DATA SCIENCE

AULA 6 - Aprendizagem Não- supervisionada II

Prof. Gabriel Resende Machado

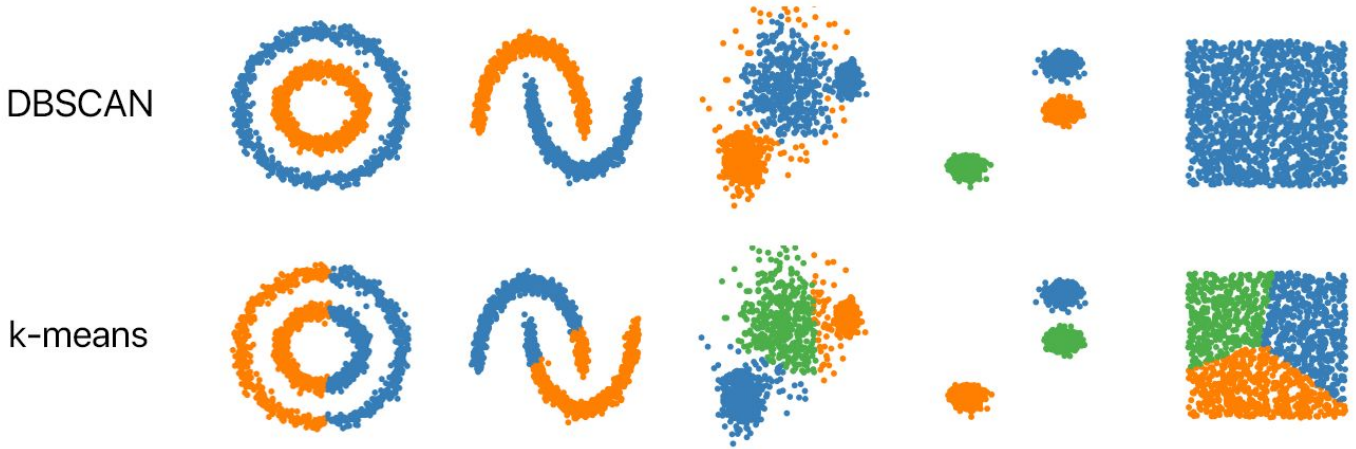
gabrielmachado@unifeso.edu.com

https://www.linkedin.com/in/machadogabriel

https://github.com/UNIFESO-Gabriel/data-science

Relembrando:
Pontos Negativos do k-Means

- *k-Means* é bastante suscetível a problemas quando os *clusters* são de diferentes tamanhos.
- k-*Means* é também bastante suscetível a problemas quando os *clusters* têm formatos globulares ou diferentes densidades.



O algoritmo DBSCAN

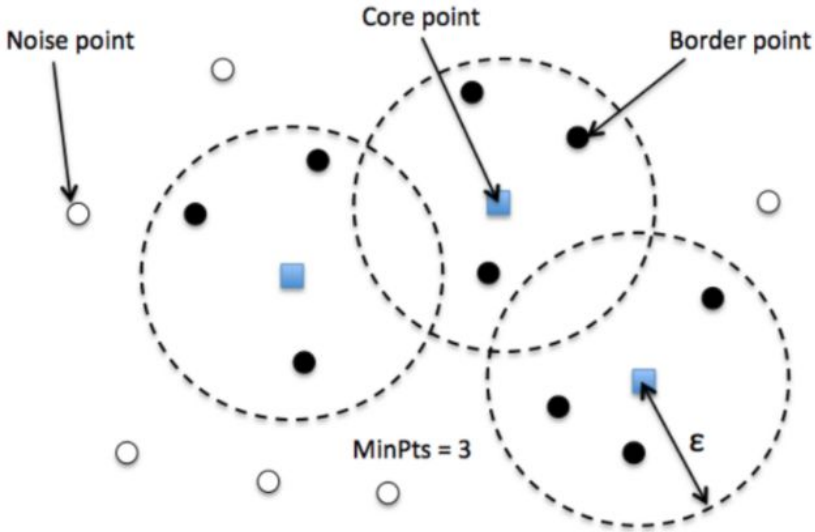
- DBSCAN é um acrônimo para ***D**ensity-**B**ased **S**patial **C**lustering **A**pplications with **N**oise;*
- Trabalha com a definição de **densidade** para realizar os agrupamentos, identificando diferentes tamanhos e formatos **mesmo com a presença de ruídos e *outliers***;
- Não necessita da definição do número ***k*** de *clusters*, como no *k-Means*;
- Ao invés, necessita que sejam definidos dois parâmetros:
  - ***eps:*** uma medida de distância utilizada para localizar pontos circunvizinhos;
  - ***minPts****:* o número mínimo de pontos aglomerados em uma região para considerá-la densa

