Algoritmos Não Supervisionados – Clustering Hierárquico

ROCKETSEAT

Plataforma completa de aprendizado contínuo em programação.

#BoostingPeople

rocketseat.com.br

Algoritmos Não Supervisionados

Clusterização Hierárquica

O objetivo deste módulo é apresentar o algoritmo de clusterização hierárquica e trabalharemos num projeto para um marketplace de laptops onde faremos o processo completo desde o EDA até a entrega de aplicação que irá consultar os clusters obtidos pelo modelo.

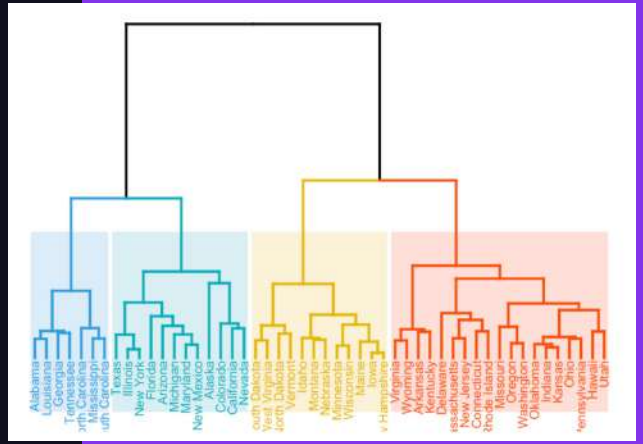
Agenda

- O que é Clusterização Hierárquica
- Etapas da Clusterização Hierárquica
- O que é o algoritmo de Clusterização Hierárquica Aglomerativo
- O que é o algoritmo de Clusterização Hierárquica Divisivo
- O que é um Dendrograma
- Projeto – Clusterização Hierárquica.

O que é clusterização hierárquica

A clusterização hierárquica é uma técnica distinta dentro dos métodos de agrupamento em aprendizado de máquina não supervisionado, que se diferencia principalmente pela forma como os clusters são estruturados e pela flexibilidade no número de clusters.

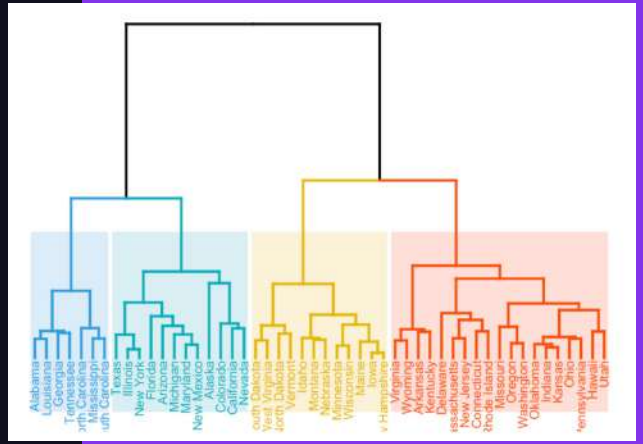
Além disso, é particularmente útil para análises exploratórias e situações onde a relação entre os pontos de dados é mais importante do que a formação de grupos distintos com fronteiras claras. É também apropriado para conjuntos de dados onde as relações entre os dados podem ser representadas em múltiplas escalas ou níveis de agregação.



O que é clusterização hierárquica

Estrutura Hierárquica

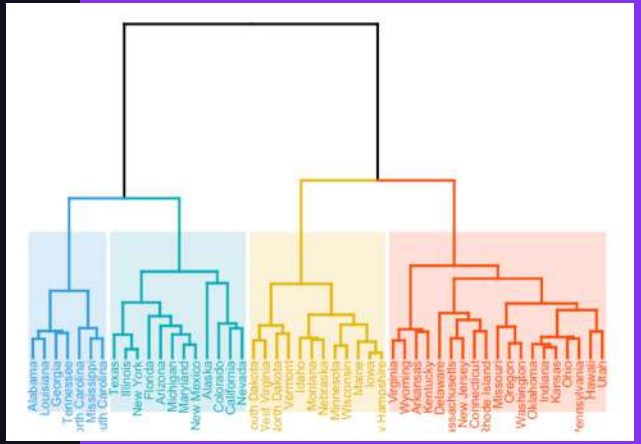
Diferentemente de métodos como K-means, que partem de uma definição inicial de quantos clusters serão formados, o agrupamento hierárquico cria uma árvore de clusters, conhecida como dendrograma. Esta árvore pode ser interpretada em diferentes níveis de granularidade, permitindo uma visão mais detalhada ou mais agregada conforme necessário.



