MÓDULO:
Clustering - Modelos Não Supervisionados

Todos os exercícios e colabs do módulo podem ser acessados

CLICANDO AQUI

Obs: os mesmos exercícios e colabs acima seguem anexados em cada aula ao longo do módulo.

data expert.®

Data Science & Machine Learning

Boas vindas ao módulo

Consultor: Tulio Souza

Túlio Souza

Data Cientist
@Avenue Code

Consultor de Projetos de Machine Learning no Mercado Nacional e Internacional

Co-fundador da comunidade Machine Learning Experience

Machine Learn Experience: Milhares de pessoas impactadas com projetos em mais de 50 eventos.

data expert.®

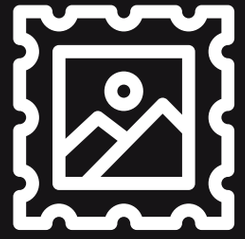
Data Science & Machine Learning

Recapitulando conceitos e objetivos do módulo

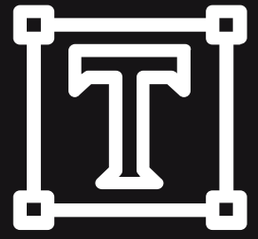
Consultor: Tulio Souza

Recapitulando

Tipos de dados



Texto



Imagem

Dados Tabulares

Recapitulando Problemas

Texto	Imagem	Dados Tabulares

Classificação	Recomendação	Regressão	Clustering

Recapitulando

Abordagens

Texto

Imagem

Dados Tabulares

Classificação

Recomendação

Regressão

Clustering

Supervisionados

Não supervisionados

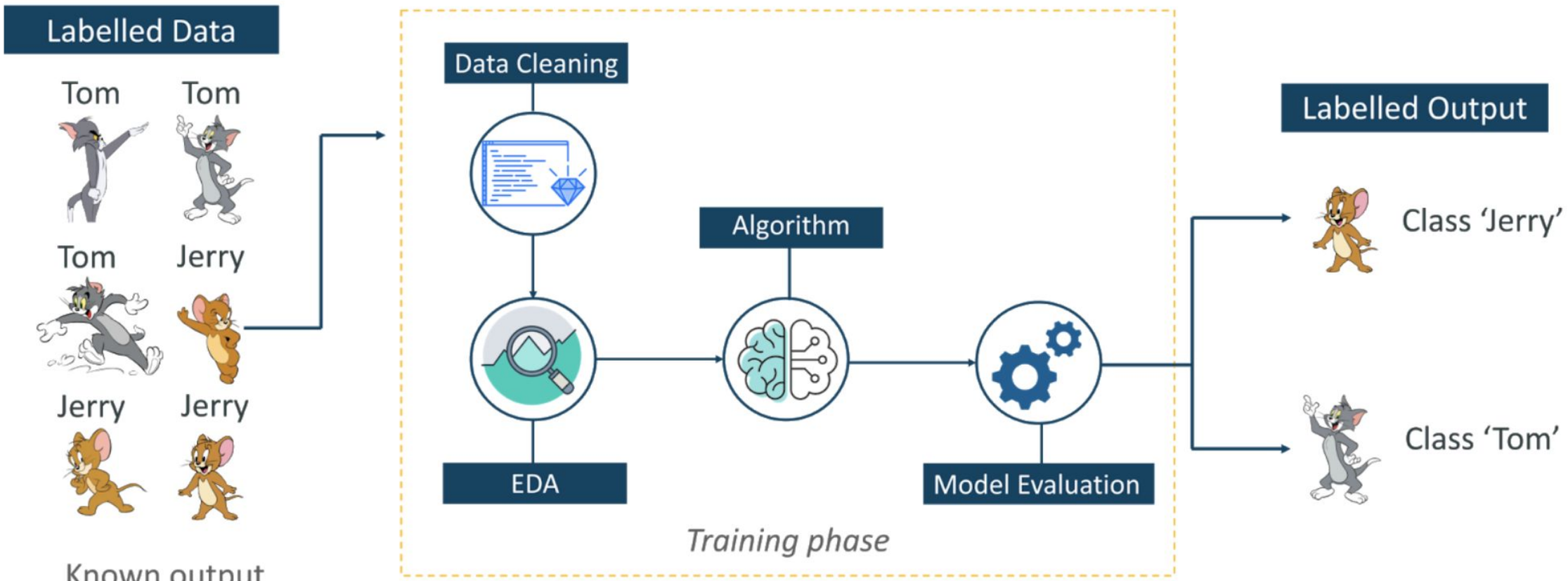
Apredizagem por reforço

Classificar

Tom e Jerry

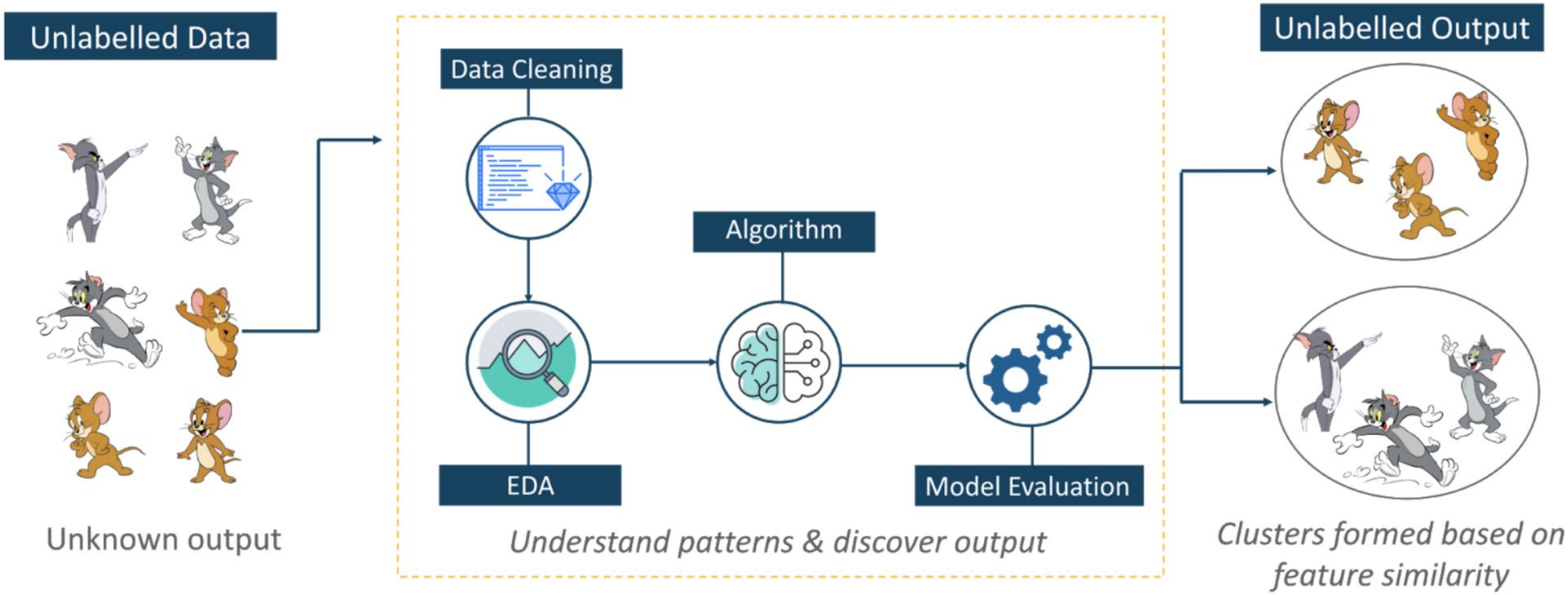
Classificar

Tom e Jerry



Classificar

Tom e Jerry



Objetivos do Módulo

1. Técnicas não supervisionadas: Kmeans, DBScan, Hierarchical Clustering, Mean shift e Gaussian Mixture.
2. Aplicar técnicas de agrupamento por aprendizagem não supervisionada em diferentes texto, imagens e dados tabulares.
3. Entender como estas técnicas podem ser utilizadas em cenário real.

data® expert.

Data Science & Machine Learning

Introdução aos problemas de Clusterização

Consultor: Tulio Souza

O que veremos neste aula:

01

O que é Análise de Cluster?

02

Análise de Cluster x Tipos de Análises

03

Problemas Comuns na Indústria

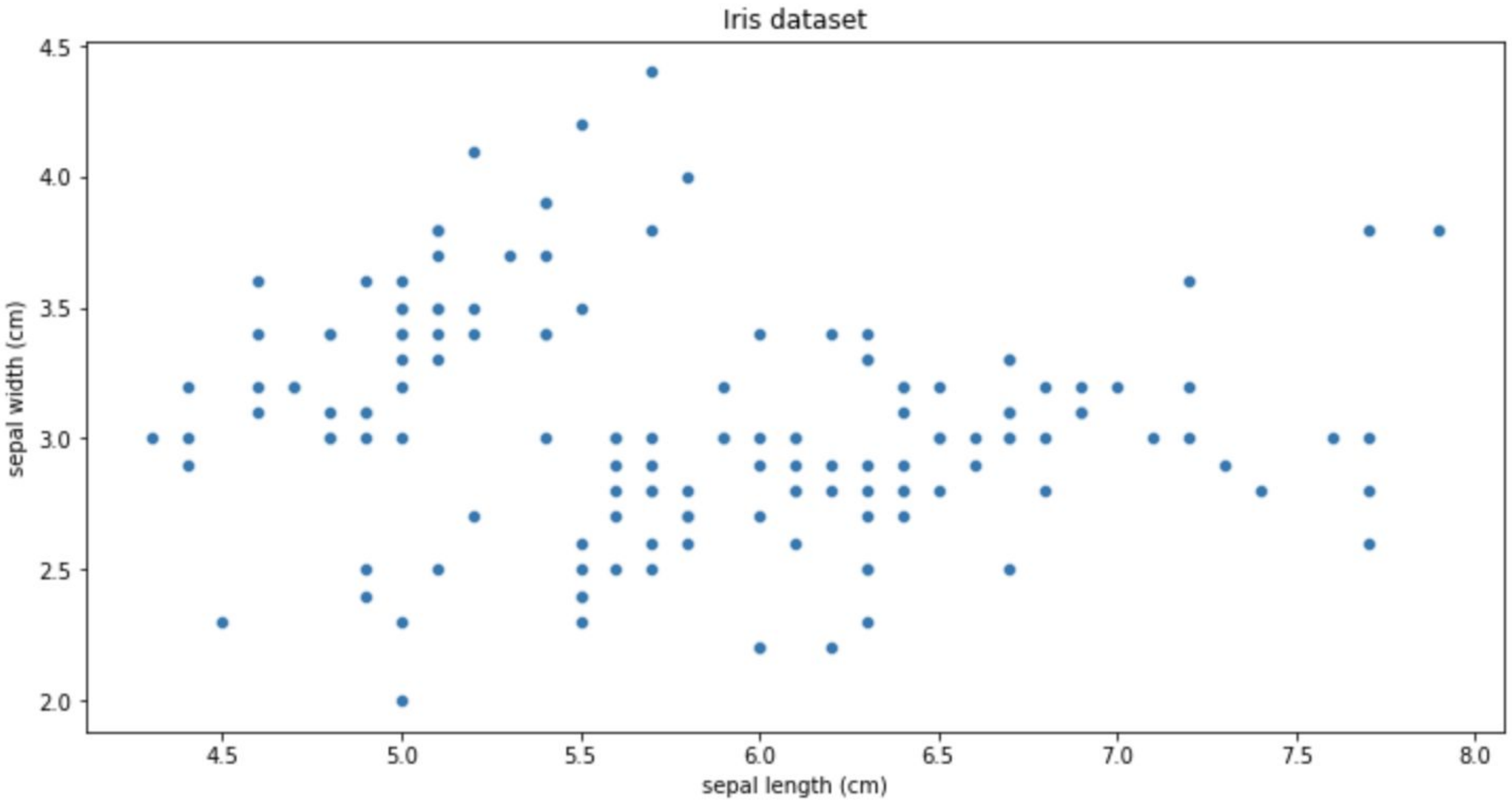
O que é Análise de Cluster

- Clustering é o conjunto de técnicas de mineração de dados que visa fazer agrupamentos de dados segundo o seu grau de semelhança.