O algoritmo DBSCAN

- As definições dos parâmetros do DBSCAN podem ser melhor entendidas a partir de dois conceitos:
  - **Alcançabilidade:** um ponto é alcançável por outro ponto se ele está dentro de uma distância *'eps'* dele;
  - **Conectividade:** determina que determinados pontos formam um *cluster* em particular caso haja uma conexão em cadeia. Por exemplo, *p* e *q* são pontos que podem ser conectados caso *p -> r -> s -> t -> q*, onde *'a -> b'* significa 'b é vizinho de a'.

O algoritmo DBSCAN

- Há três categorias de pontos após o processo de clusterização do DBSCAN:
  - **Núcleo (*Core*):** um ponto com pelo menos *'m'* pontos dentro de uma distância *'eps'* dele mesmo;
  - **Fronteira (*Border*):** um ponto com pelo menos um ponto de núcleo até uma distância *eps;*
  - **Ruído (*Noise*):** um ponto que nem é núcleo nem fronteira.

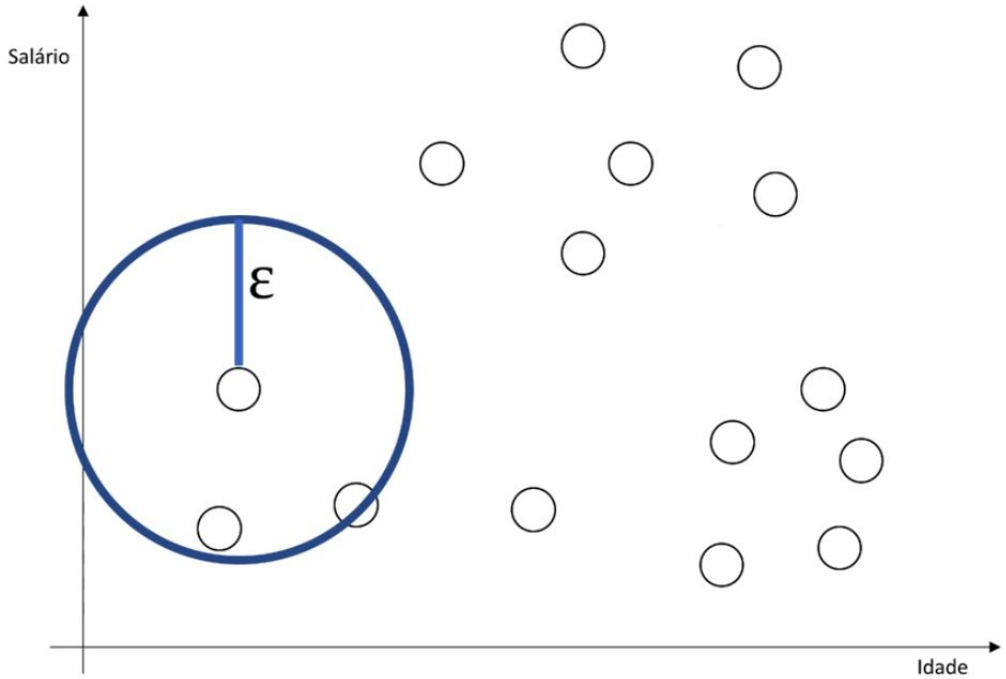


O algoritmo DBSCAN
Etapas

1. O algoritmo escolhe aleatoriamente um ponto não selecionado no conjunto de dados;
2. Se houver pelo menos ‘*minPts*’ dentro de um raio ‘*eps*’ do ponto escolhido, considere todos os pontos como pertencentes ao mesmo *cluster;*
3. Após os *clusters* são expandidos aplicando recursivamente os cálculos anteriores para cada ponto circunvizinho.

