Aprendizado Não-Supervisionado

O seguinte trabalho a seguir tem por objetivo explorar técnicas de aprendizado não-supervisionado, útil quando não temos features rotuladas em nosso dataset e queremos encontrar padrões e agrupamentos no nosso conjunto de dados.

Importação dos dados

Nessa etapa foi feito o carregamento dos dados que serão utilizados para a tarefa.

Temos, no total:

total_amostras	total_features
150	4

1. Redução de Dimensionalidade

A redução de dimensionalidade é útil para reduzirmos o número de features, que representam dimensões, do nosso conjunto de dados, seja para visualização dos dados, análise, clusterização, entre outras finalidades.

Vamos utilizar o PCA, como pedido no exercício, e também fiz um teste com o UMAP (**Uniform Manifold Approximation and Projection for Dimension Reduction),** que é muito útil para reduzir dimensões de conjuntos de dados que não possuem relação linear entre suas variáveis.

PCA

Abaixo está ilustrado a redução de dimensão dos dados com PCA, passando para 2 variáveis para plotar em 2 dimensões. Já é possível visualizar certa possível separação entre o dataset, indicando um possível padrão para a clusterização.

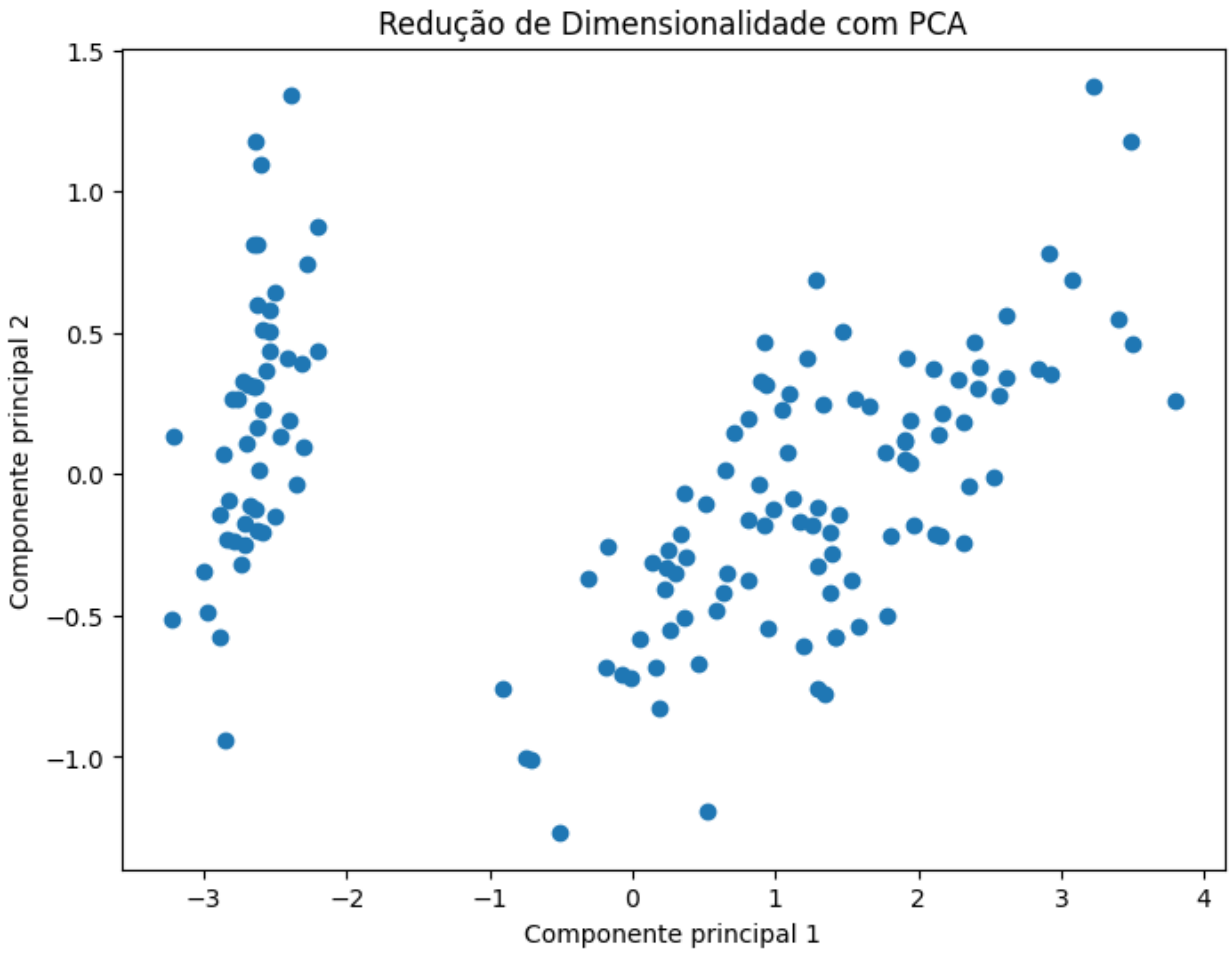

Figura 1.