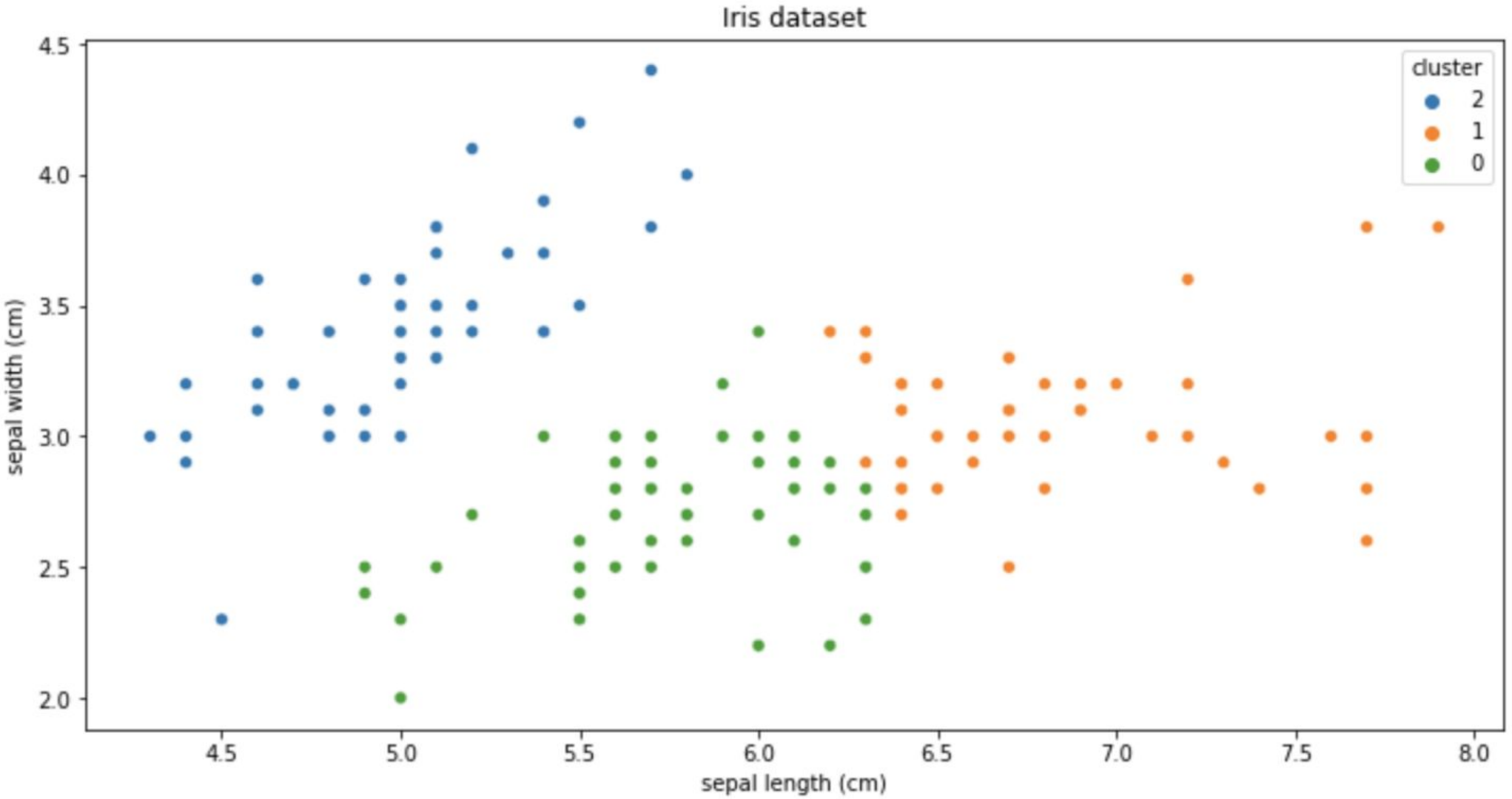
O critério de semelhança faz parte da definição do problema e do algoritmo.

A cada conjunto de dados do processo dá-se o nome de agrupamento (cluster).

Dados Brutos



Dados Clusterizados



Tipos de Análises

Passado

Futuro

Descritiva
“O que?”

Preditiva
“O que vai acontecer?”

Diagnóstica
“Por que?”

Prescritiva
“O que fazer?”

Tipos de Análises

Passado

Descritiva
“O que?”

Diagnóstica
“Por que?”

A análise de cluster ajuda a descrever o que aconteceu no passado e agrupar observações com características/comportamento similares.

Tipos de Análises

Futuro

Preditiva
“O que vai acontecer?”

- Análise de clustering é utilizada para gerar as labels (target) do modelo em uma etapa intermediária do supervisionamento.
- Também pode ser utilizada como feature para ajudar o modelo a entender a melhor o agrupamento dos dados.

Tipos de Análises

Futuro

Prescritiva
“O que fazer?”

- Análise de clustering é comumente aplicada em sistemas de recomendação para agrupar indivíduos de comportamentos similares ou fazer ofertas personalizadas.

Problemas comuns Na Indústria

Segmentação de cliente

Agrupar documentos

Agrupamento de performance

Problemas envolvendo geolocalização

Recapitulando

- **Análise de cluster pode ser aplicada em todos tipos de análise no seu dia a dia.**
- **Análise de cluster permeia as fases de modelagem ou preparação dos dados, depende do problema que estamos trabalhando.**

data expert.®

Data Science & Machine Learning

Clustering - Cases

Consultor: Tulio Souza

Uber Hack 2019

ICarros

AvenueCode

End-to-end digital transformation solutions across every vertical. US, Brazil, Canada, & the Netherlands.

iCarros

iCarros is one of the largest car marketplace in Brazil.

INDUSTRY
Marketplace

PRACTICE AREA & SCENARIO
Cloud

SOLUTION
Smart Analytics

TECH STACK
Google Cloud AI Platform, Machine Learning, Google Cloud, Exploratory Data Analysis, Python, Scikit-Learn, Pandas, Matplotlib, Numpy

Avenue Code + iCarros

Marketing Analytics by State - Clustering model

MULTI-CLOUD

Opportunity

As a car marketplace, iCarros needs to balance how much people are buying and selling cars inside their platform. Several databases tell us information about the process but they were not integrated.

Solution

- Structuring Google Analytics 360 and AWS Data Lake(S3) information using BigQuery as Data Warehouse
- Integration of data sources in a new table in granularity of the business problem
- Creation of dashboards containing descriptive analysis on the newly created table
- Modeling using unsupervised learning(K-means) to create a cluster of states with similar characteristics

Results

+ Better Insights

Deep understand about product performance in different states in Brazil.

+ Analytical Maturity

More data visualization and information sharing between different teams.

+ Data Driven Decisions

Possibility to create reliable strategies that were founded on data.

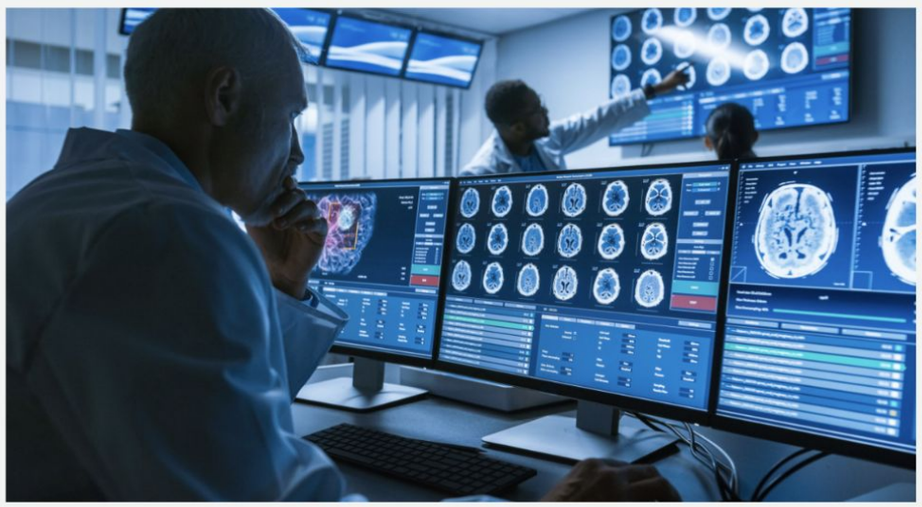
Hermes Pardini

HERMES PARDINI
Medicina, Saúde e Bem-estar

Hermes Pardini utiliza tecnologia AWS para otimizar envio de insumos para mais de 6 mil laboratórios conveniados

2020

Com mais de 60 anos de atuação no mercado, o Laboratório Hermes Pardini é hoje referência no segmento de Apoio Laboratorial, estando entre os três maiores laboratórios do país em volume de análises e em faturamento.



Temos volumes de dados exponenciais. São patamares assustadores que consomem muitos recursos de hardware e não seria razoável manter uma estrutura para realizá-los internamente. O uso dos recursos da AWS nos permite escala e disponibilidade sob medida. Usamos quando precisamos e da forma que precisamos. Conseguimos escalar quando necessário e obter os resultados que o projeto requer."

Lucas Santana
gerente corporativo de TI, responsável pela vertical Analytics do Laboratório Hermes Pardini

Recapitulando

- **Cases 1 e 2 -> Modelo Unsupervised era a entrega final.**
- **Cases 3 -> Modelo Unsupervised era usado como feature.**

data expert.®

Data Science & Machine Learning

Clustering Methods

Consultor: Tulio Souza

O que veremos neste aula:

01

Introdução

02

Centroid Clustering

03

Distribution Clustering

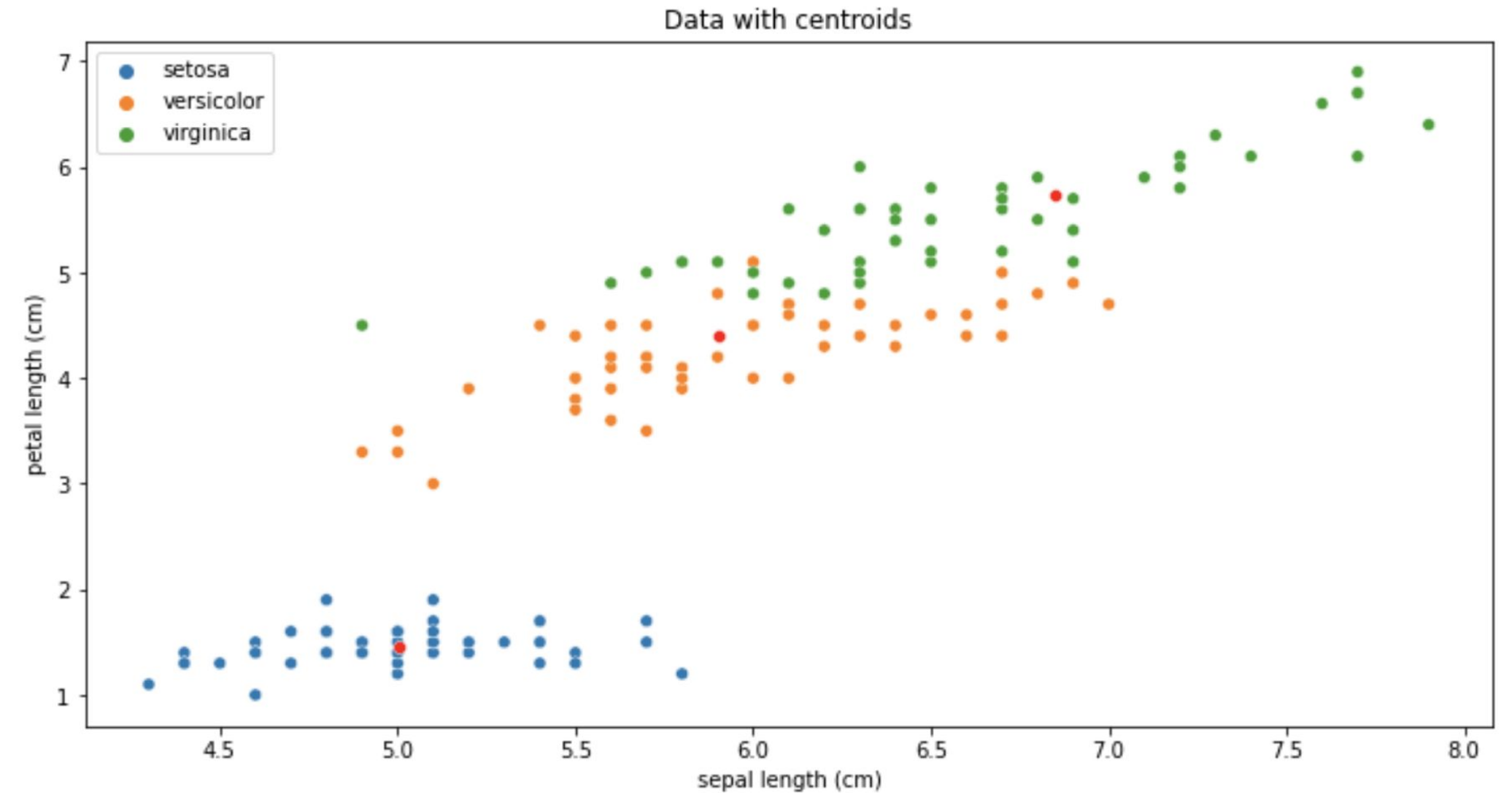
04

Density Clustering

05

Hierarchical Clustering

Centroid Clustering

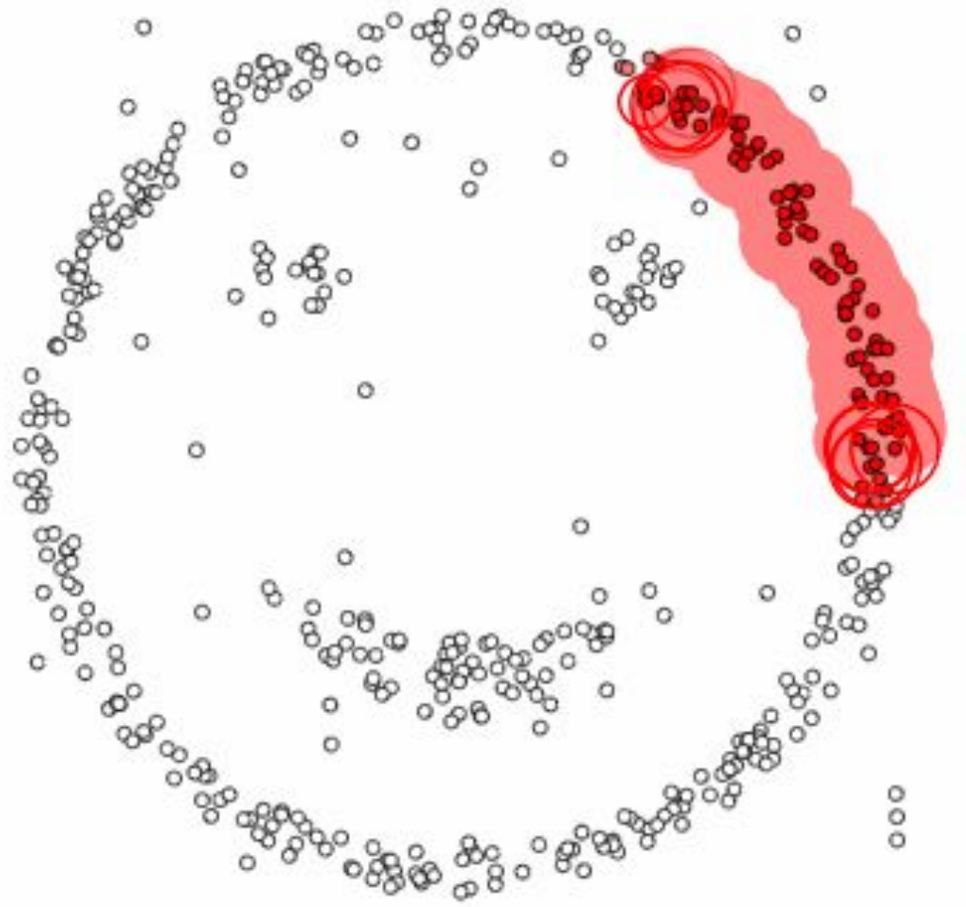


Distribution Clustering

O agrupamento baseado em distribuição está diretamente relacionado ao uso de modelos de distribuição (Ex: Gaussiano / Normal) em estatísticas.

Fundamentalmente, os clusters são definidos com base na probabilidade de os objetos incluídos pertencerem à mesma distribuição.

Distribution Clustering



epsilon = 1.00
minPoints = 4

Restart

Pause

Density Clustering

os clusters são definidos com base na identificação de áreas de maior densidade do que o que pode ser encontrado no restante do espaço de dados.

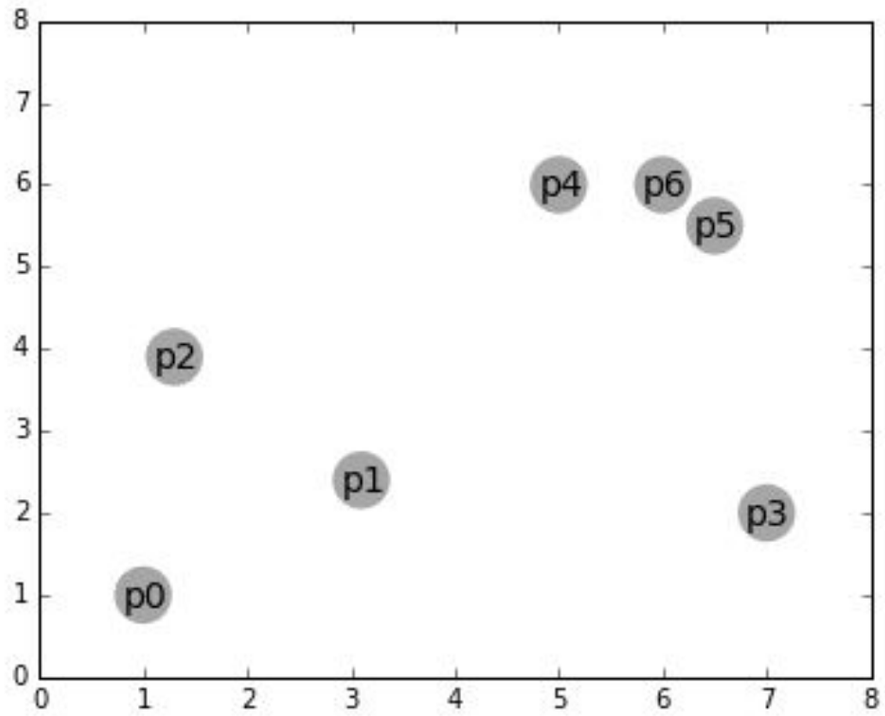
O agrupamento por densidade é capaz de lidar com o ruído se o resultado do ruído forem objetos em áreas do espaço de dados que são esparsas

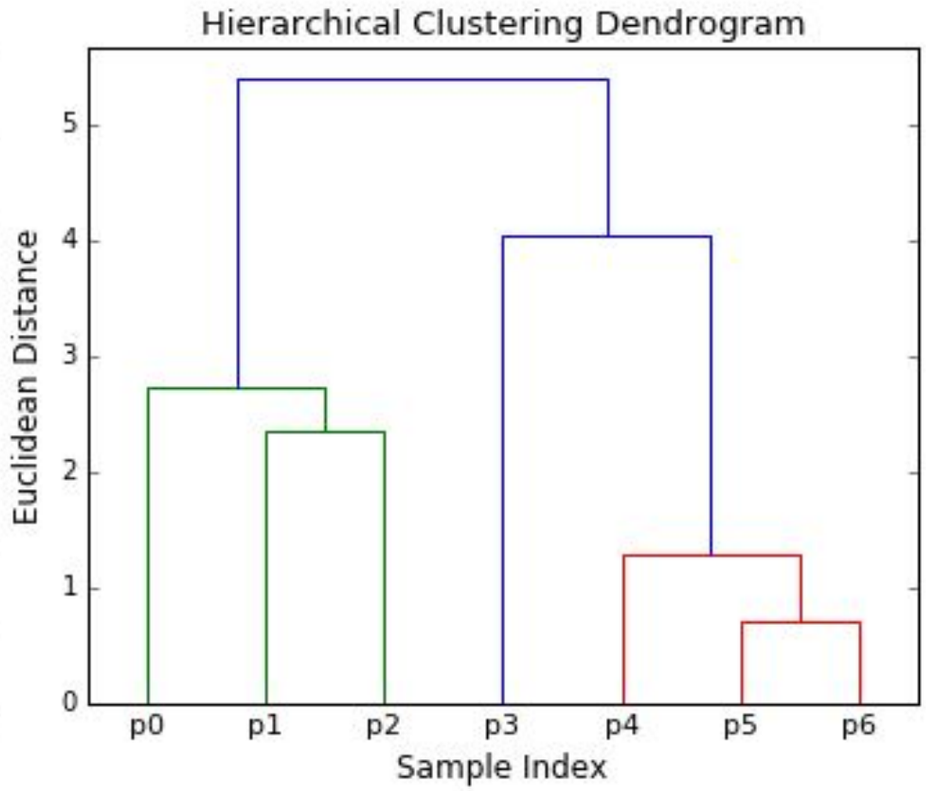
Distribution Clustering

Hierarchical Clustering

Agrupamento hierárquico, ou Hierarchical clustering no inglês, é uma técnica de clusterização de dados que baseia-se no tamanho e distância dos dados em um conjunto.

Hierarchical Clustering





data expert.®

Data Science & Machine Learning

KMeans

Consultor: Tulio Souza

O que veremos neste aula:

01

O que é Kmeans?

02

Como funciona o Algoritmo?

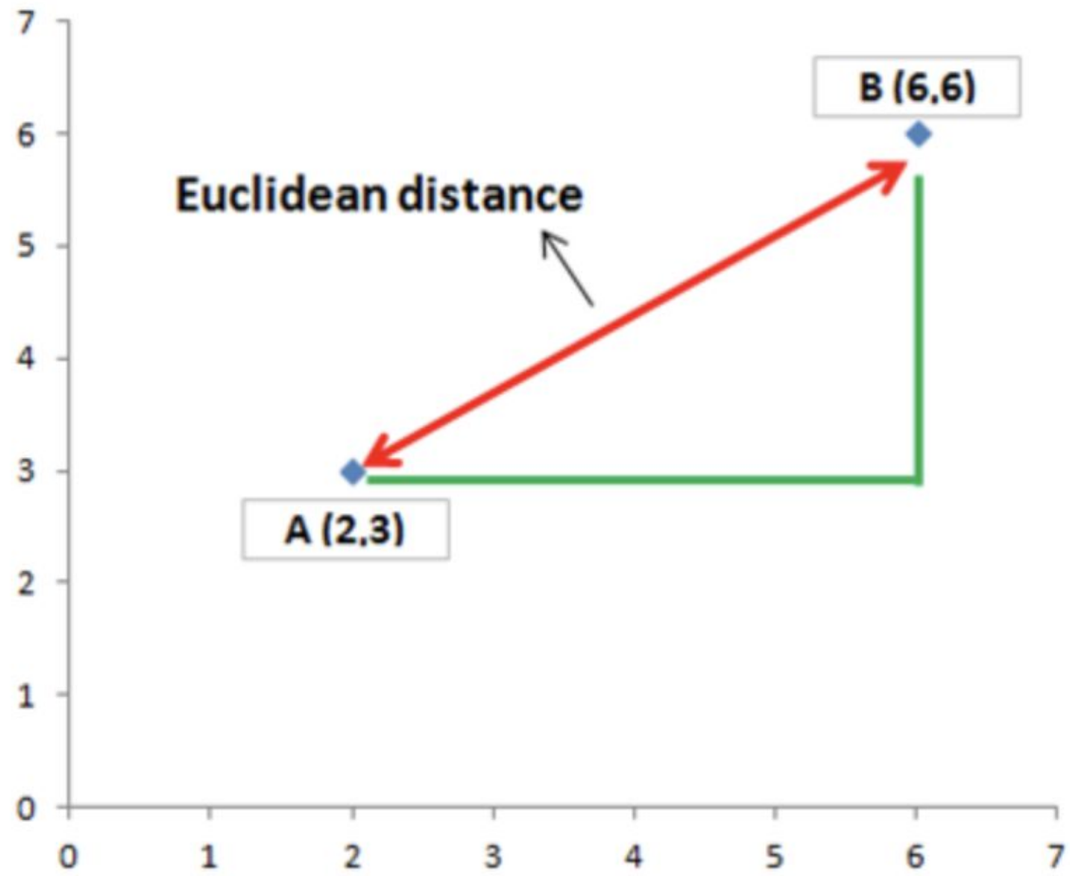
03

Prós e Contras

KMeans

- É um método de clusterização baseado em centróides, como em seu nome: “K número de centros”.
- O centro de cada cluster terá a média dos valores neste cluster.
- A tarefa do algoritmo é encontrar o centróide mais próximo há um ponto utilizando alguma métrica de distância e atribuir o ponto ao cluster.

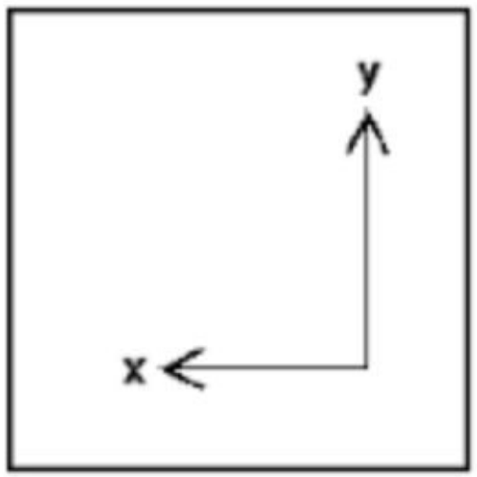
Distância Euclidiana



$$Euclidean\ distance\ (a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

Distância de Manhattan

- A distância de Manhattan é a soma das diferenças absolutas entre os pontos em todas as dimensões.
- De forma simples, a soma total da diferença entre as coordenadas x as coordenadas y.



Manhattan

KMeans - Treinamento

- Selecionamos um ‘K’, ou seja, um número de clusters.
- Inicia-se, definindo aleatoriamente, um centróide para cada cluster.
- Calcular, para cada ponto, o centróide de menor distância.

Cada ponto pertencerá ao centróide mais próximo.

