

**Insper: Instituto de Ensino e Pesquisa**

Ellen Shen

Enzo Neto

Gabriel Huerta

Giulia Castro   
   
   
   
 

**CIÊNCIA DOS DADOS**

 Projeto 3: Análise de dados do Spotify  
   
   
   
 

**São Paulo**

**2019**

**Introdução:**

Falar da pergunta

Falar de como baixou a base de dados

Falar dos 3 parâmetros escolhidos

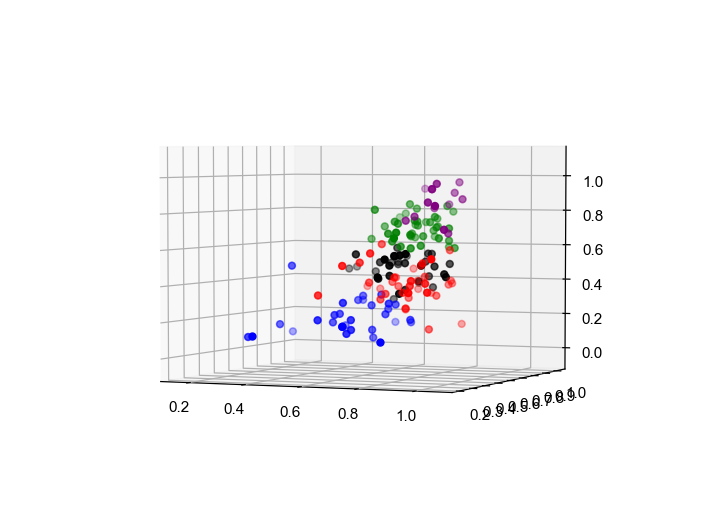
**Clusterização:**

Para a técnica de clusterização inicialmente foram importadas as seguintes bibliotecas: matplotlib.pyplot, pandas, numpy, mpl\_toolkits.mplot3d, seaborn e sklearn. Logo após foram importadas as bases de dados e divididas entre a base que seria utilizada para o treinamento – chamada de top\_2017\_2018\_treinamento – e a base que seria utilizada para o teste do código criado – chamada de top\_2017\_2018\_teste.

Depois desse processo inicial, começamos uma análise exploratória básica dos dados. Desse modo, foi plotado um scatterplot com as três variáveis escolhidas para o nosso estudo, como mencionado na introdução, para a visualização da distribuição das músicas que compunham a base de dados no espaço.

Através do comando *cluster.MiniBatchKMeans* a equipe definiu por trabalhar com 5 grupos distintos, e logo após foi realizado um *fit* passando como argumento as colunas das variáveis que foram definidas anteriormente (danceability, energy e valence). Utilizando o comando *.predict*, o grupo obteve a separação das músicas nos diferentes grupos pré-estabelecidos. O resultado foi guardado na variável *saida\_2017\_2018* e copiados para construir um novo dataset no qual possuía a coluna de saída com os seus respectivos clusters, o qual foi chamado de: *musicas\_2017\_2018*.

Após esse processo foi plotado um outro gráfico de dispersão que separava as músicas pertencentes a cada grupo por cores diferentes. Ou seja, nesse novo gráfico obtivemos a mesma distribuição das músicas no espaço, porém com pontos de cores diversas onde o Grupo 1 foi representado pela cor vermelha, o Grupo 2 pela cor verde, o Grupo 3 pelo azul, Grupo 4 pelo preto e o Grupo 5 pelo roxo. Chegando ao resultado representado na Figura 1 abaixo.



**Figura 1** – Gráfico de dispersão das músicas separadas nos grupos

Apesar do gráfico possibilitar uma ótima observação da distribuição dos Grupos, ele não possibilita uma precisão nos valores utilizados para a construção da divisão. Tendo isso em vista, foi realizado um Pairplot para cada grupo para entendermos a lógica utilizada pelo computador para fazer a divisão. Sendo assim foi concluído que:

**- Grupo 1:** Apresenta dançabilidade alta, porém a valência e a energia médias. Portanto, são músicas dançantes, mas sem muito barulho e não são nem muito alegres nem tristes.

**- Grupo 2:** Possui as três variáveis com valores altos. Conclui-se que são músicas dançantes, barulhentas e alegres.

**- Grupo 3:** Pode-se observar que esse grupo apresenta as músicas com valência muito baixa. Assim, conclui-se que as músicas desse grupo são negativas e mais deprimentes.

**- Grupo 4:** Como pode ser visto as músicas pertencentes a esse grupo apresentam dançabilidade e energia altas e valores de valência medianos. Conclui-se, então, que as músicas são dançáveis, bem rápidas e barulhentas, porém não são tão positivas.

**- Grupo 5:** Nota-se pelo gráfico de dispersão que essas músicas apresentam valência, energia e dançabilidade altos. Ou seja, as músicas desse grupo são no geral bem dançantes, bem energéticas e muito positivas.

Com a conclusão dos resultados obtidos, começamos a trabalhar com a base