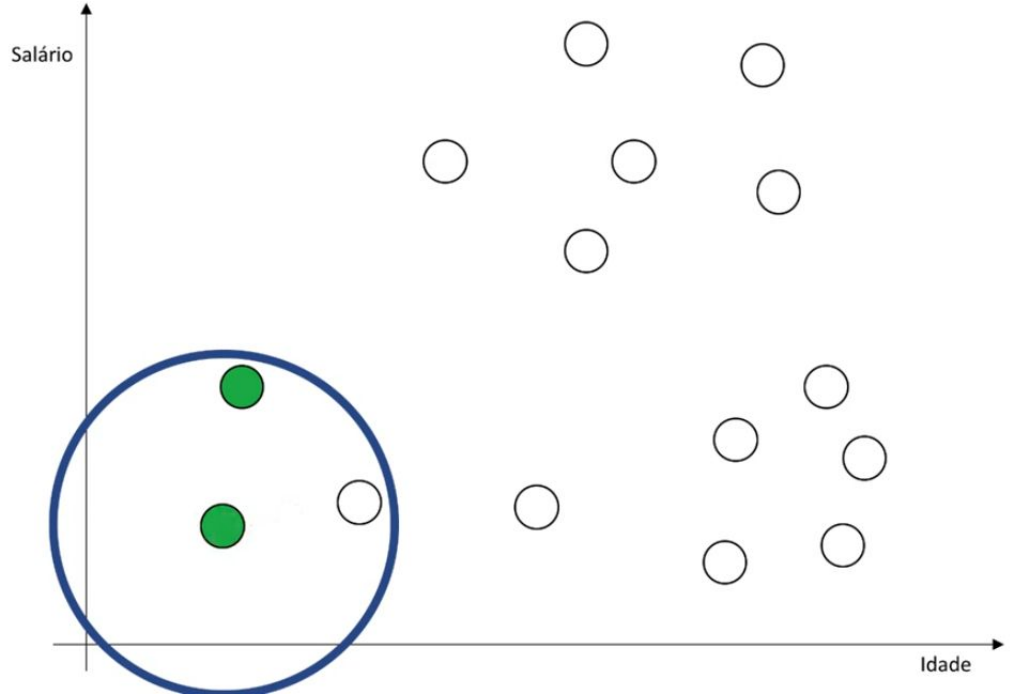
O algoritmo DBSCAN

Exemplo 1: minPts = 1



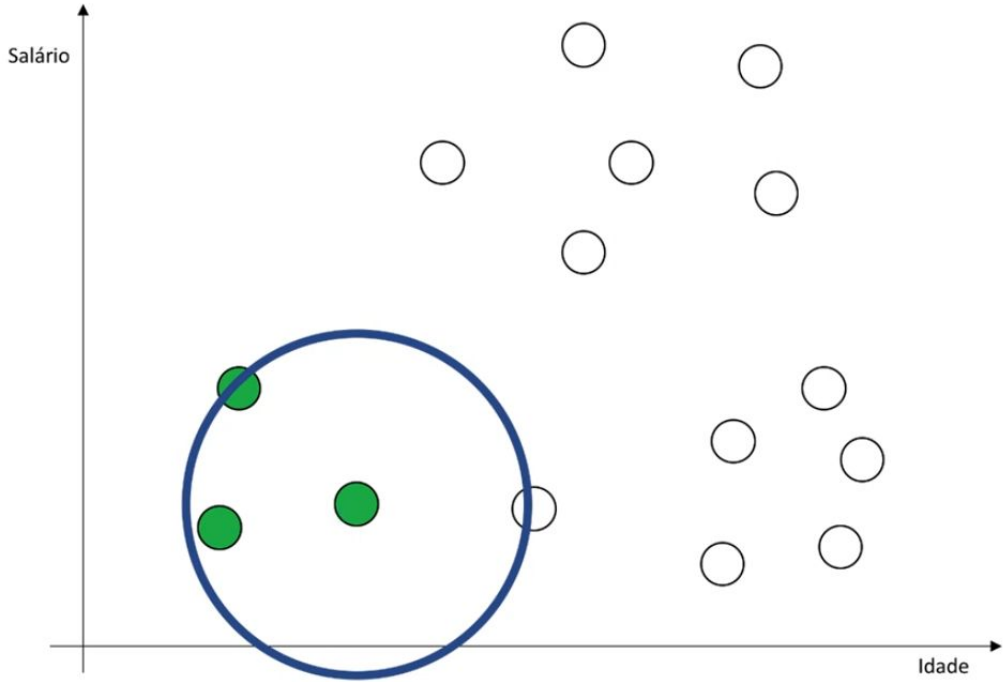
O algoritmo DBSCAN