O que é clusterização hierárquica

Não requer a especificação do número de clusters

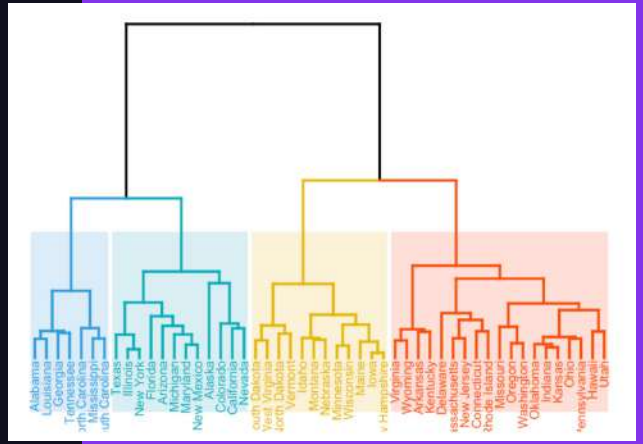
Enquanto métodos como o K-means exigem que o número de clusters seja definido a priori, no agrupamento hierárquico o número de clusters pode ser determinado após a construção do dendrograma, através de sua análise e do ponto de corte escolhido.



O que é clusterização hierárquica

Sensibilidade a mudanças nos dados

O agrupamento hierárquico constrói clusters baseado em etapas sucessivas de aglomeração/fusão ou divisão, o que o torna sensível à ordem dos dados e a outliers. Isso pode resultar em clusters diferentes se a ordem dos dados for alterada, diferentemente de métodos como DBSCAN que são mais robustos a outliers e à ordem dos pontos.

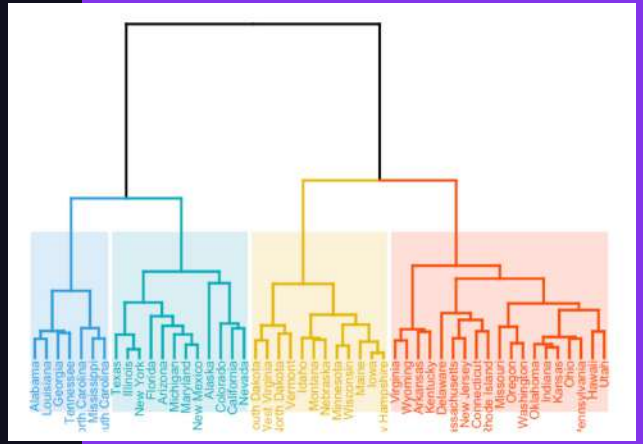


O que é clusterização hierárquica

Comparação com K-means:

K-means é eficiente para grandes conjuntos de dados e para clusters de forma esférica, mas falha em capturar clusters de formas complexas ou tamanhos variados. Por outro lado, o agrupamento hierárquico pode identificar essas estruturas mais complexas, embora seja menos eficiente em termos de tempo computacional para grandes conjuntos de dados.

K-means pode resultar em diferentes soluções dependendo dos centroides iniciais. A clusterização hierárquica proporciona uma saída mais estável em termos de hierarquia construída, mas é mais suscetível a ser influenciada por outliers.

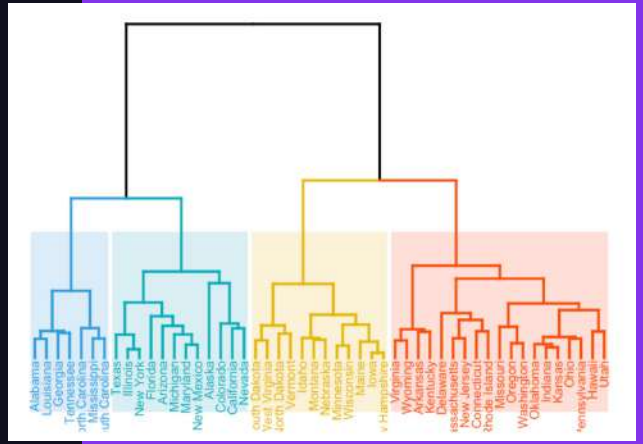


O que é clusterização hierárquica

Etapas

1) Definição de Similaridade: Primeiramente, é necessário definir uma métrica de similaridade. Comumente, distâncias como a Euclidiana, Manhattan, ou outras distâncias específicas do domínio são usadas para quantificar quão similares ou distintos são os objetos.