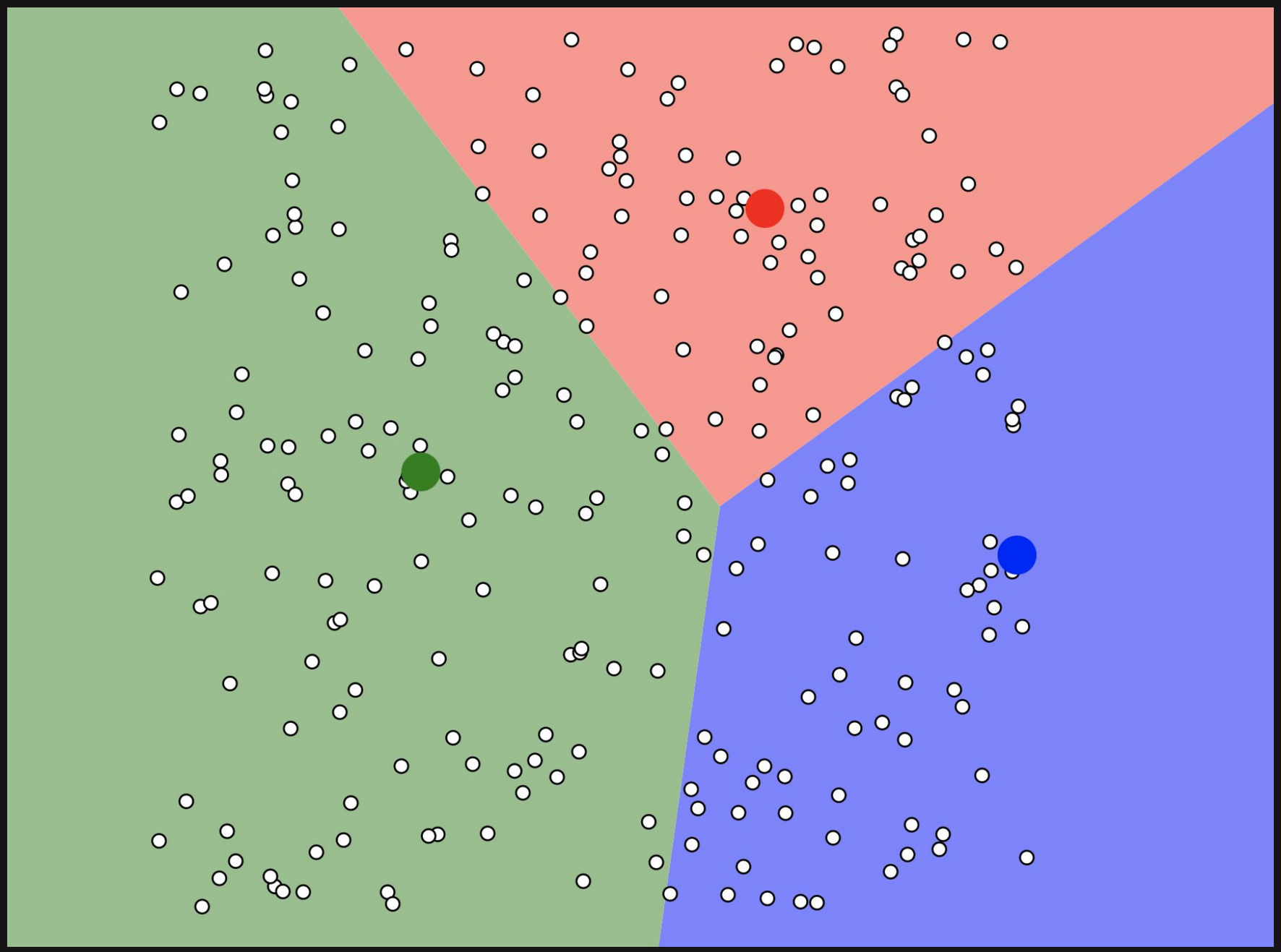
KMeans - Treinamento

- Reposicionar o centróide.

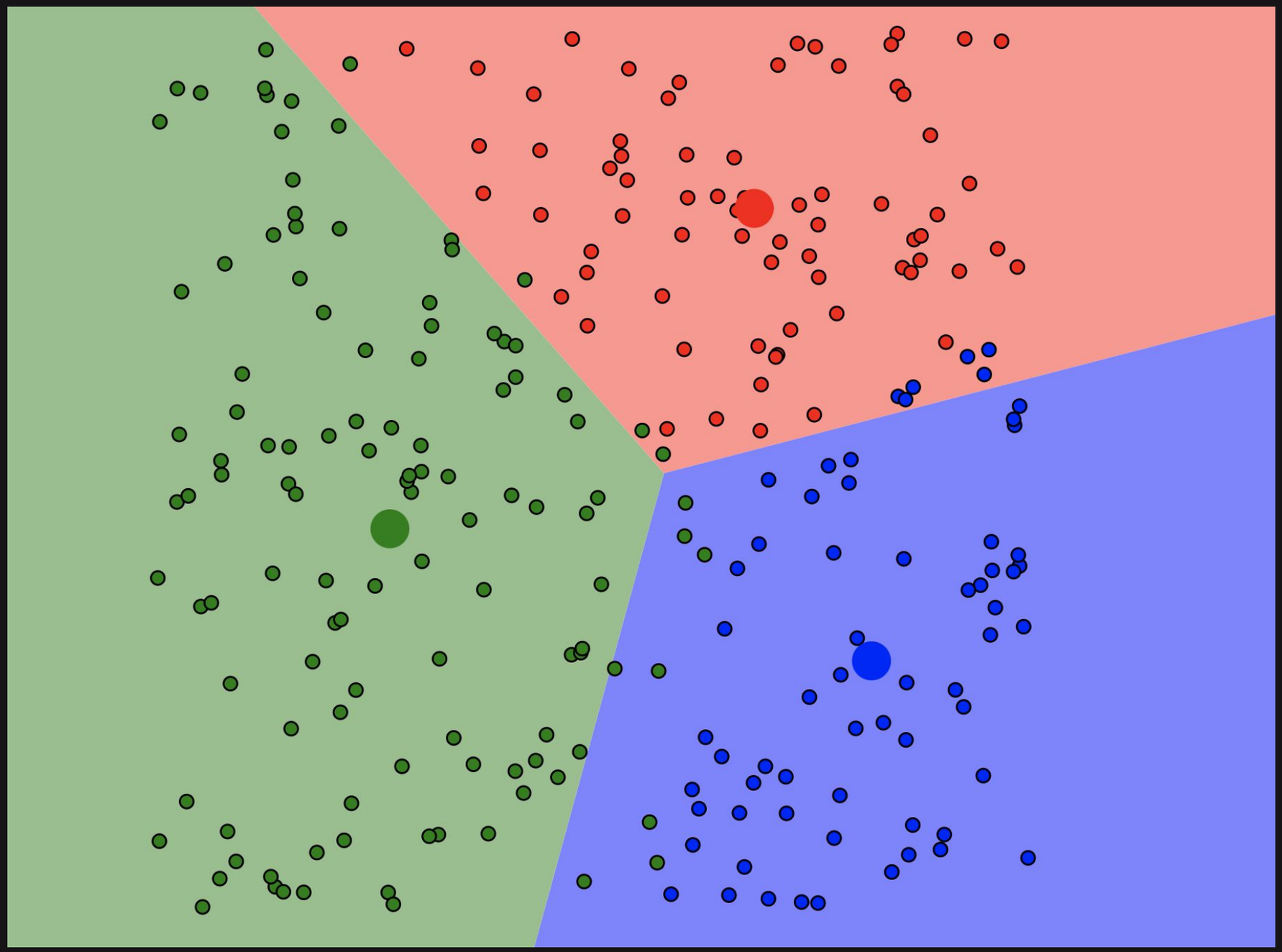
A nova posição do centróide será a média da posição de todos os pontos do cluster.

- Iterativamente, os últimos dois passos são repetidos até obtermos a posição ideal dos centróides.

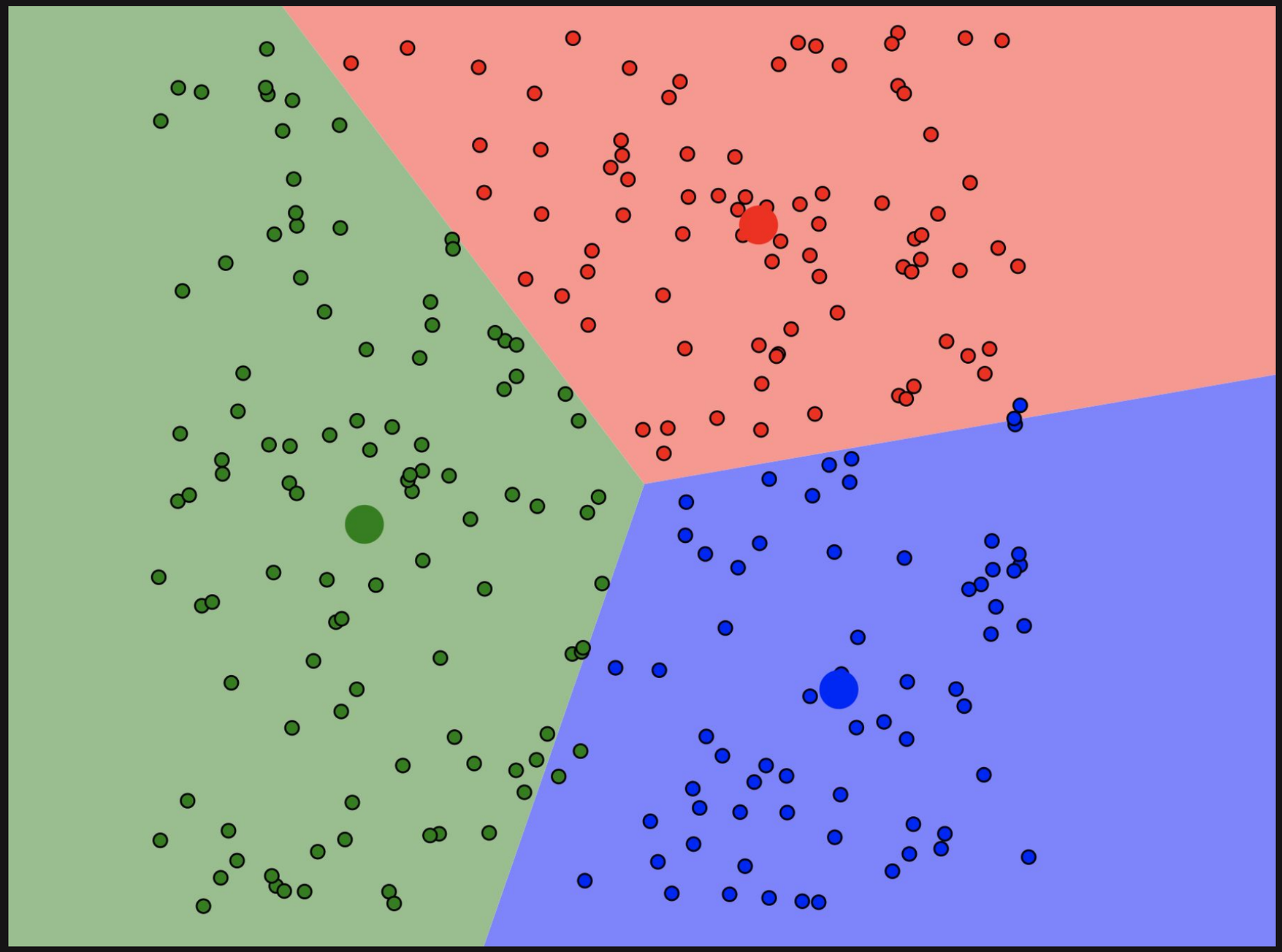
KMeans - Treinamento



KMeans - Treinamento



KMeans - Treinamento



KMeans - Inferência

- Calcula-se a distância do novo ponto para todos os centróides.
- O novo ponto pertencerá ao cluster do centróide de menor distância.

Vantagens

- É simples e intuitivo e de fácil interpretação.
- É rápido e pouco custoso computacionalmente.

Desvantagens

- Precisamos saber antes os números de clusters.
- Muito sensível a outliers.
- Não funciona bem com distribuições não convencionais.
- Tenta gerar clusters de tamanhos iguais.

Avaliando o Modelo

- Como é um problema de agrupamento não supervisionado não existe certo e errado (Clusterização != classificação)
- A interpretabilidade do modelo se dará no cruzamento das features com o target.
- Inertia: é um indicativo de quão estáveis os clusters estão. Ou seja, caso eu promova mais n-iterações, quão diferentes os clusters vão ser entre si a cada nova rodada.

