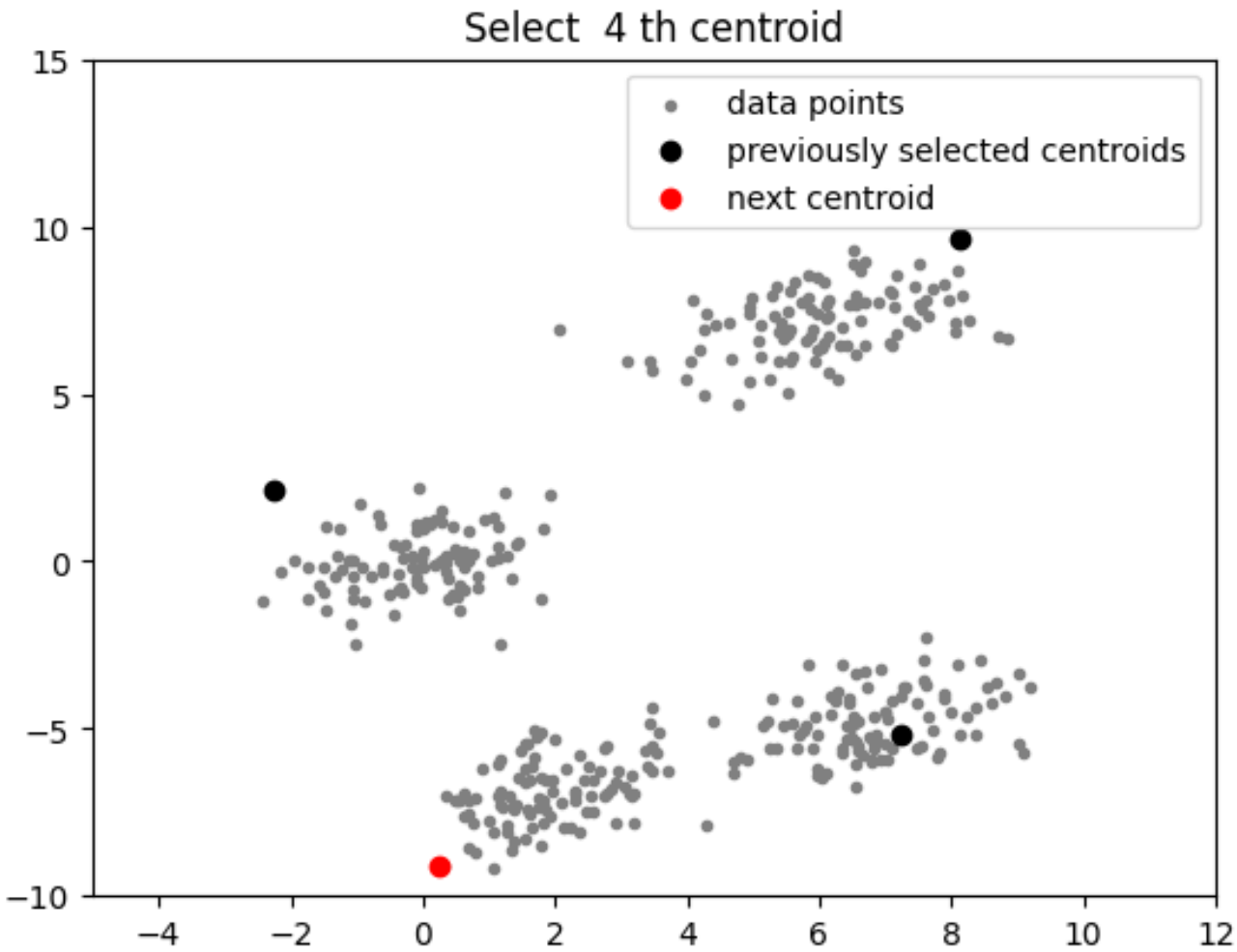


Figura: Quarto centroide selecionado[**geeksforgeeks_kmeans**].