UMAP

Abaixo, será ilustrado a redução de dimensionalidade com UMAP, que, como mencionado, possui um diferencial para reduções de dimensionalidade não lineares, podendo capturar uma nuância maior entre as variáveis do dataset.

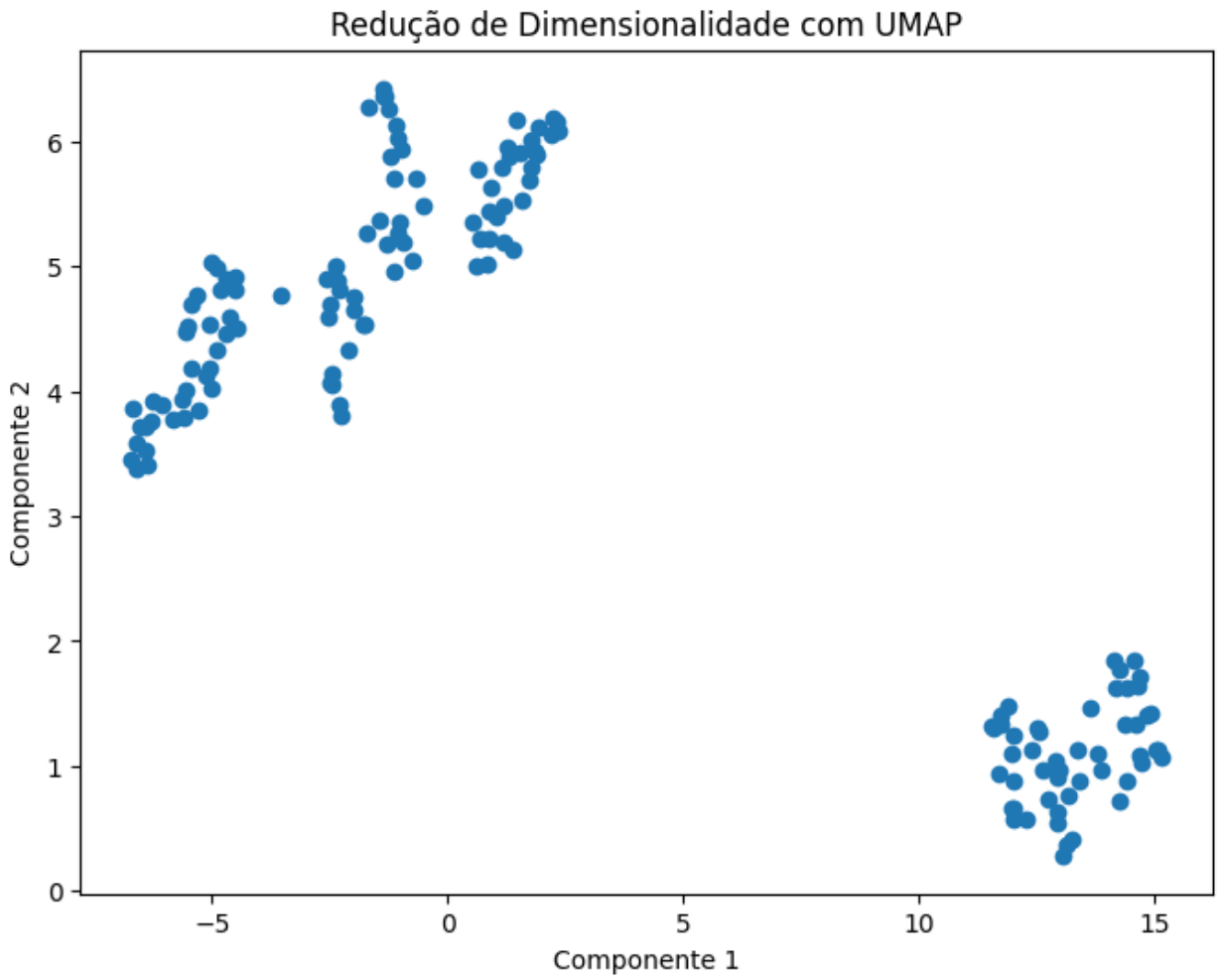

Figura 2.