Recapitulando

- **O que é KMeans;**
- **Como funciona;**
- **Avaliando o modelo.**

data expert.®

Data Science & Machine Learning

Hierarchical Clustering

Consultor: Tulio Souza

O que veremos neste aula:

01

O que é Hierarchical Clustering

02

Como funciona o Algoritmo

03

Parâmetros Principais

04

Medidas de Distâncias

05

Vantagens e Desvantagens

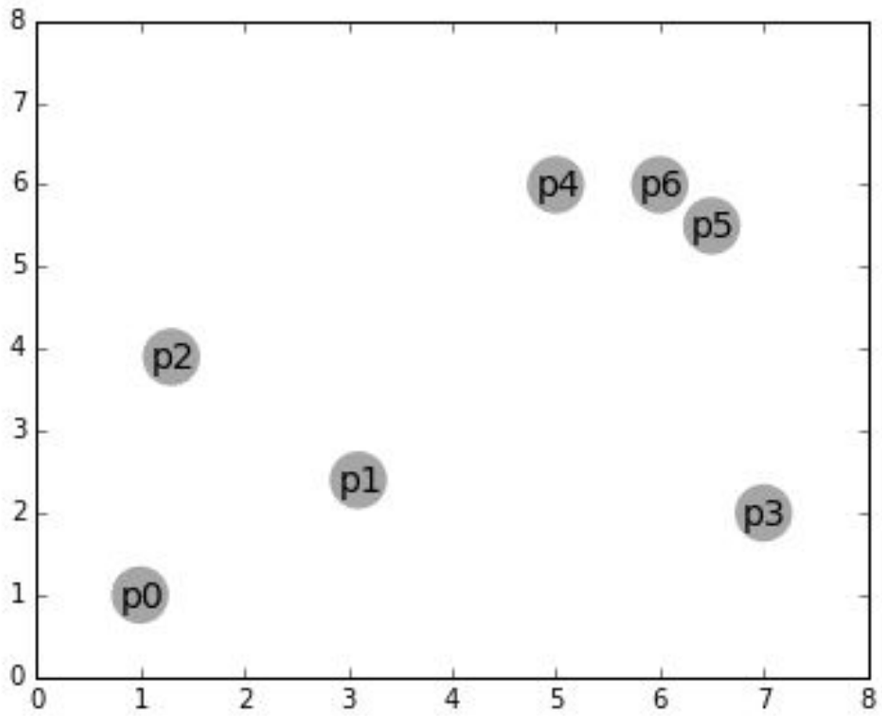
Hierarchical Clustering

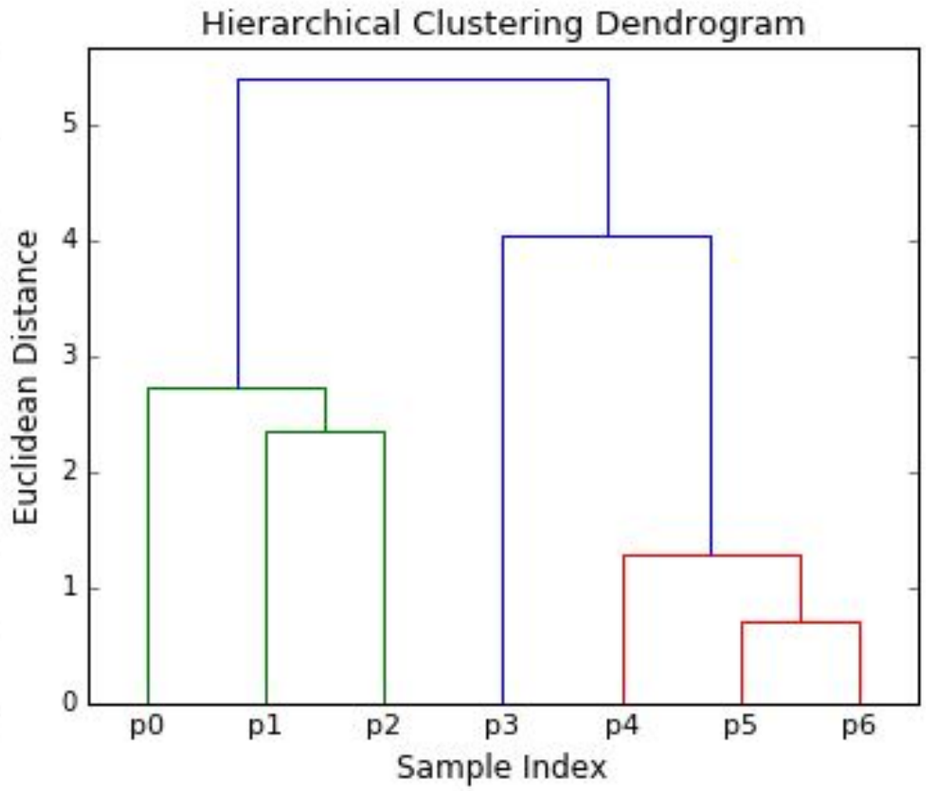
- Método de agrupamento de dados baseado em hierarquia.
- Temos duas estratégias para o clustering hierárquico: bottom up ou top down.
- Bottom up (Agglomerative): Considera-se no primeiro momento que cada observação nos dados é um cluster.
- Top down (Divisive): Considera-se no primeiro momento um único cluster e a partir daí são realizadas as divisões.

Agglomerative H. Cluster

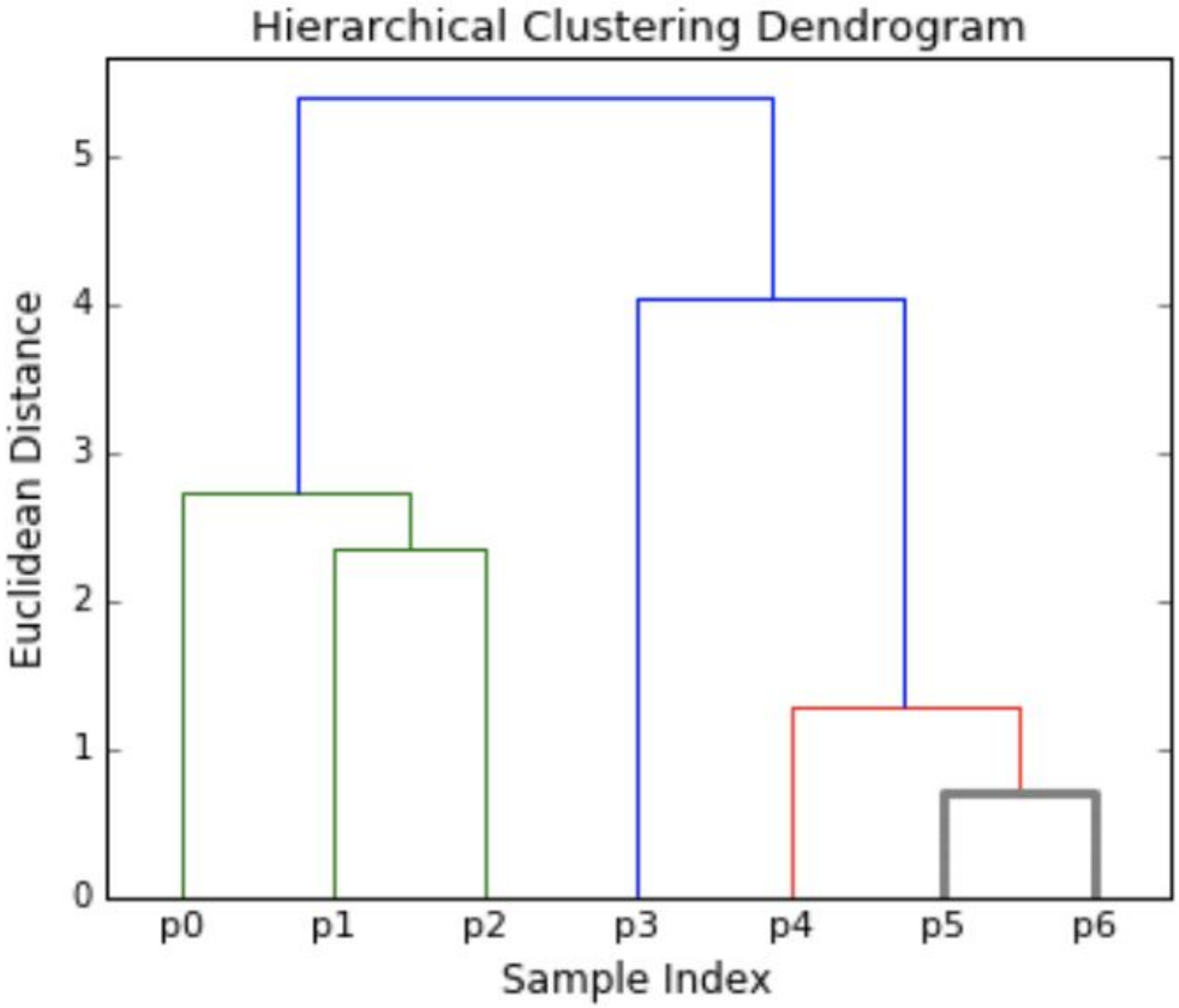
- Considera-se cada ponto como um cluster.
- Seleciona-se uma métrica de calcular as distâncias entre clusters.
- Em cada iteração combinamos os 2 clusters mais próximos, até todos os dados serem agregados.

Agglomerative H. Cluster

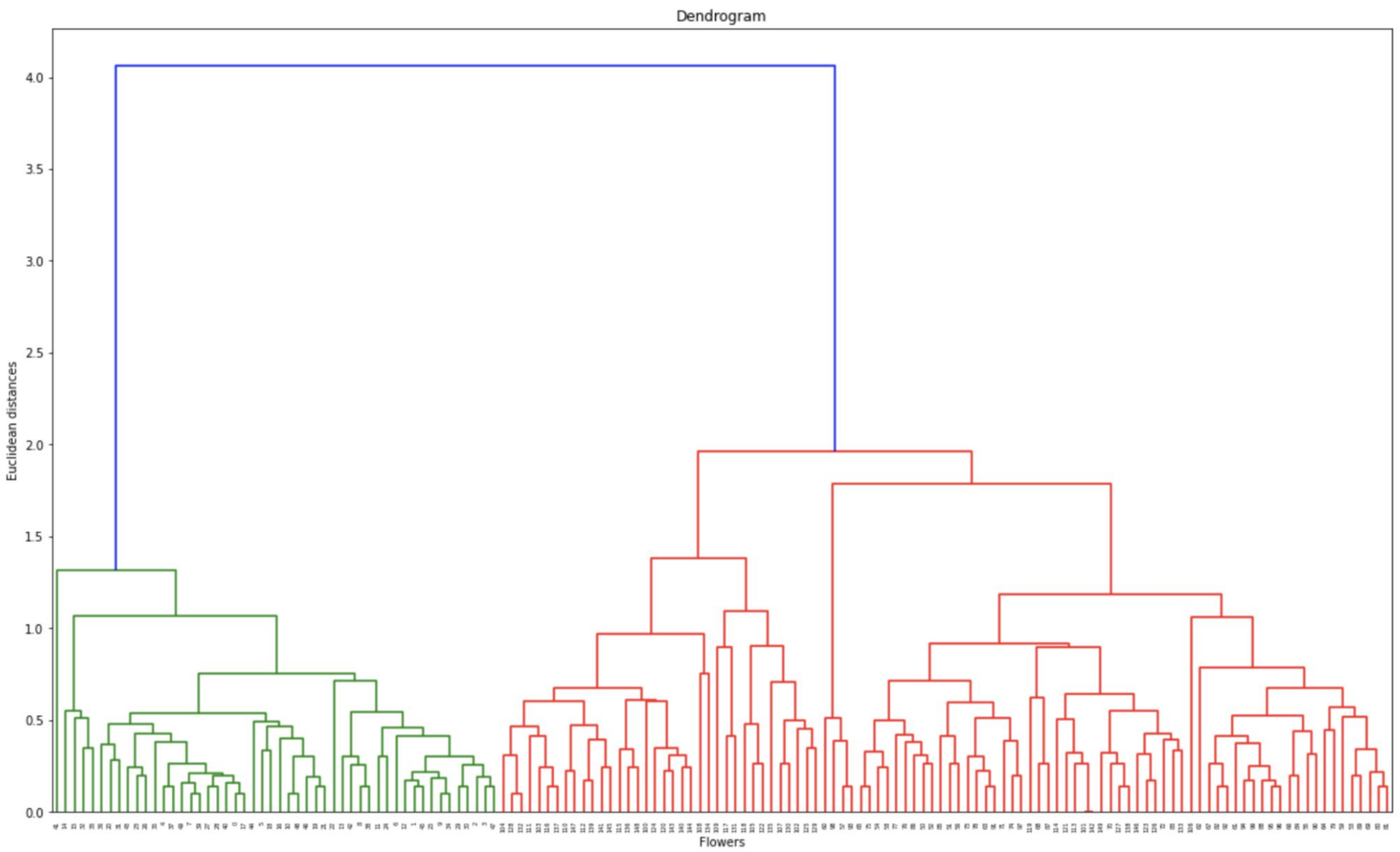




Entendendo o Dendograma



Dendograma


Dendrogram
Euclidean distances
Flowers
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Parâmetros Principais

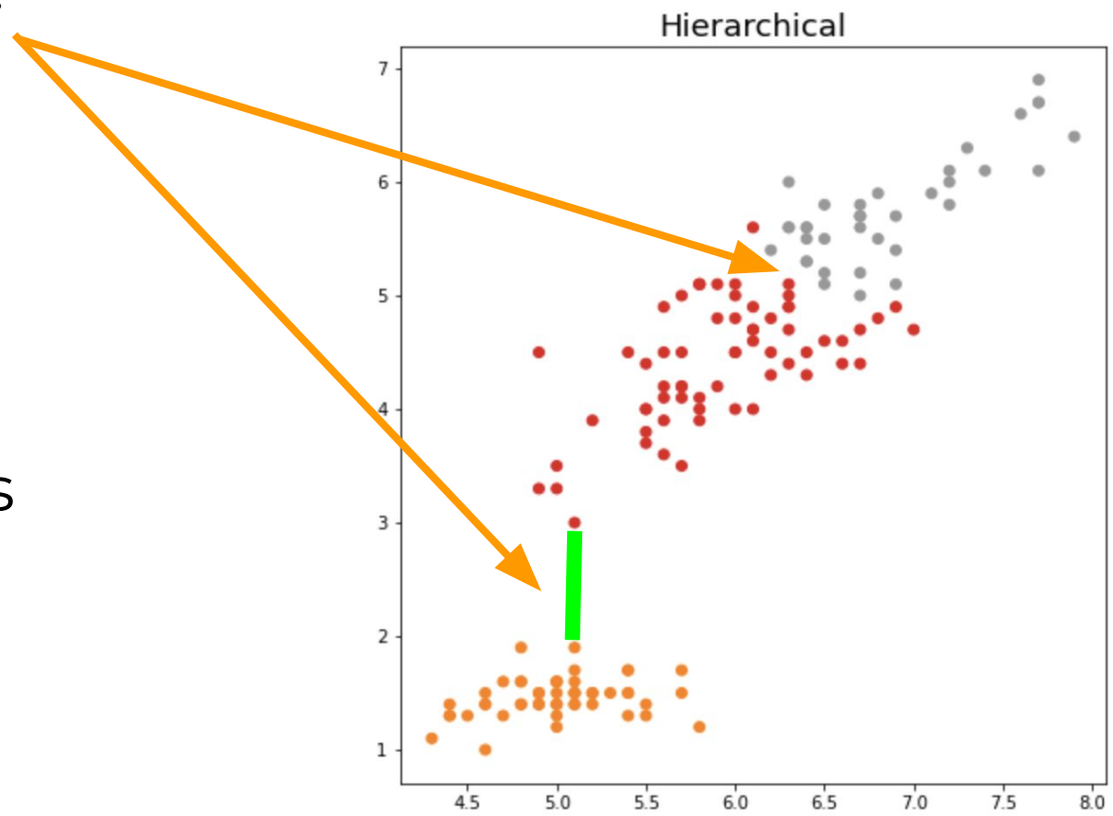
- Affinity: Medida de distância a ser considerada no cálculo de distância entre os clusters.