Quantidade de Clusters

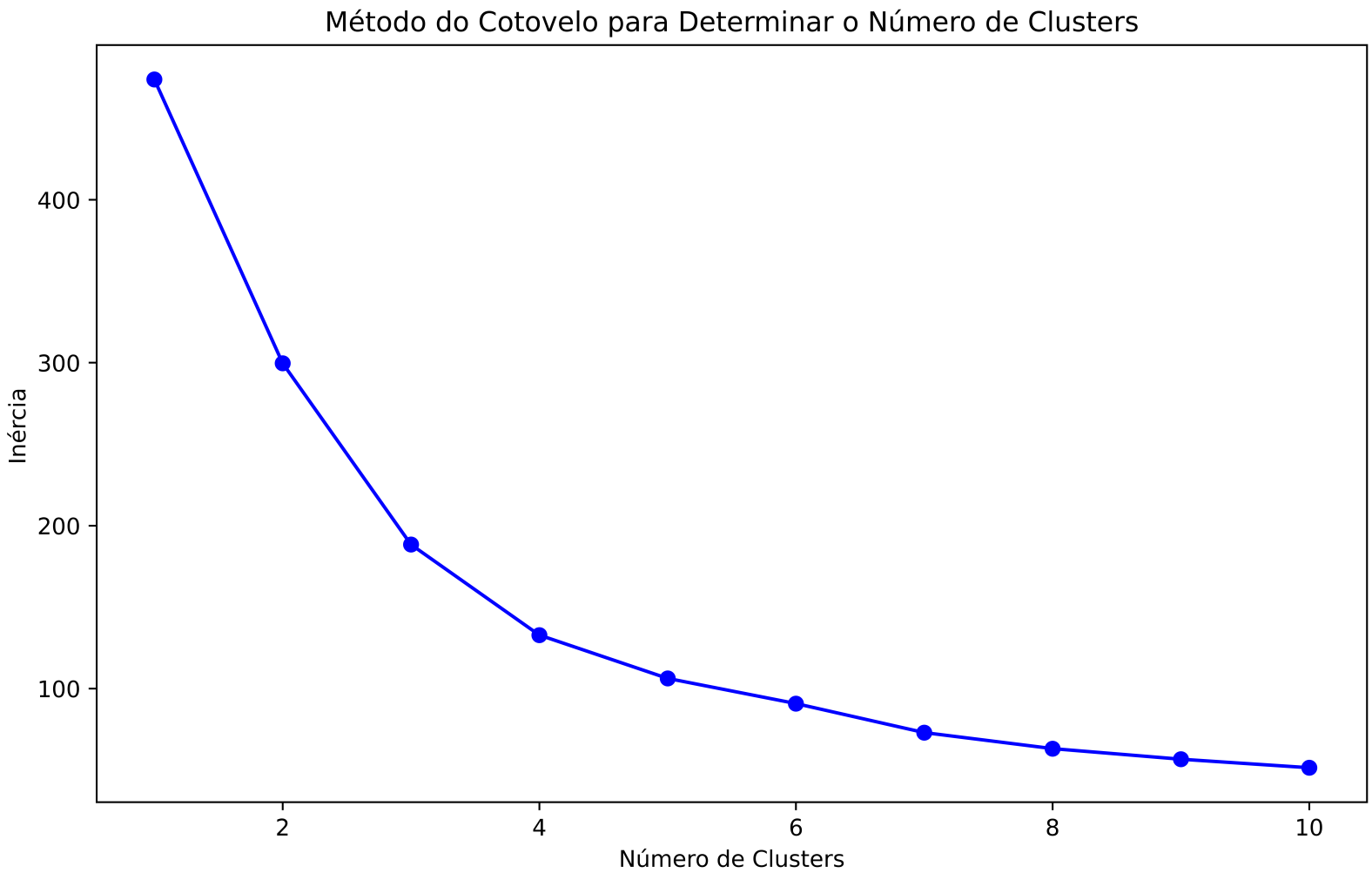
Mas, como podemos determinar a quantidade de cluster?

Quantidade de Clusters

Mas, como podemos determinar a quantidade de cluster?

- Método do Cotovelo (*Elbow Method*)
- Análise da Silhueta (*Silhouette Analysis*)

Método do Cotovelo (*Elbow Method*)



Análise da Silhueta (*Silhouette Analysis*)