Exemplo 1: minPts = 1



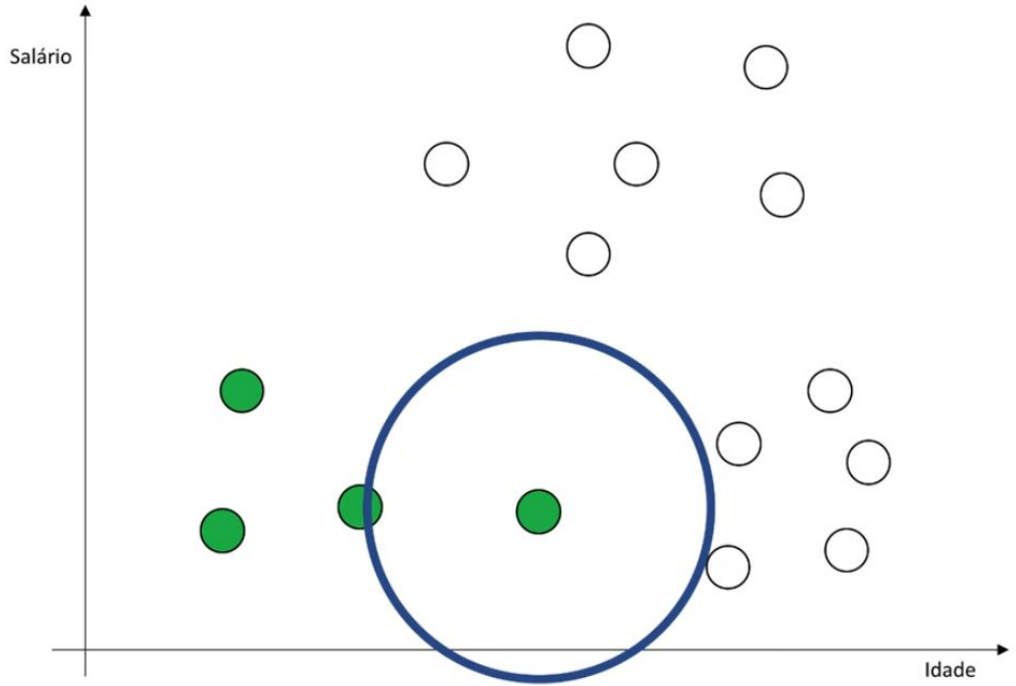
O algoritmo DBSCAN

Exemplo 1: minPts = 1



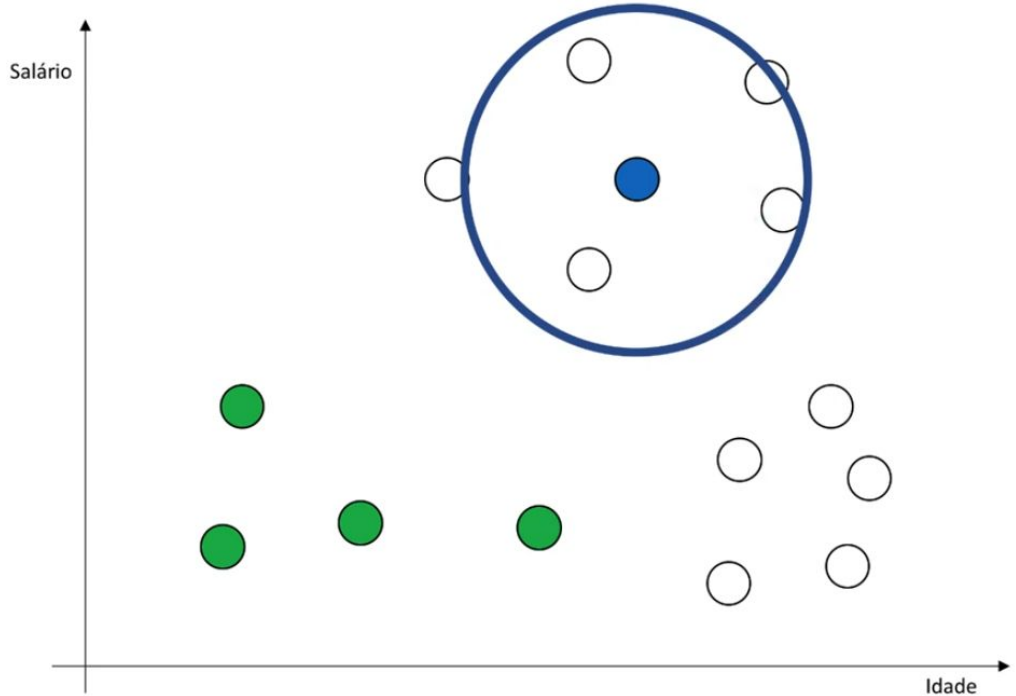