(Exemplo: Euclidiana, Manhattan, Cossenos)

- Linkage: Medidas de distâncias entre clusters.

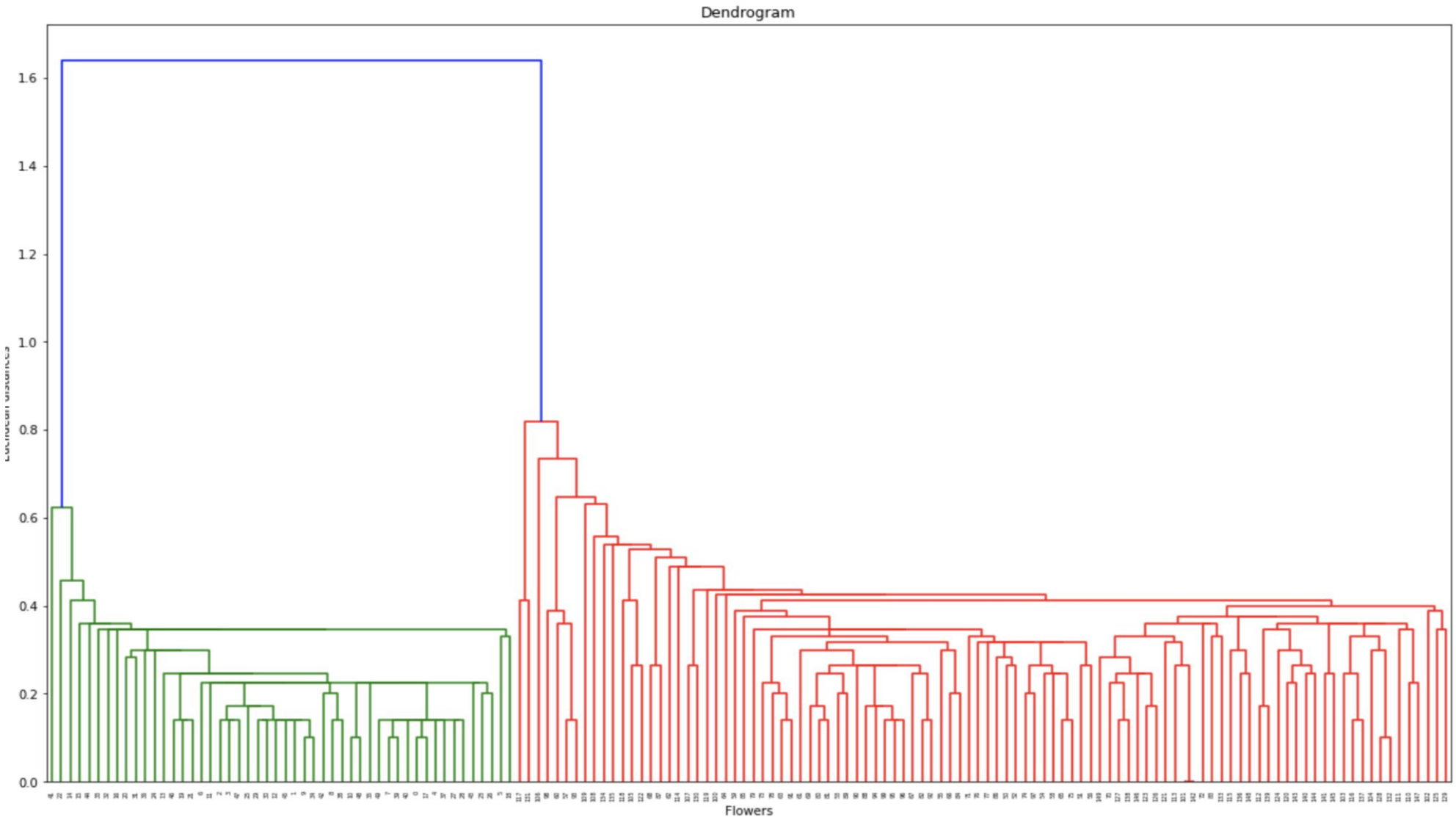
Cluster Distance Measures

- Sigle Link: Distância entre os elementos mais próximos.



- Caracteristicas: Tende a formar corrente nos nossos dados.

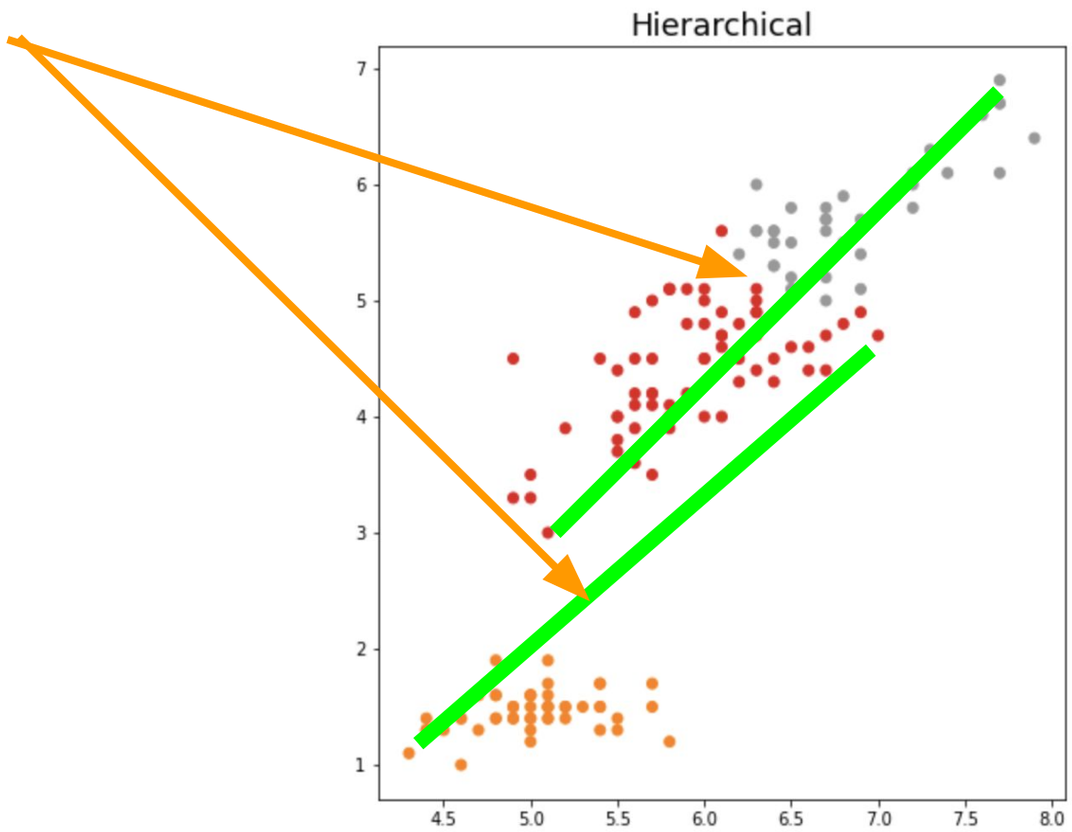
Single Link



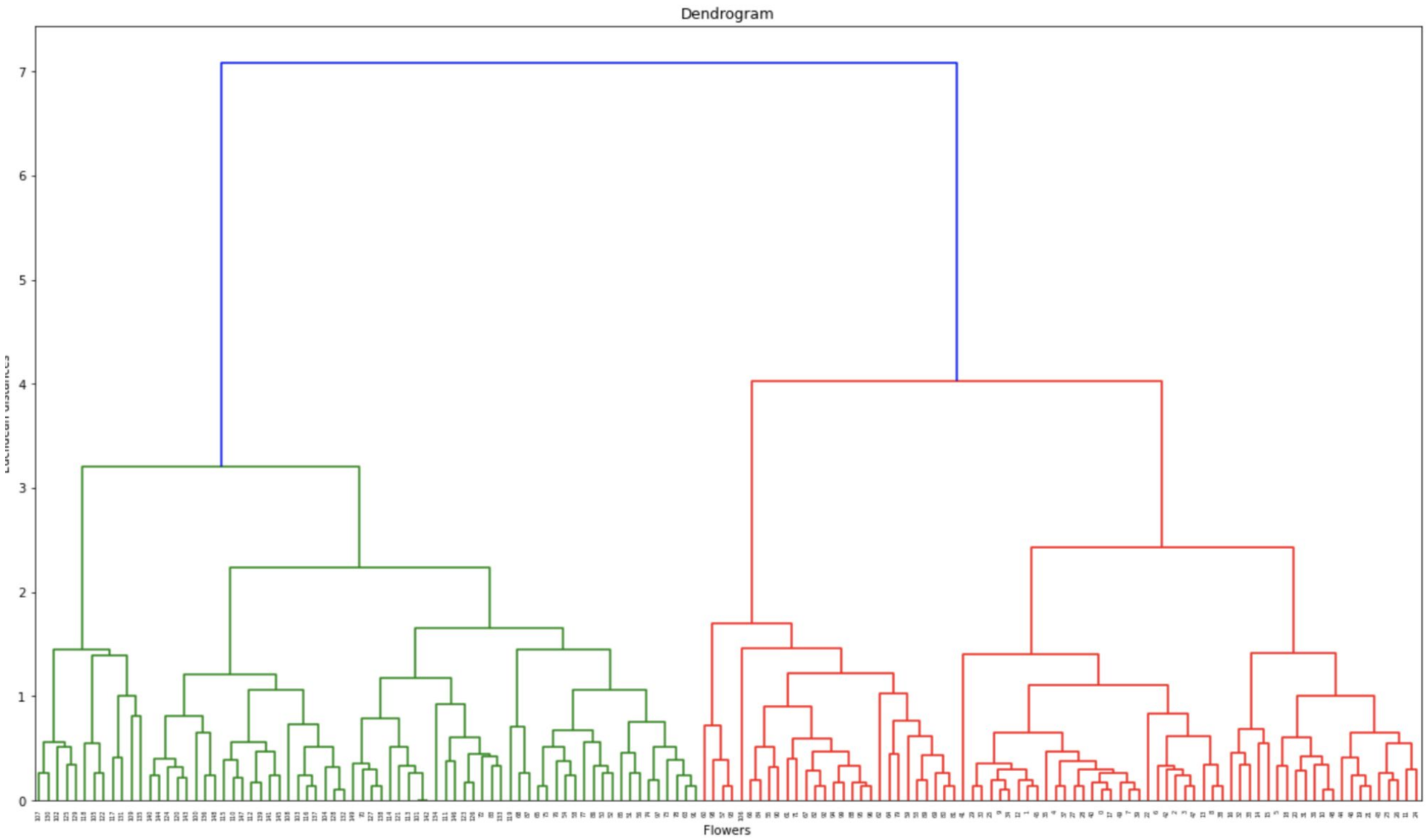
Cluster Distance Measures

- Sigle Link: Distância entre os elementos mais Longes.

- Caracteristicas: Tende a formar clusters esféricos.