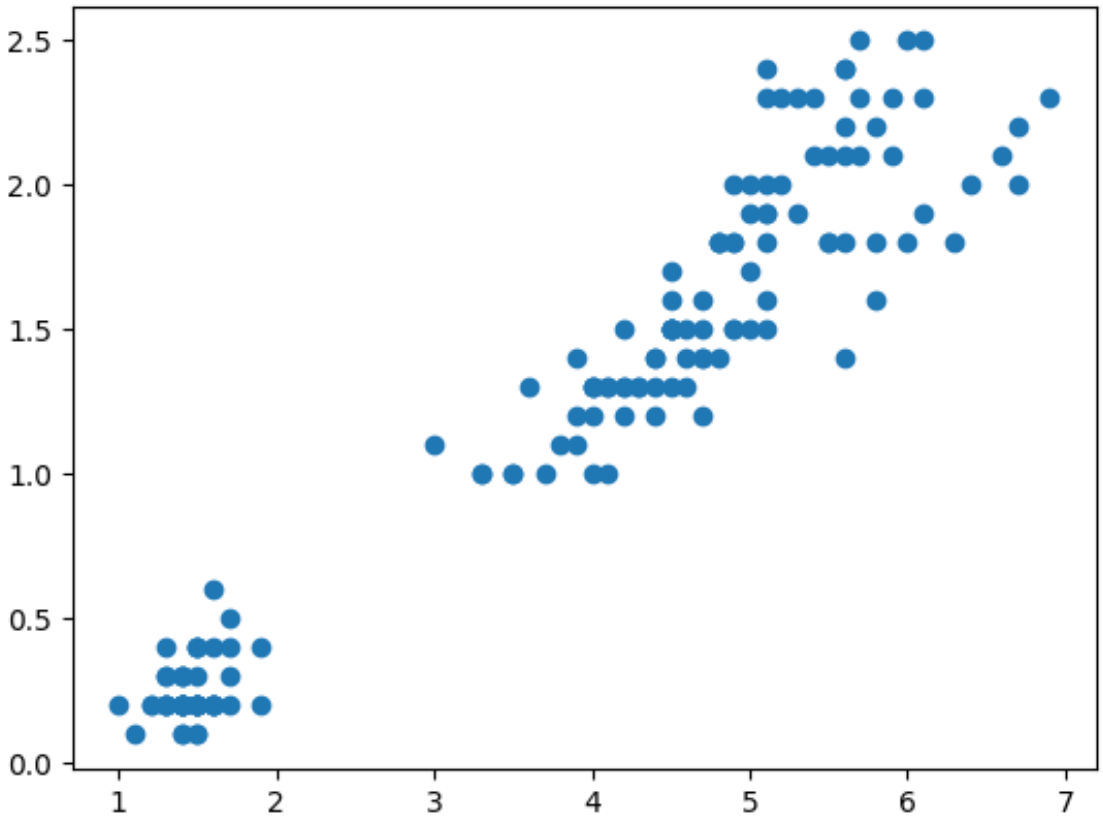


Figura: Pontos iniciais.

$K = 5$

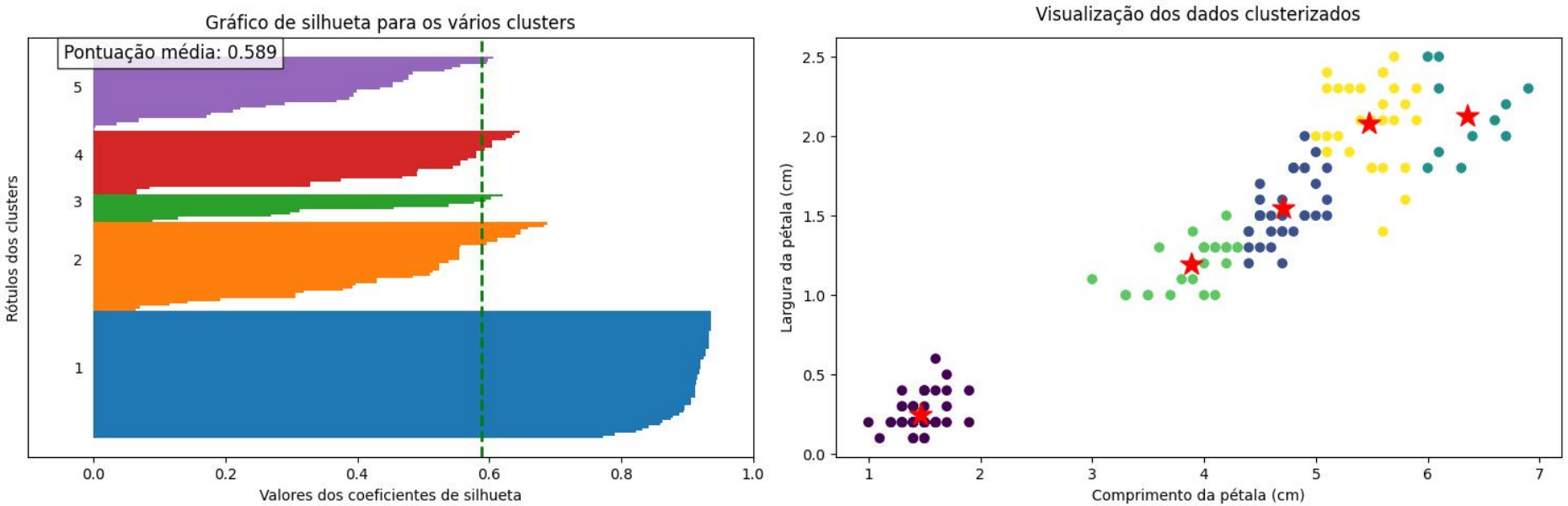


Figura: Pontuação média para cinco cluster.