2. Clusterização

K-Means

Para a clusterização do dataset, foi utilizado o K-Means, que é um método de agrupamento baseado em centroides, em que particiona os dados em K subgrupos distintos, onde cada ponto de dados pertence ao subgrupo (cluster) com a média mais próxima. Com isso, inicializa aleatoriamente os centroides, atribui o subgrupo mais próximo a esse centroide, recalcula os centroides e repete esse processo até convergir.

Abaixo, temos nosso gráfico depois da clusterização, com os centroides plotados no gráfico:

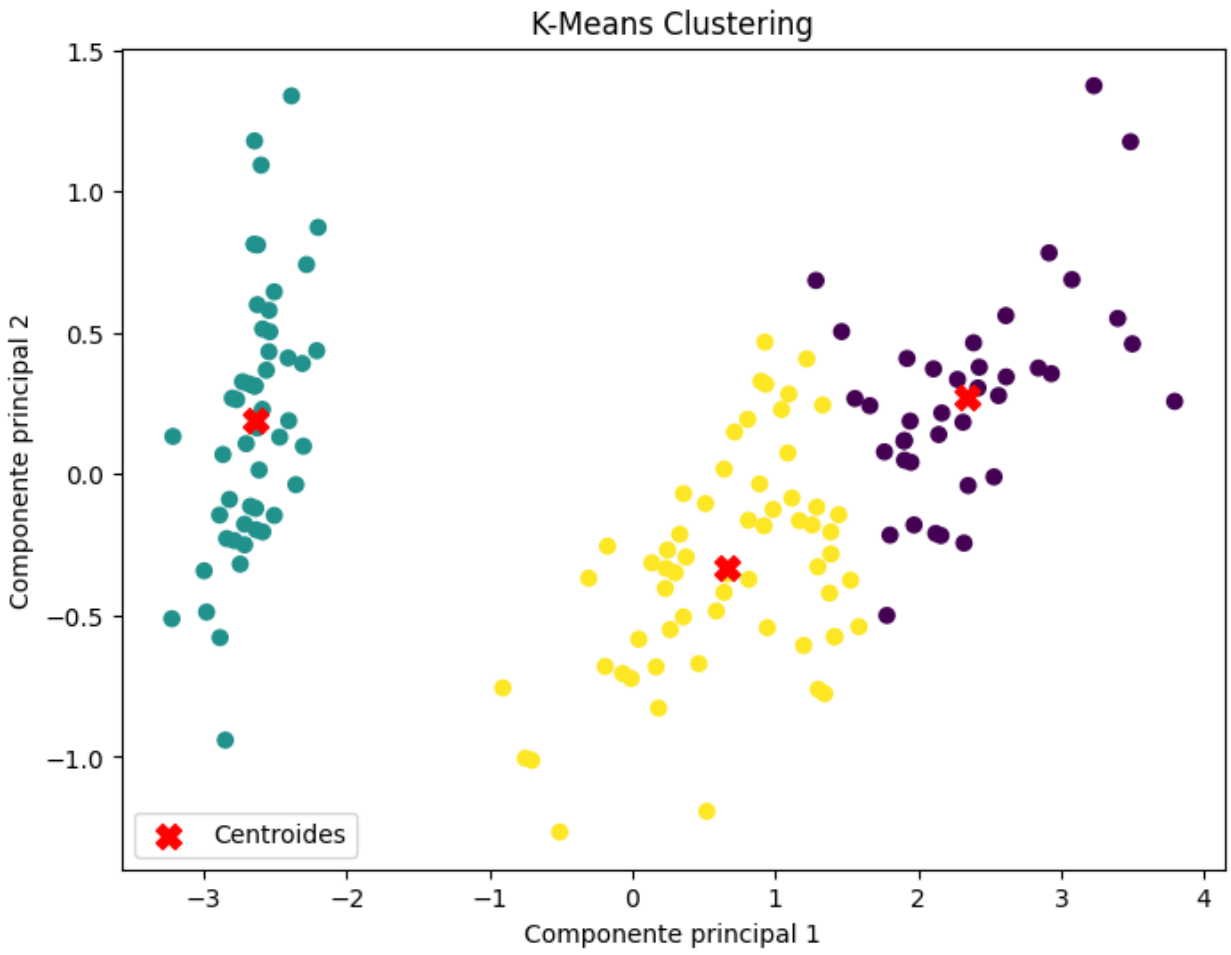

Figura 3.