2) Construção da Matriz de Distância: Calcula-se a distância entre cada par de objetos no conjunto de dados, resultando em uma matriz de distância.



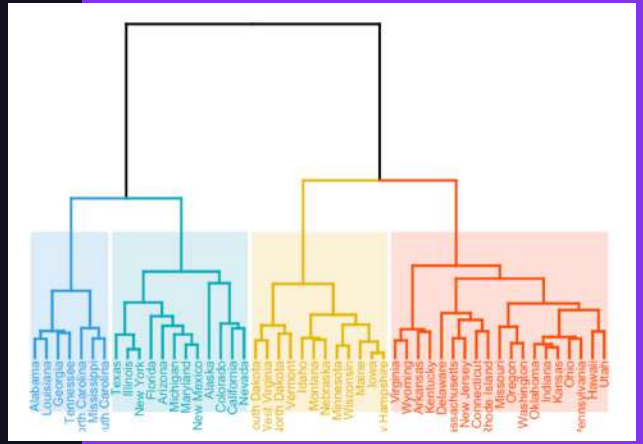
O que é clusterização hierárquica

Etapas

3) Construção do Dendrograma:

Aglomerativo: Inicialmente, cada objeto é tratado como um cluster individual e os clusters mais próximos são fundidos.
Divisivo: Inicia-se com um único cluster que inclui todos os objetos e o dendrograma é construído dividindo sucessivamente os clusters.

4) Corte do Dendrograma: A escolha do número de clusters é feita cortando o dendrograma em uma certa altura, que define o número final de clusters.



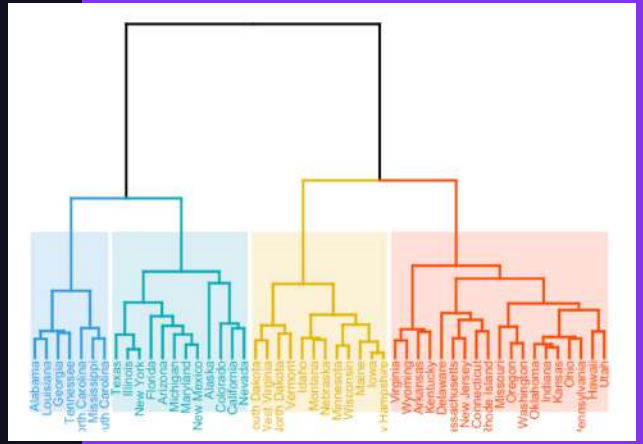
O que é o algoritmo de clusterização hierárquica aglomerativo

Etapas

1) Inicialização: Cada ponto de dado é tratado como um cluster individual.

2) Cálculo da Matriz de Distância: Antes de iniciar o processo de fusão, calcula-se a matriz de distância que contém as distâncias entre todos os pares de pontos.

3) Fusão de Clusters: Encontre os dois clusters que estão mais próximos um do outro com base na matriz de distância e combine em um novo cluster.



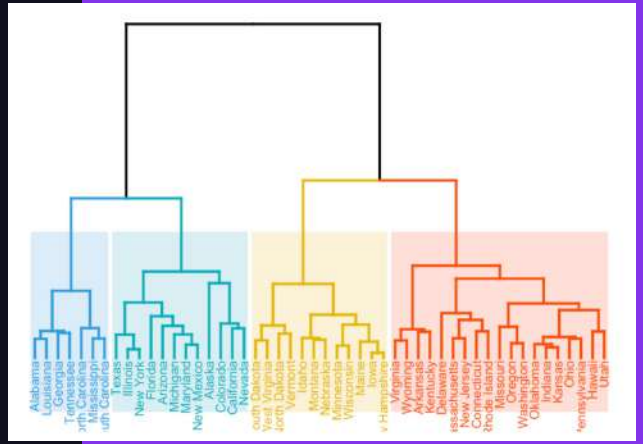
O que é o algoritmo de clusterização hierárquica aglomerativo

Etapas

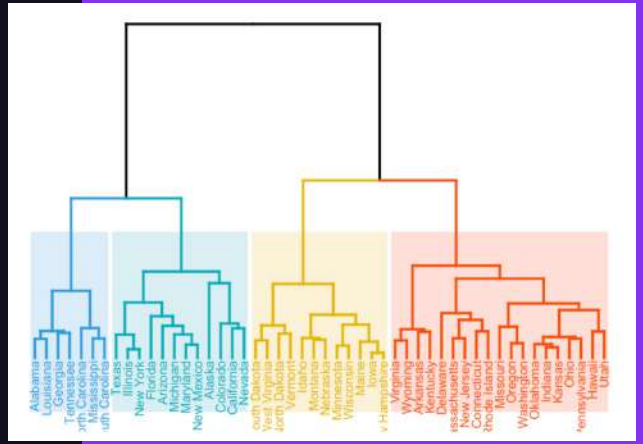
4) Atualização da Matriz de Distância: Após cada fusão, é necessário atualizar a matriz de distância para refletir a distância entre o novo cluster formado e os demais clusters.