$K = 4$

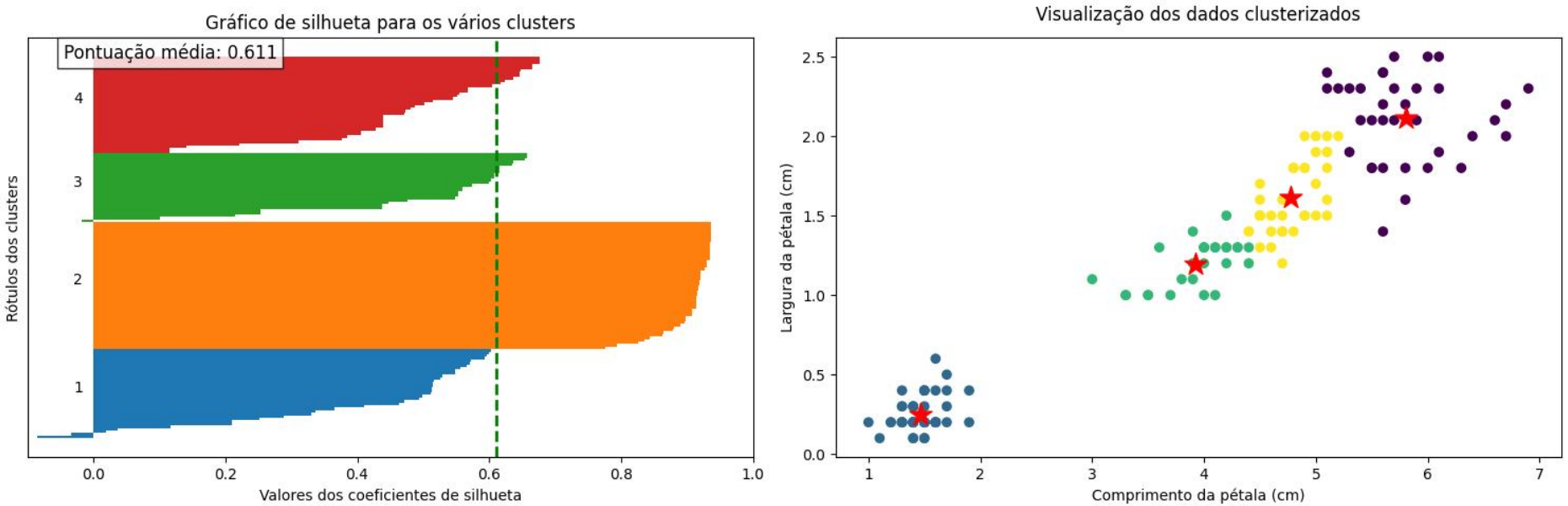


Figura: Pontuação média para quatro cluster.