Clusterização Hierárquica

A clusterização hierárquica, diferente do K-Means, forma uma estrutura de hierarquia em seu agrupamento, podendo ser feito da seguinte forma: cada ponto é, a princípio, atribuído um cluster, e a partir dai clusters são **sucessivamente mesclados** a partir de suas similaridades, formando a espécie de **hierarquia** mencionada. É muito útil quando não sabemos de antemão a quantidade ideal de clusters para nosso problema, podendo visualiza-los em uma espécie de árvore (dendograma) para estimar a quantidade ideal de clusters para nosso problema.

Abaixo, temos o Dendograma gerado a partir da clusterização hierárquica feita no nosso problema:

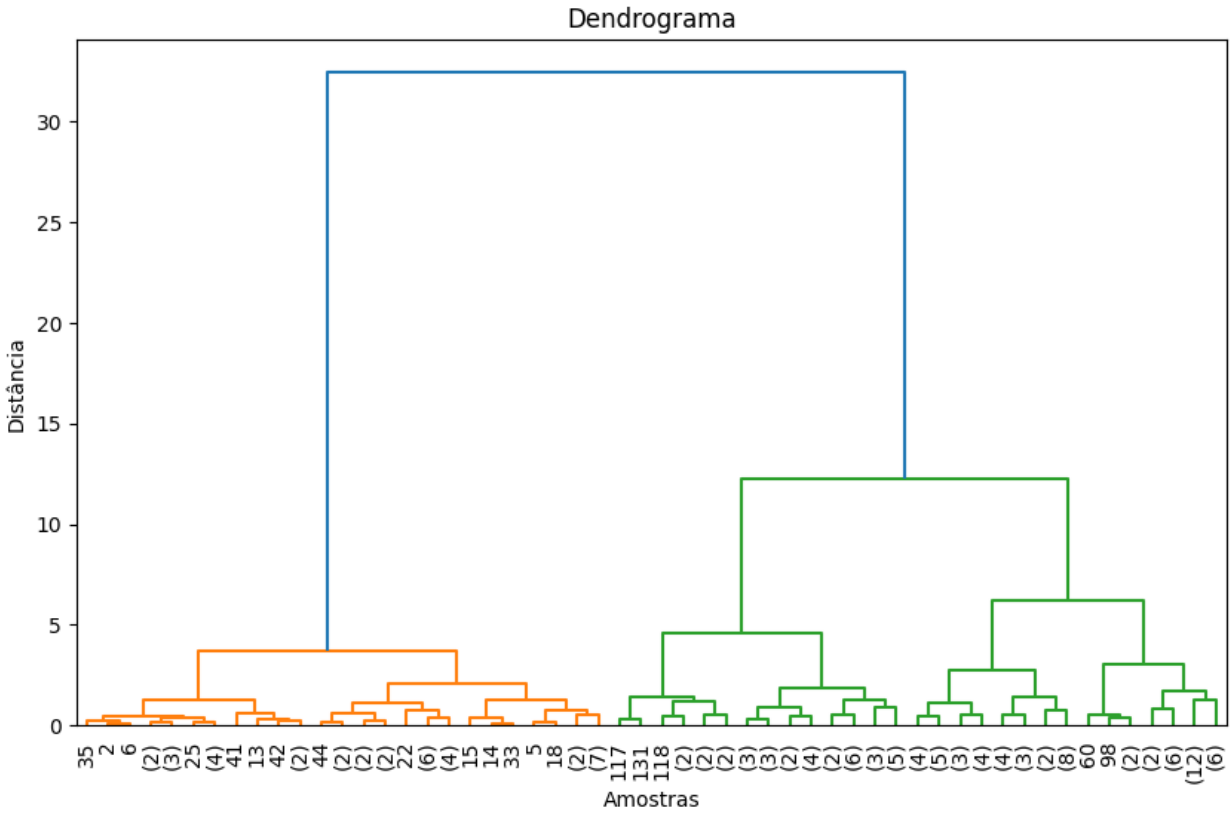

Figura 4.

Pelo dendograma acima identificamos, principalmente, 2 clusters principais (laranja e verde).

Vamos selecionar 3 clusters para compararmos com o PCA, fazendo o corte em uma distância de, aproximadamente, 10 de acordo com o dendograma.