5) Repetição: O processo de encontrar os clusters mais próximos, fundi-los e atualizar a matriz de distância é repetido até que todos os pontos de dados estejam no mesmo cluster

6) Construção do Dendrograma: Ao longo do processo, é possível construir um dendrograma, para visualizar a formação de clusters e decidir sobre o ponto de corte que determina o número final de clusters.



O que é o algoritmo de clusterização hierárquica divisivo



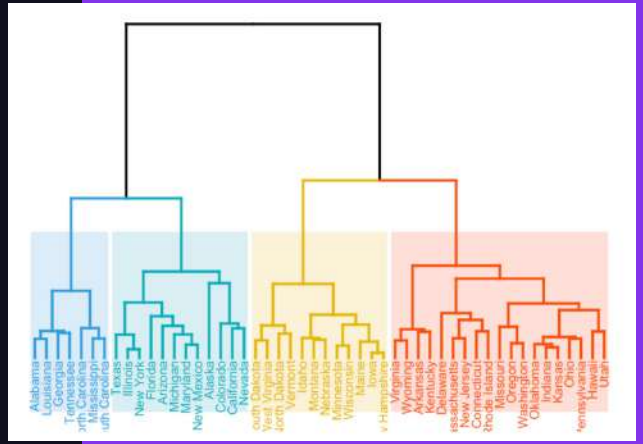
Etapas

1) Inicialização: Comece com um único cluster que inclui todos os pontos de dados. Este é o cluster de nível mais alto.

2) Cálculo da Matriz de Distância: Em cada etapa, escolha um cluster para ser dividido. A escolha pode ser baseada em vários critérios, como o tamanho do cluster, a heterogeneidade interna, ou uma métrica específica que indique a "divisibilidade" do cluster.

3) Identificação do Ponto de Corte: Determine como dividir o cluster escolhido em dois subclusters.

O que é o algoritmo de clusterização hierárquica divisivo



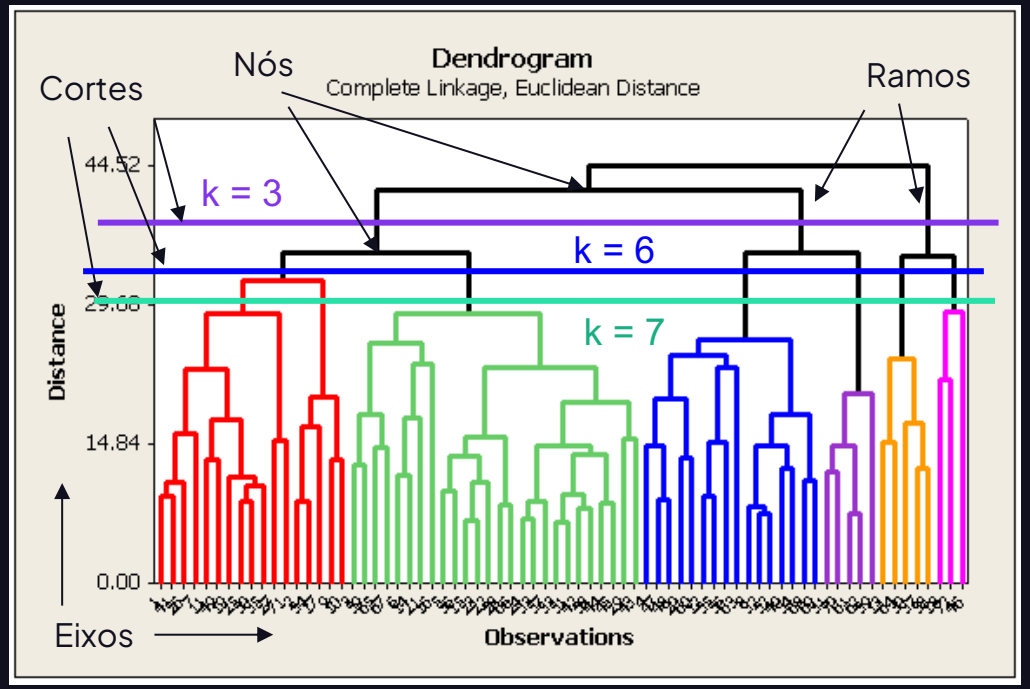
Etapas

4) Execução da Divisão: Aplicar o critério de divisão para separar o cluster em dois. Pode se usar K-Means ou DBSCAN pra isso.