$K = 3$

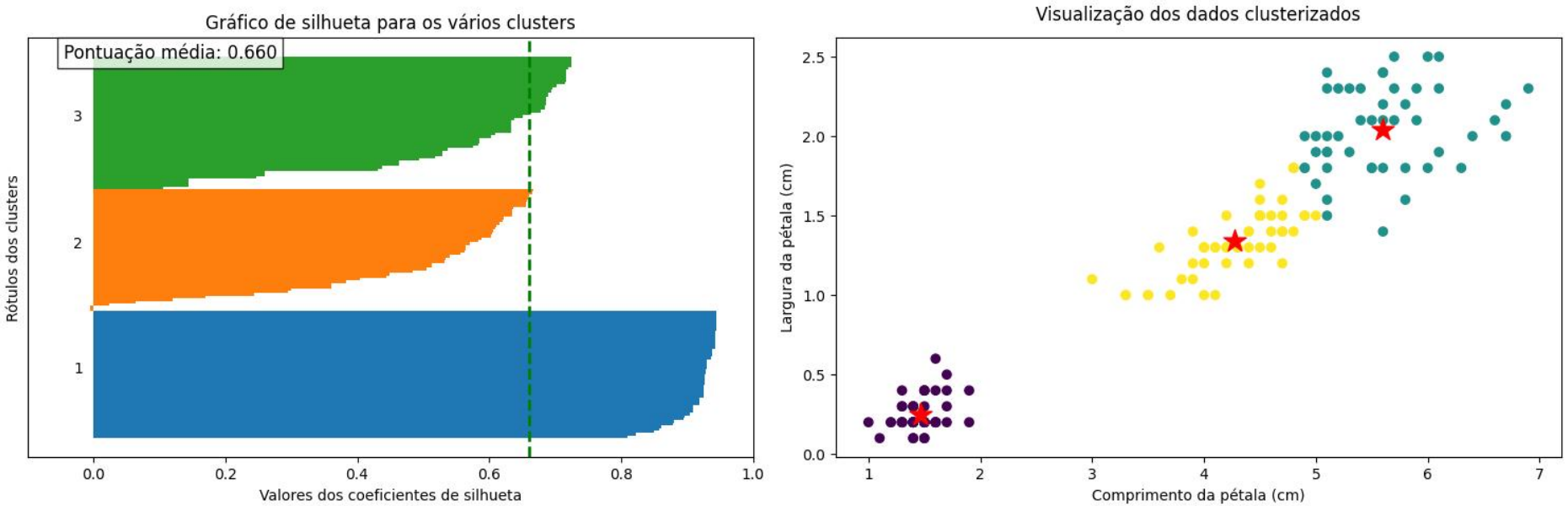


Figura: Pontuação média para três cluster.

$K = 2$

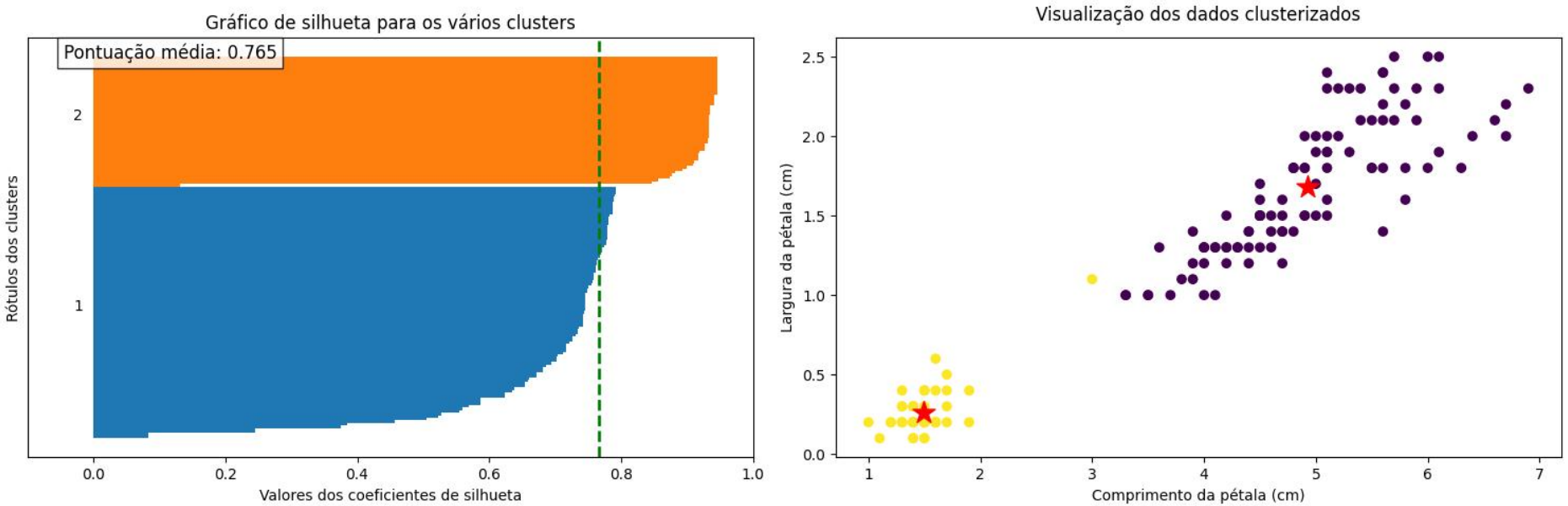


Figura: Pontuação média para dois cluster.

Comparação com Outros Métodos

K-Means	**DBSCAN (Baseado em Densidade)**
Requer K pré-definido.	Não requer K.
Assume clusters esféricos.	Encontra clusters de formas arbitrárias.
Particiona todos os pontos.	Robusto a outliers (classifica-os como ruído).
Rápido em dados de baixa dimensão.	Não atribui todos os pontos.

Tabela: Comparação entre entre técnicas, mostrando suas diferenças.

Resumo Final do K-Means

Pontos Fortes

- **Simplicidade e Rapidez:** Fácil de implementar e computacionalmente eficiente.
- **Escalabilidade:** Funciona bem em datasets grandes.

Pontos Fracos

- **Fraquezas:** Vulnerável à inicialização e a outliers. Inadequado para clusters não-esféricos.