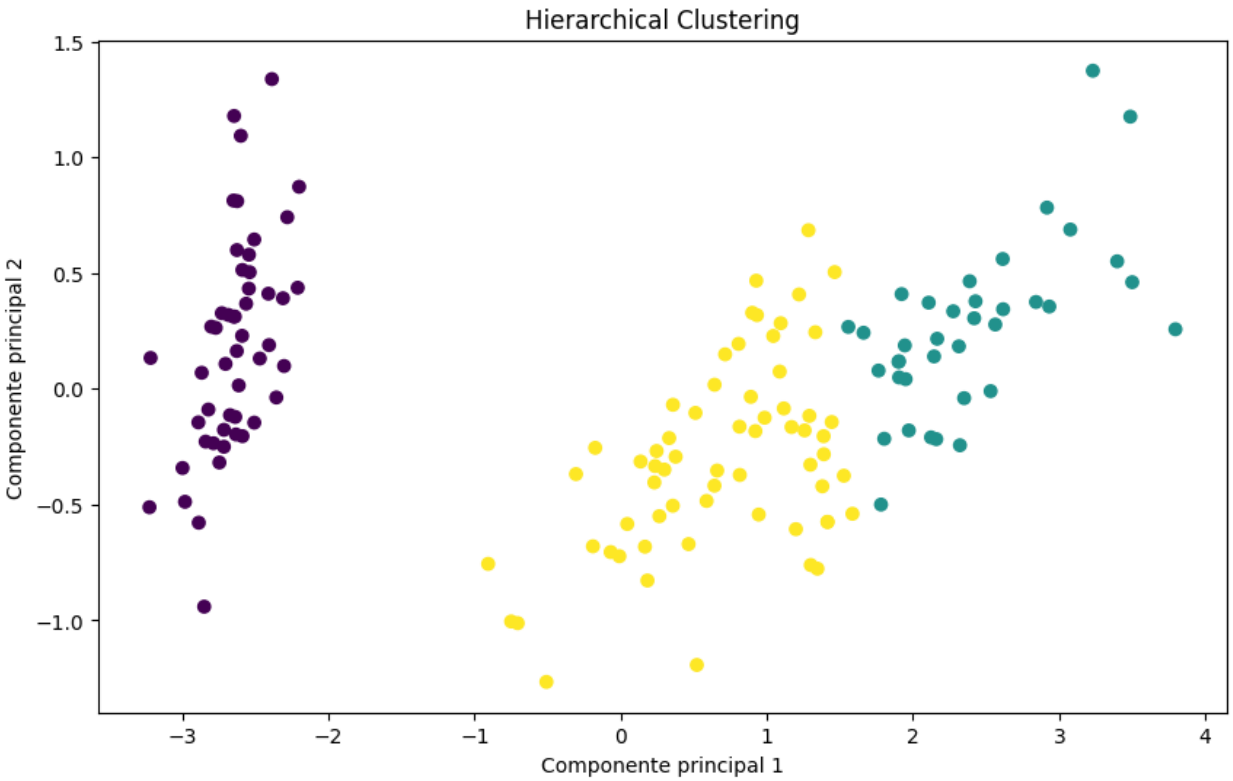

Figura 5.

3. Comparação entre métodos

Para comparação entre os métodos de clusterização, foi utilizado o Indice Randômico Ajustado (ARI), obtendo os seguintes resultados para cada exemplo:

```
Adjusted Rand Score - K-Means: 0.716
Adjusted Rand Score - Agrupamento Hierárquico: 0.744
```

Vemos que o Agrupamento Hierárquico se aproximou mais dos valores reais do que o K-Means.

Inicialmente, temos as principais características dos dois:

- O **clustering hierárquico** é um método que busca construir uma hierarquia de clusters. Essa técnica começa tratando cada ponto de dados como um único cluster e, em seguida, mescla iterativamente os pares mais próximos de

clusters até que todos os pontos sejam mesclados em um único cluster abrangente ou até que uma estrutura desejada seja alcançada. É isso que, basicamente, mostra o nosso **Dendograma,** sendo muito útil quando não conhecemos a estrutura dos dados de antemão e a quantidade de clusters.

- O **agrupamento K-means**, por outro lado, particiona os dados em **K subgrupos distintos,** onde cada ponto de dados pertence ao subgrupo (cluster) com a média mais próxima. O processo envolve inicializar aleatoriamente centroides K, atribuir pontos ao centroide mais próximo, recalcular centroides e repetir essas etapas até a convergência.

Portanto, vemos que a clusterização hierárquica captura melhor a estrutura dos nossos dados e nos permite uma análise mais abrangente, contribuindo também para que sua taxa de ARI, no trabalho, seja maior para essa técnica.