5) Repetição: Continue o processo de selecionar e dividir clusters até que cada ponto seja seu próprio cluster ou até que se atinja o número desejado de clusters

6) Construção do Dendrograma: Ao longo do processo, é possível construir um dendrograma, que ilustra como os clusters são divididos progressivamente.

O que é um dendrograma



O que é um dendrograma

Aplicações do Dendrograma

Análise Exploratória: O dendrograma ajuda os analistas a entenderem a estrutura dos dados, visualizando como os grupos são formados e quão semelhantes ou diferentes eles são entre si.

Determinação do Número de Clusters: É uma ferramenta crucial para decidir o número adequado de clusters ao realizar a clusterização hierárquica, permitindo ajustar o nível de granularidade da análise.

Identificação de Outliers: Outliers ou pontos anômalos muitas vezes aparecem como ramos isolados no dendrograma, facilitando sua identificação e análise.

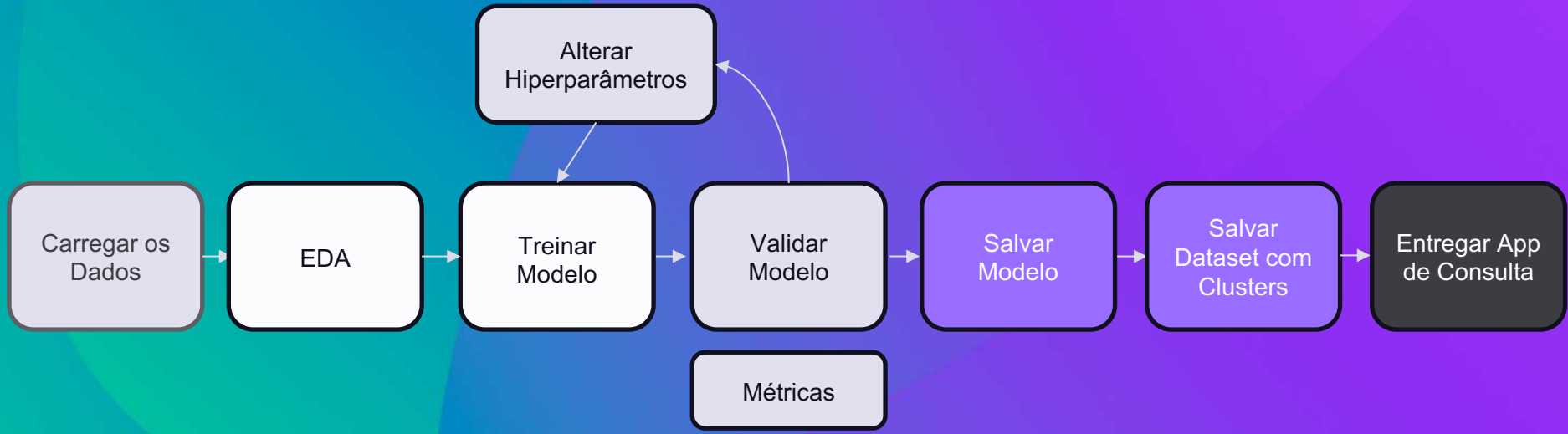
Projeto – Clustering Hierárquico

Um Marketplace especializado oferece um amplo catálogo de laptops e notebooks de diversas configurações e marcas, mantendo um estoque limitado para pronta entrega. Quando um item não está em estoque, o pedido junto ao fabricante resulta em atrasos, o que pode frustrar os clientes.

Para mitigar esse problema, o Marketplace planeja implementar um sistema de recomendação que sugira alternativas com configurações similares aos produtos indisponíveis.

Para isso, desenvolveremos um algoritmo de clusterização hierárquica para agrupar equipamentos semelhantes com base em suas especificações. Além disso, uma interface de consulta baseada na clusterização será criada, permitindo aos clientes encontrar facilmente produtos alternativos dentro do mesmo grupo do item selecionado.

Estrutura do Projeto



Code Time ...

rocketseat.com.br