Roteiro da Apresentação

Introdução ao Estudo de Caso

Imagine que somos um banco e precisamos entender o comportamento dos nossos clientes. Como o K-Means e o KNN podem nos ajudar a segmentá-los para um marketing mais eficiente?

Introdução: K-Means vs. KNN em Finanças

- **K-Means**: Algoritmo de clusterização não supervisionada que agrupa dados por similaridade, sem necessidade de rótulos prévios.
- **KNN**: Método de classificação supervisionada que faz previsões com base nos K vizinhos mais próximos, utilizando métricas de distância.
- **Relevância**: Ambos são amplamente utilizados em finanças para segmentação de clientes, previsão de comportamentos e gestão de riscos.

K-Means: Aplicações Práticas em Finanças

- **Segmentação de Clientes**: Agrupar clientes por comportamento para marketing direcionado.
- **Gestão de Risco**: Classificar candidatos a empréstimo por perfil de risco.
- **Detecção de Fraude**: Identificar transações anômalas como outliers.
- **Otimização de Caixas Eletrônicos**: Posicionar recursos em áreas de alta demanda com base em padrões de uso.

Cenário Hipotético com K-Means: Segmentação de Clientes

Desafio: Otimizar o serviço bancário para diferentes perfis de cliente.

- **Passo 1**: Coletar dados demográficos e de transações (e.g., idade, renda anual, gastos mensais) de milhões de contas.
- **Passo 2**: Normalizar os dados para mitigar o impacto de outliers e escalas diferentes.
- **Passo 3**: Aplicar o método do cotovelo para determinar o número ideal de clusters ($K = 5$).
- **Passo 4**: Executar o K-Means para agrupar clientes em perfis, como "investidores jovens" e "poupadores aposentados".
- **Passo 5**: Desenvolver produtos específicos, como robôs-consultores para jovens e planos de previdência para idosos.

Resultado: Melhoria na retenção de clientes por meio de estratégias personalizadas.