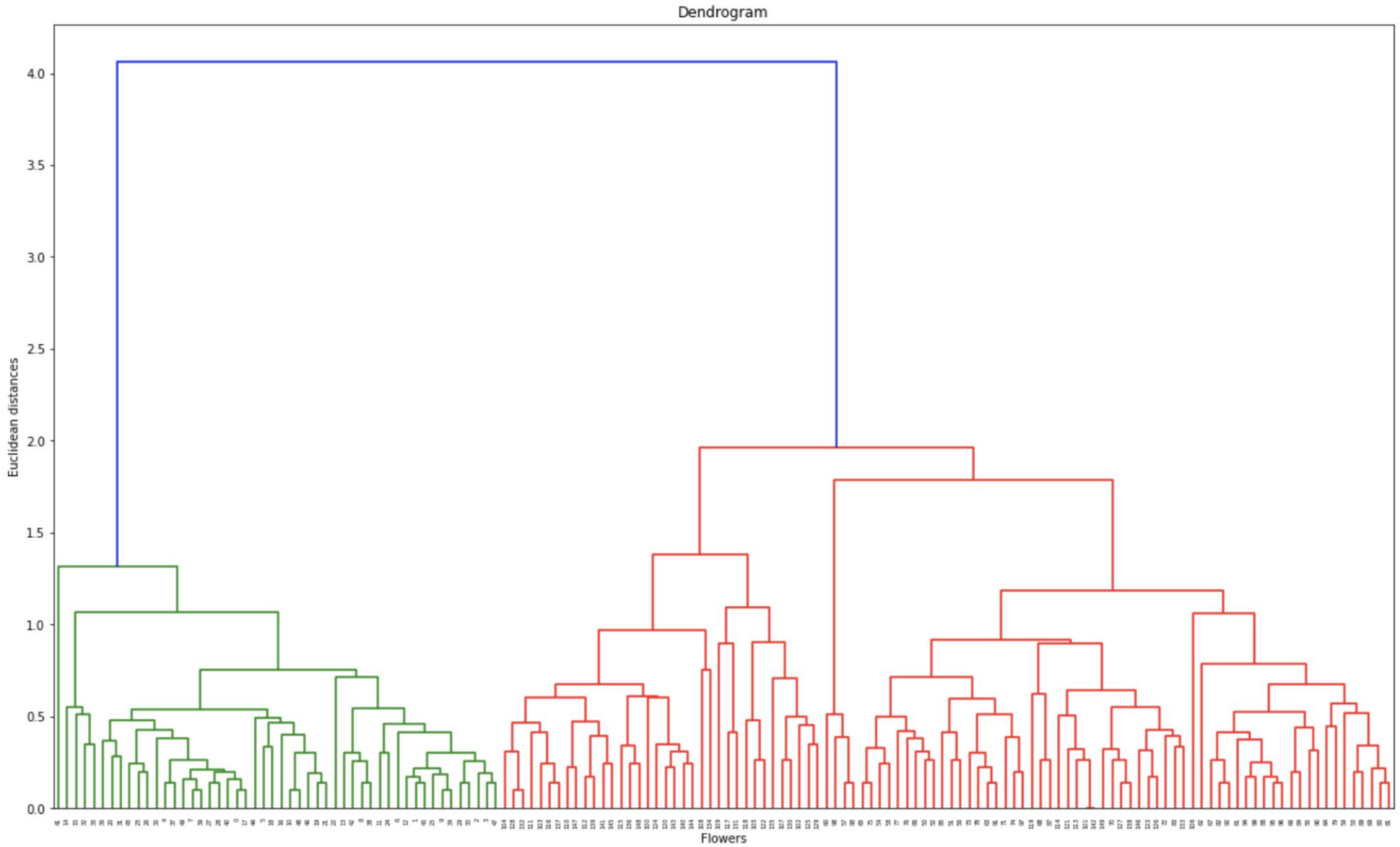
Complete Link



Cluster Distance Measures

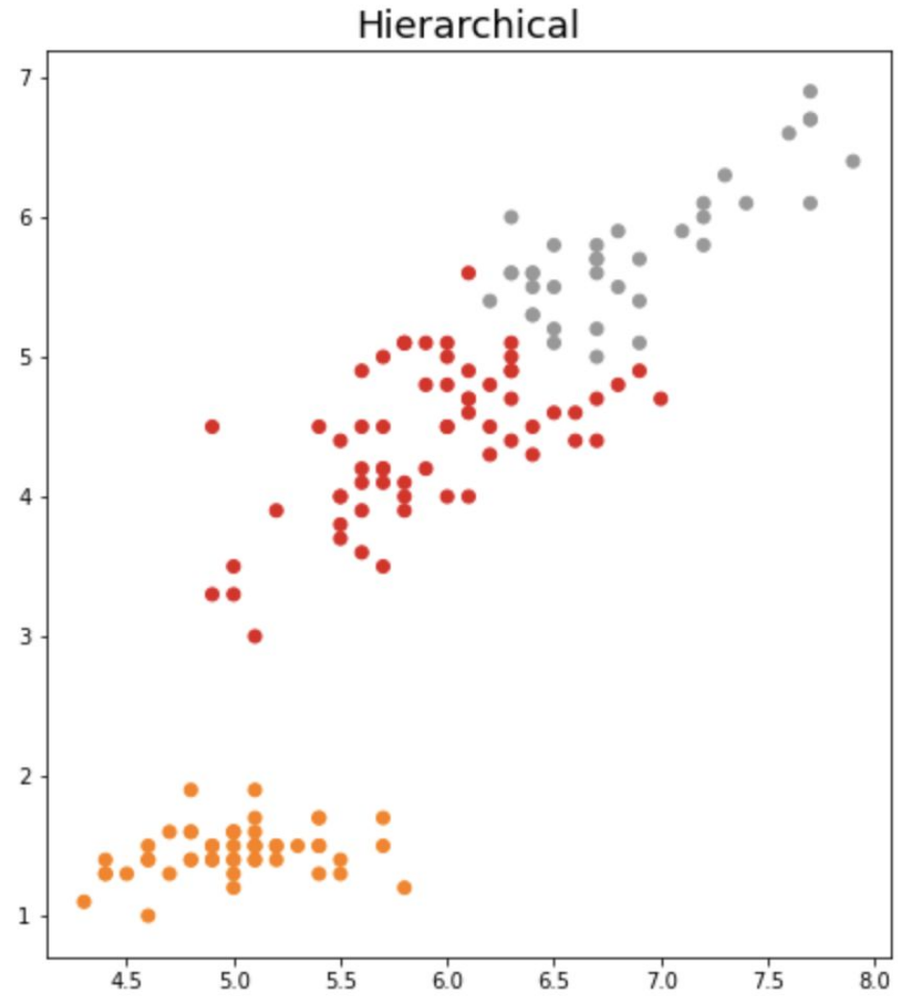
- Average Link: Distância entre todos os pares de pontos.
- Características: Menos sensível a outliers.

Average Link

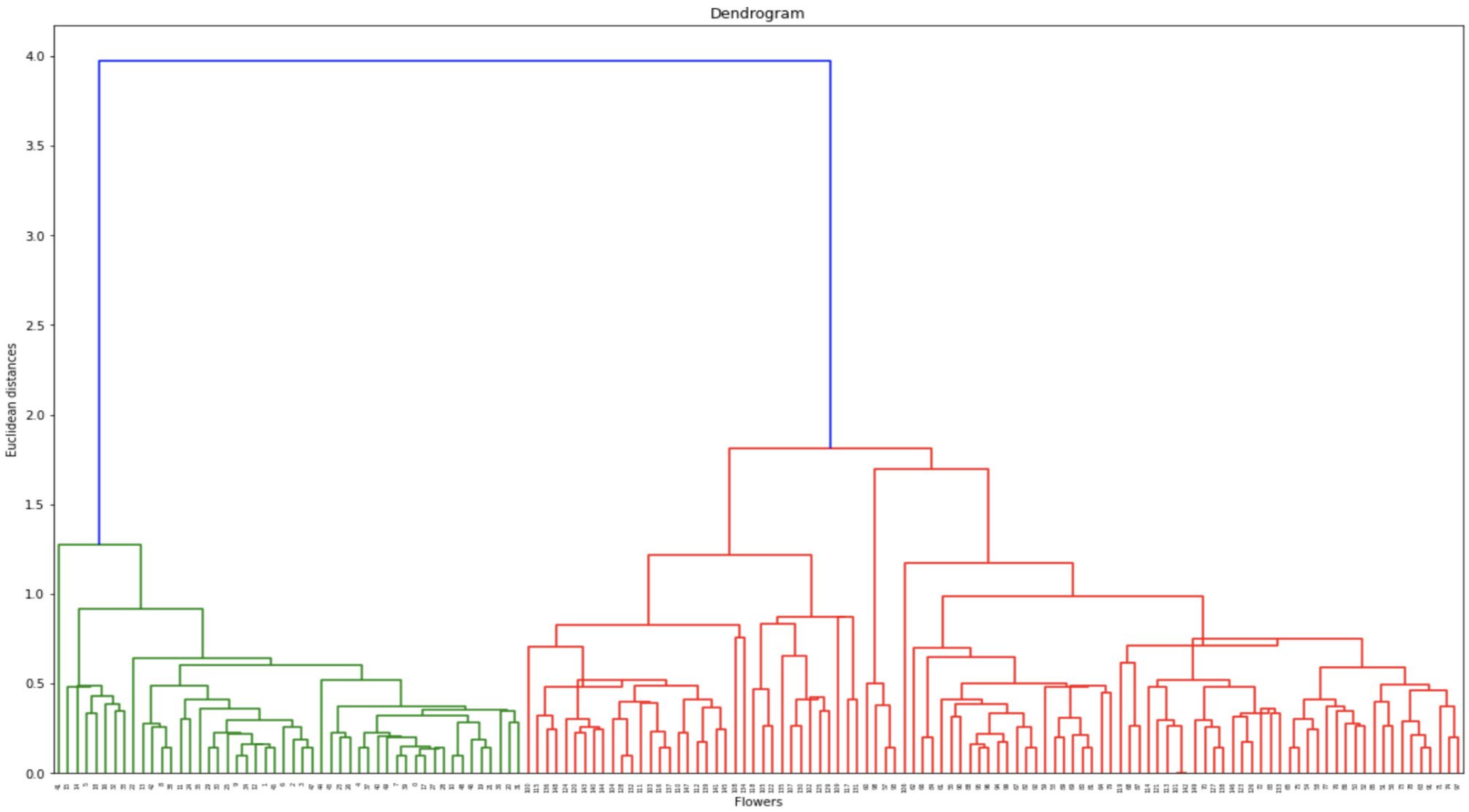


Cluster Distance Measures

- Centroid: Calcula-se o centróide dos clusters, depois juntamos os clusters com menor distância entre centróides.



Centroid Method

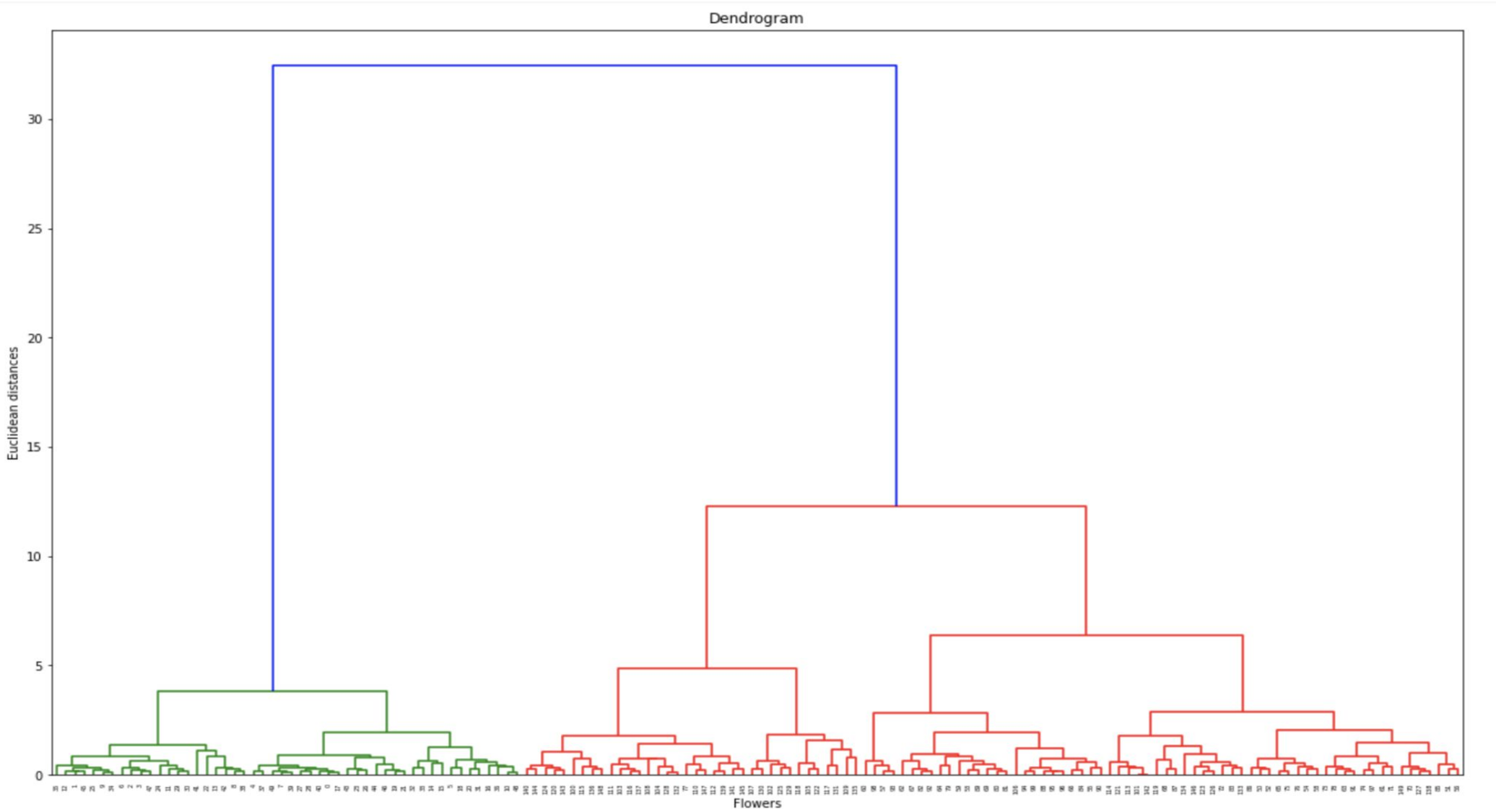


Cluster Distance Measures

- Ward Method: Para agregar dois clusters, primeiro é estima-se o centróide entre os dois clusters.