Resultados do K-Means: Método do Cotovelo

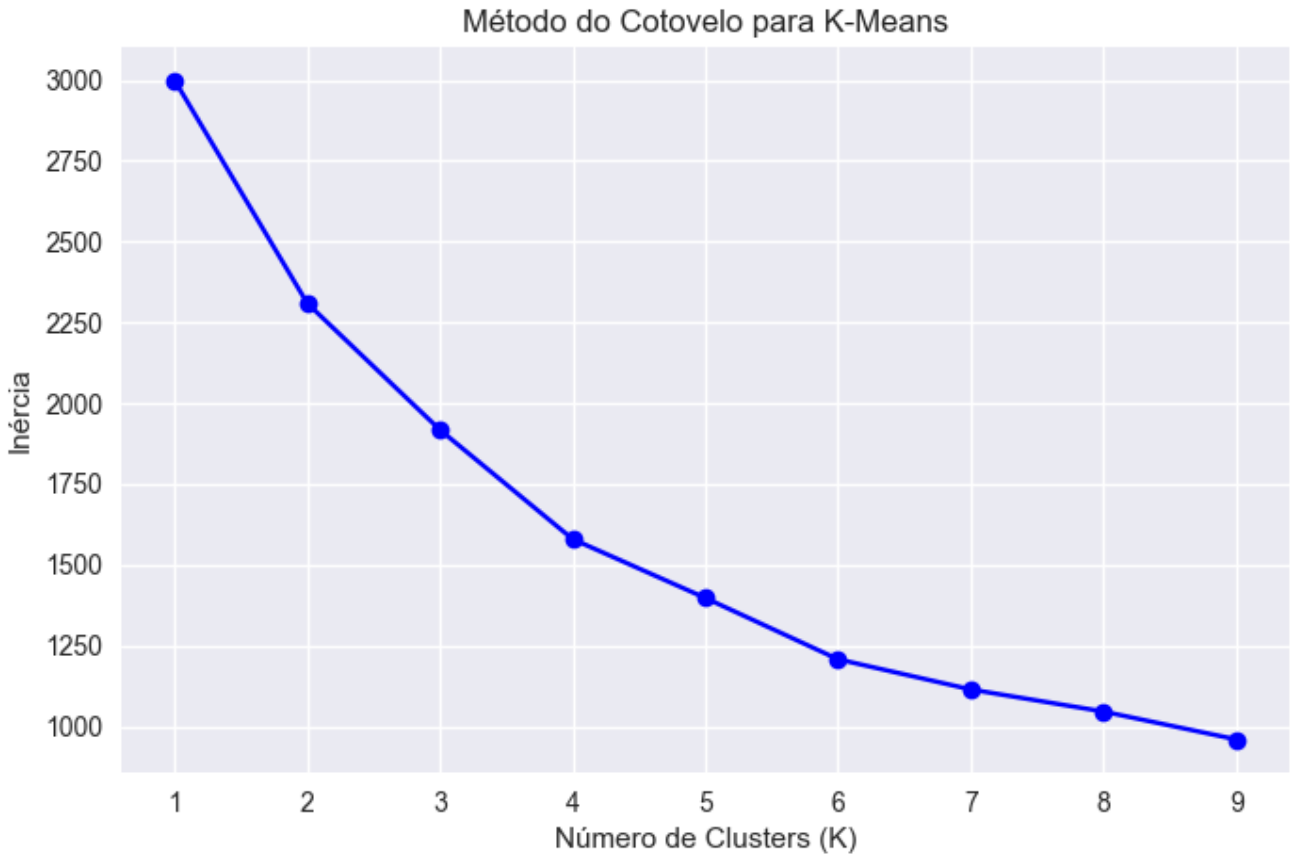


Figura: Método do cotovelo para determinar o número ideal de clusters $K = 5$.

Resultados do K-Means: Segmentação de Clientes

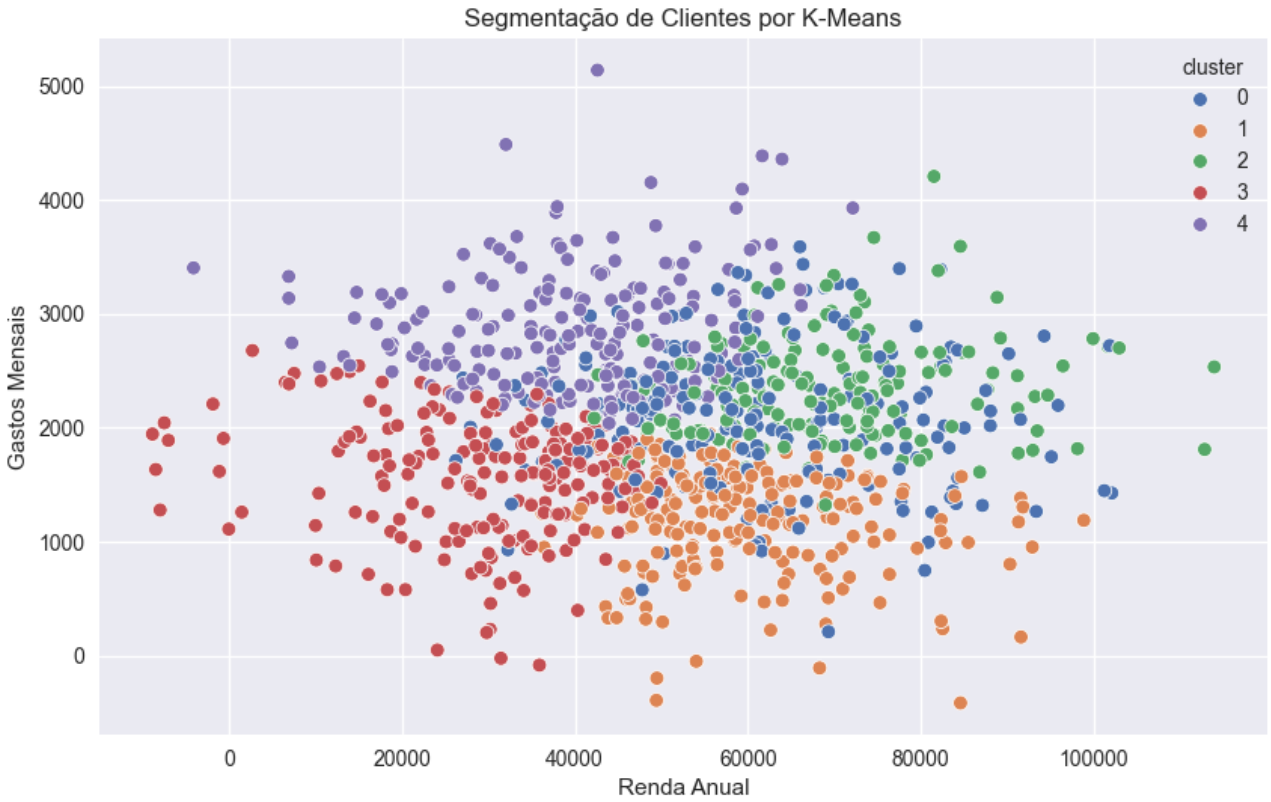


Figura: Segmentação de 1000 clientes com base em renda anual e gastos mensais.

KNN: Aplicações Práticas em Finanças

- **Previsão de Ações**: Prever preços de ativos usando similaridades históricas.
- **Score de Crédito**: Classificar a pontuação de risco de candidatos com base em dados passados.
- **Detecção de Fraude**: Identificar transações atípicas com base em desvios de padrões conhecidos
- **Previsão de Falência**: Avaliar o risco de falha de empresas por comparações históricas.

Cenário Hipotético com KNN: Previsão de Ações

Desafio: Prever a direção do preço de uma ação para o próximo dia.

- **Passo 1**: Coletar dados históricos de uma ação (preços, volume, indicadores técnicos).
- **Passo 2**: Pré-processar os dados, normalizando as características.
- **Passo 3**: Selecionar $K = 5$ e a métrica de distância Euclidiana.
- **Passo 4**: Identificar os K vizinhos mais próximos com base nas condições atuais do mercado.
- **Passo 5**: Prever o movimento (alta ou queda) com base na maioria dos vizinhos.

Resultado: Aumento na precisão das decisões de compra e venda.

Resultados do KNN: Previsão de Movimentos de Ações

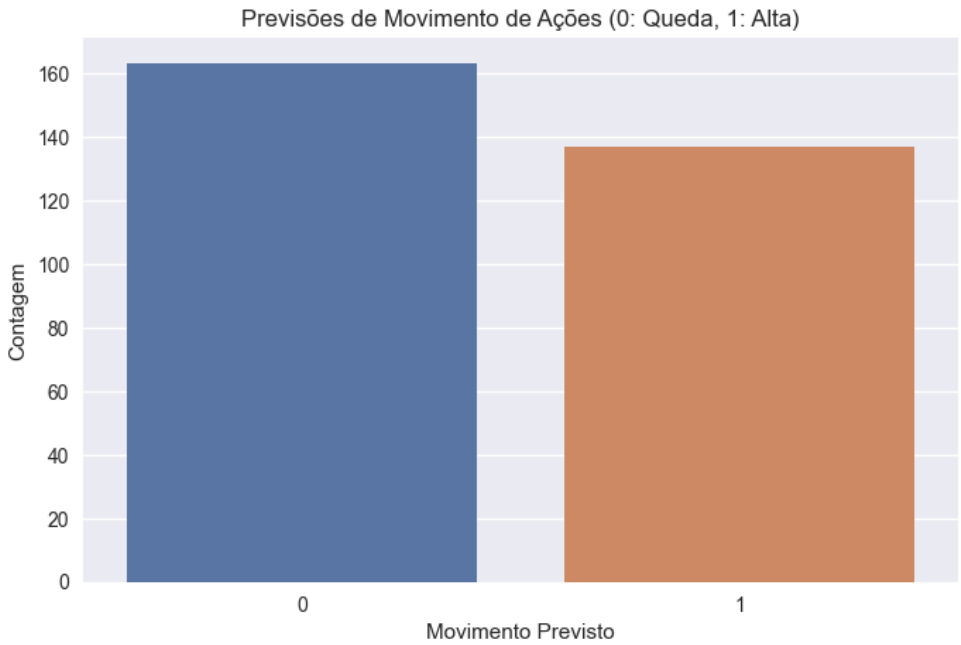


Figura: Distribuição das previsões de movimento de 300 ações de teste (0: Queda, 1: Alta).

Insights Comparativos e Futuro

- **K-Means**: Eficaz para segmentação não supervisionada, mas requer interpretação humana para nomear clusters e é sensível a inicializações.
- **KNN**: Ideal para previsões supervisionadas, mas computacionalmente intensivo em grandes datasets e sensível à maldição da dimensionalidade.
- **Modelos Híbridos**: A integração de K-Means e KNN pode melhorar a detecção de fraudes ou previsões ao combinar segmentação e classificação.
- **Desafios**: Mitigar ruídos nos dados, reduzir viés e garantir escalabilidade em grandes volumes de dados financeiros.

Roteiro da Apresentação

Resumo da Apresentação

- Exploramos a simplicidade e as complexidades dos algoritmos KNN (supervisionado) e K-Means (não supervisionado).
- Discutimos a importância das **métricas de distância**, da **escolha de K** e do **pré-processamento** para o sucesso desses modelos.
- Demonstramos como, apesar de seus desafios, eles são ferramentas poderosas e versáteis para **análise e previsão de dados**, com aplicações concretas em finanças e outras áreas.

Referências I

referencias

Obrigado pela presença de todos!

brunomendes@fisica.ufmt.br