Depois calcula-se a soma dos desvios padrões de todos pontos ao centróide, escolhe-se juntar os clusters que apresenta a menor soma em desvios padrões.

Ward Method



Vantagens

- Não é necessário especificar o número de cluster que queremos com antecedência.
- Podemos escolher o caminho de cluster que faça mais sentido para o problema que estamos atacando.
- Fácil Interpretabilidade.

Desvantagens

- Custo computacional elevado.
- Difícil de visualização em dataset's muito grandes.

Recapitulando

- O que é Hierarchical Clustering
- Como funciona o algoritmo
- Medidas de distâncias entre clusters
- Vantagens/ Desvantagens

data expert.®

Data Science & Machine Learning

Fechamento Módulo

Consultor: Tulio Souza

O que veremos neste aula:

01

Centróides

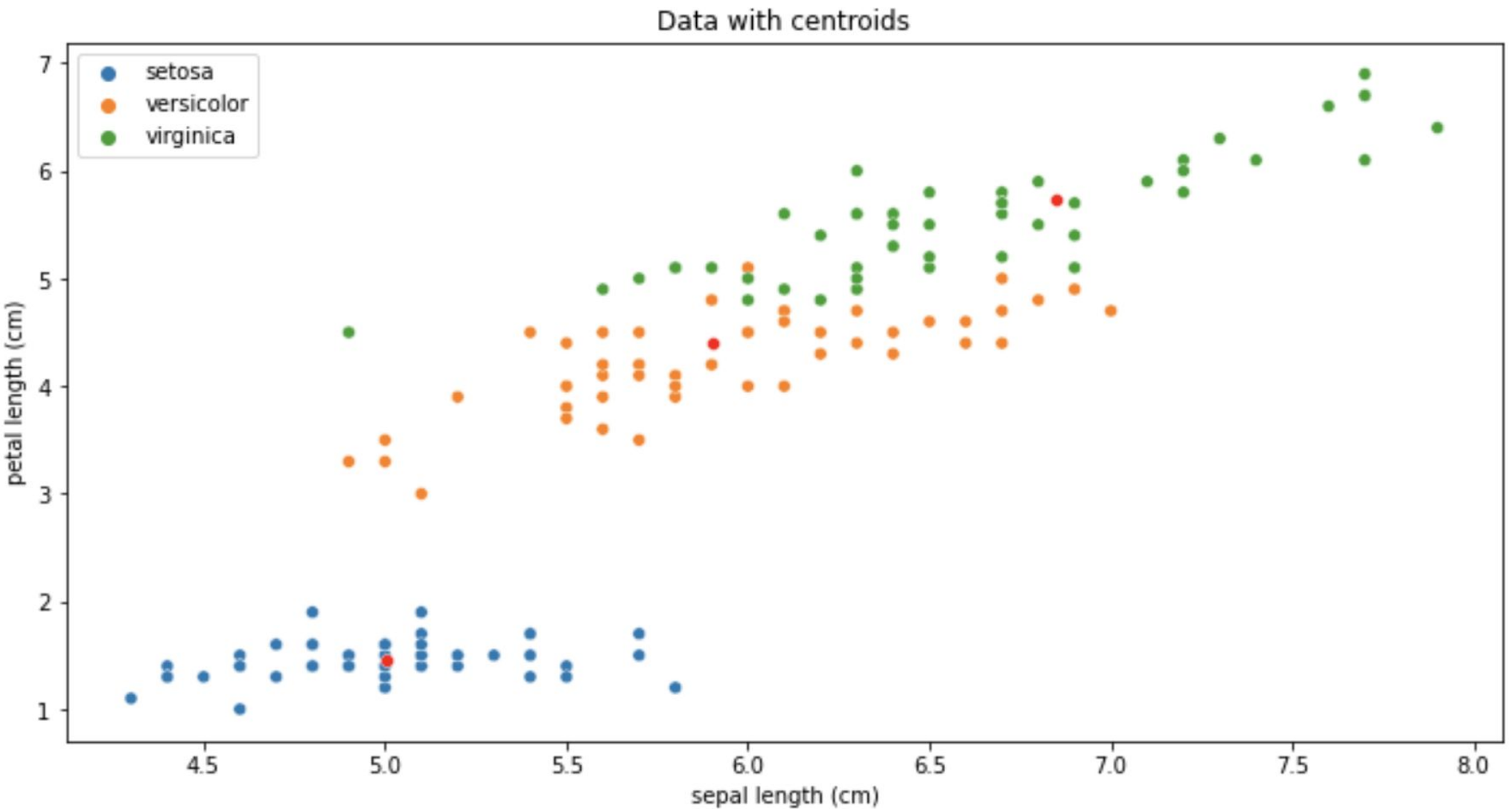
02

Densidade

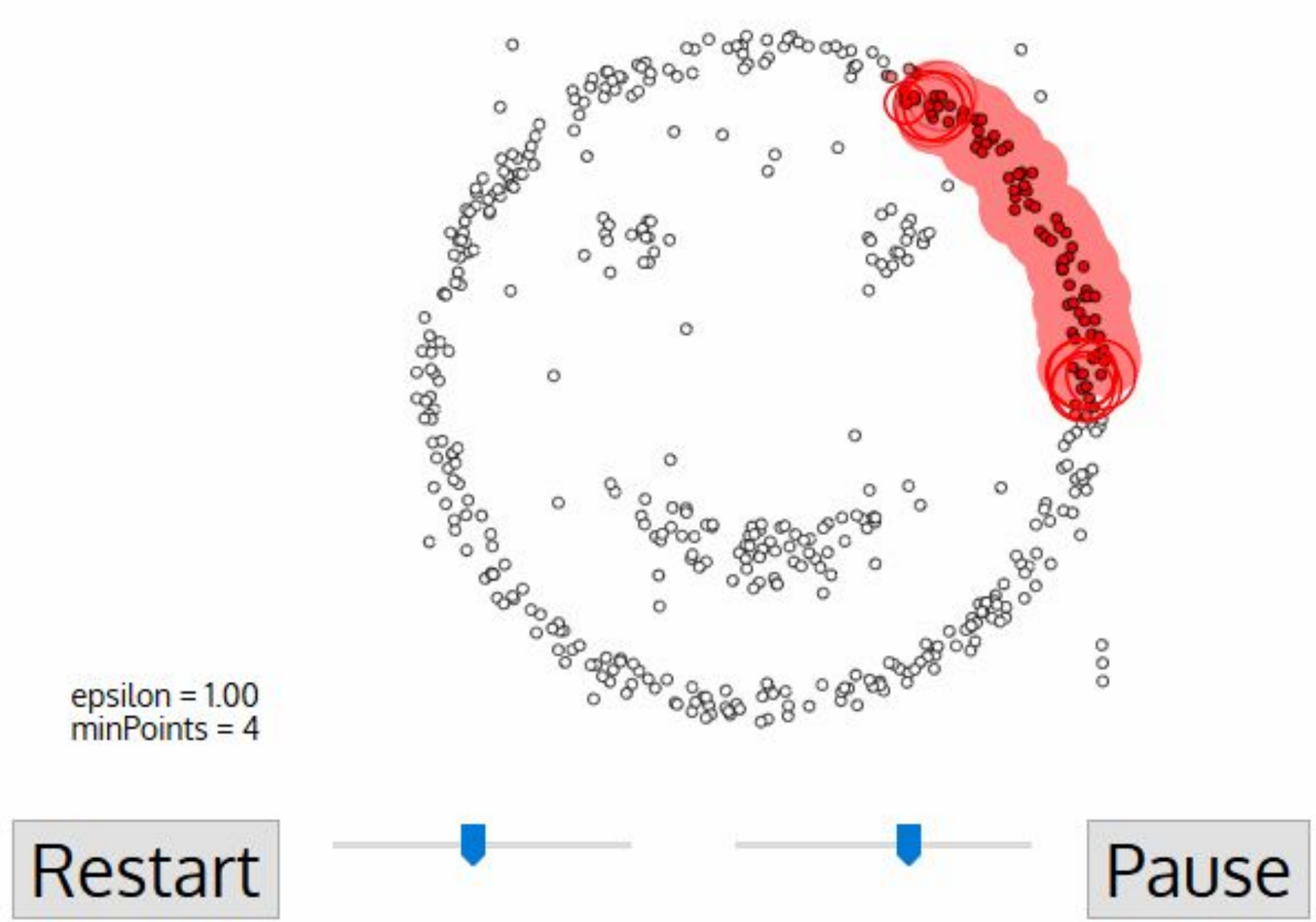
03

Hierarquia

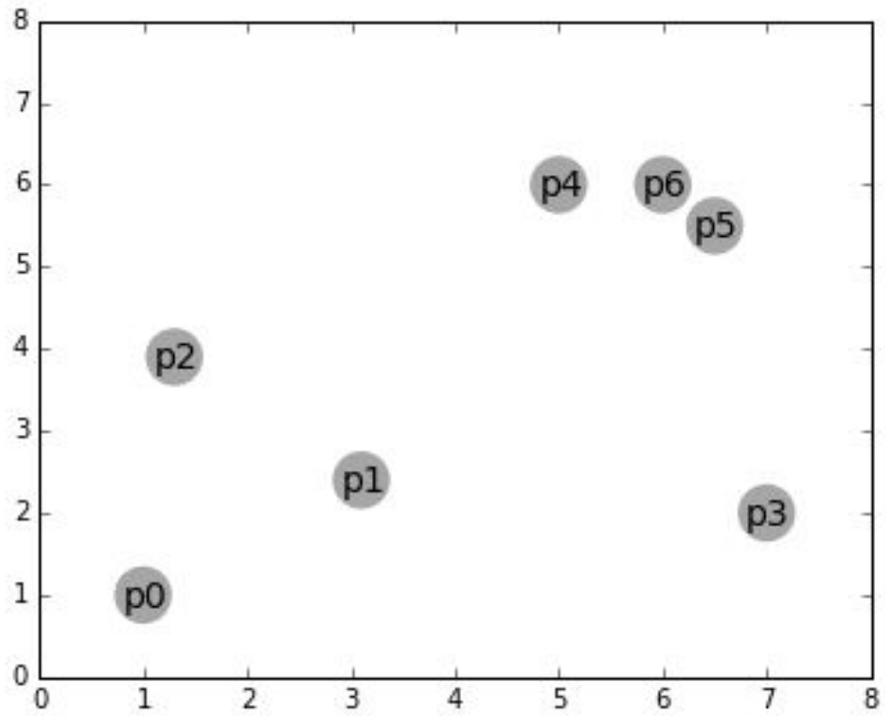
KMeans

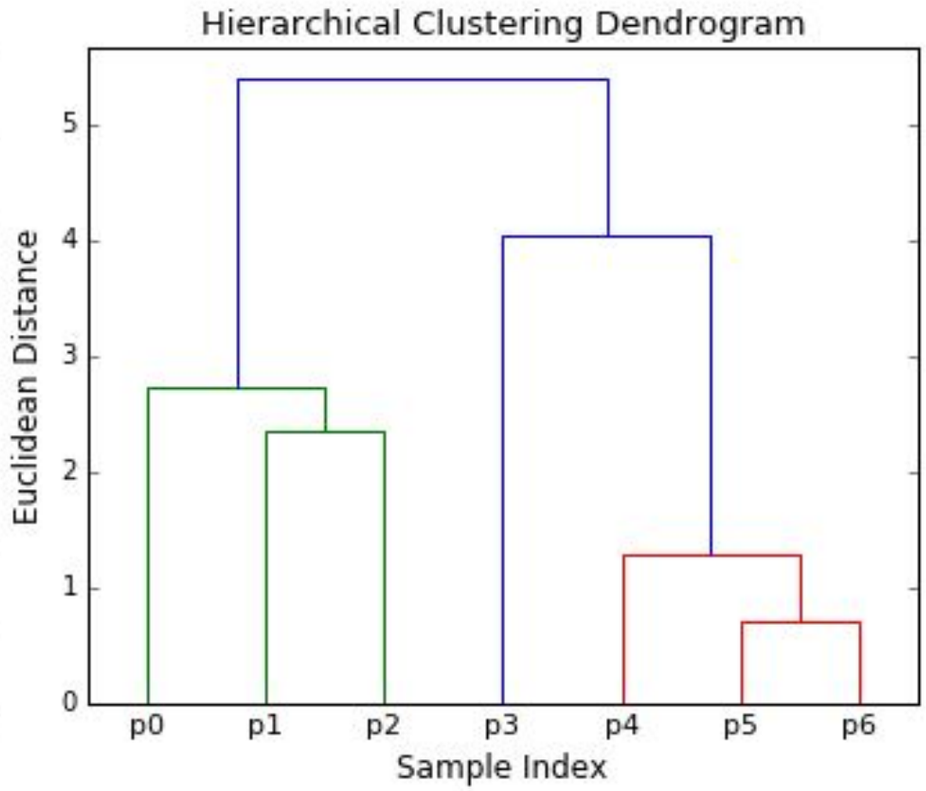

Data with centroids
setosa
versicolor
virginica
petal length (cm)
sepal length (cm)
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

DBScan



Hierarchical Clustering





Próximos passos

- Novos tipos de dados Imagem, Texto
- Distribution Clustering
- Mean Shift, Gaussian Mixture