O algoritmo DBSCAN

Exemplo 1: minPts = 1



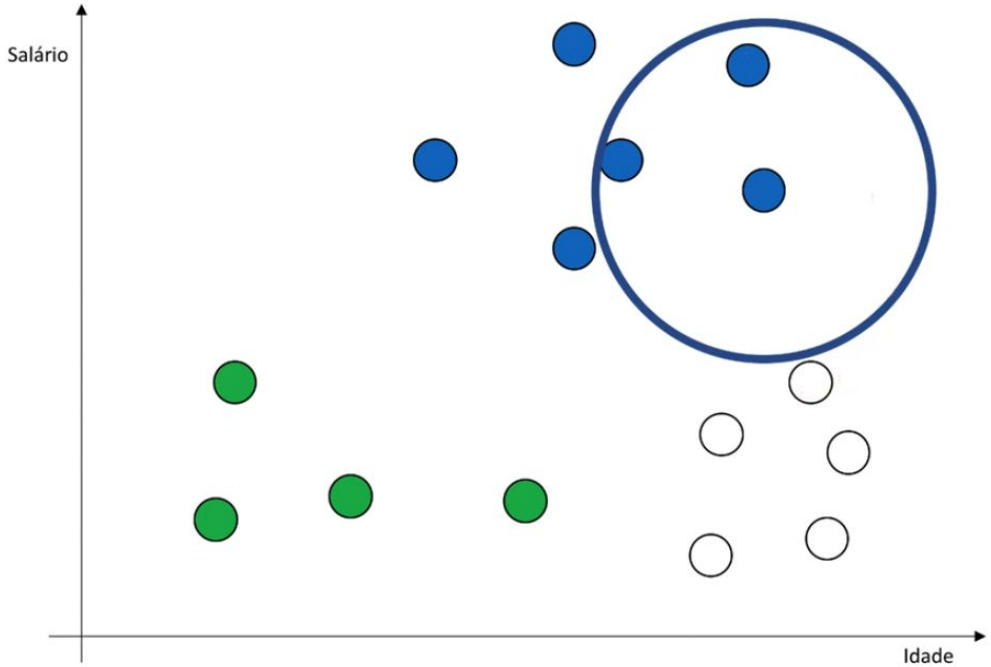
O algoritmo DBSCAN

Exemplo 1: minPts = 1



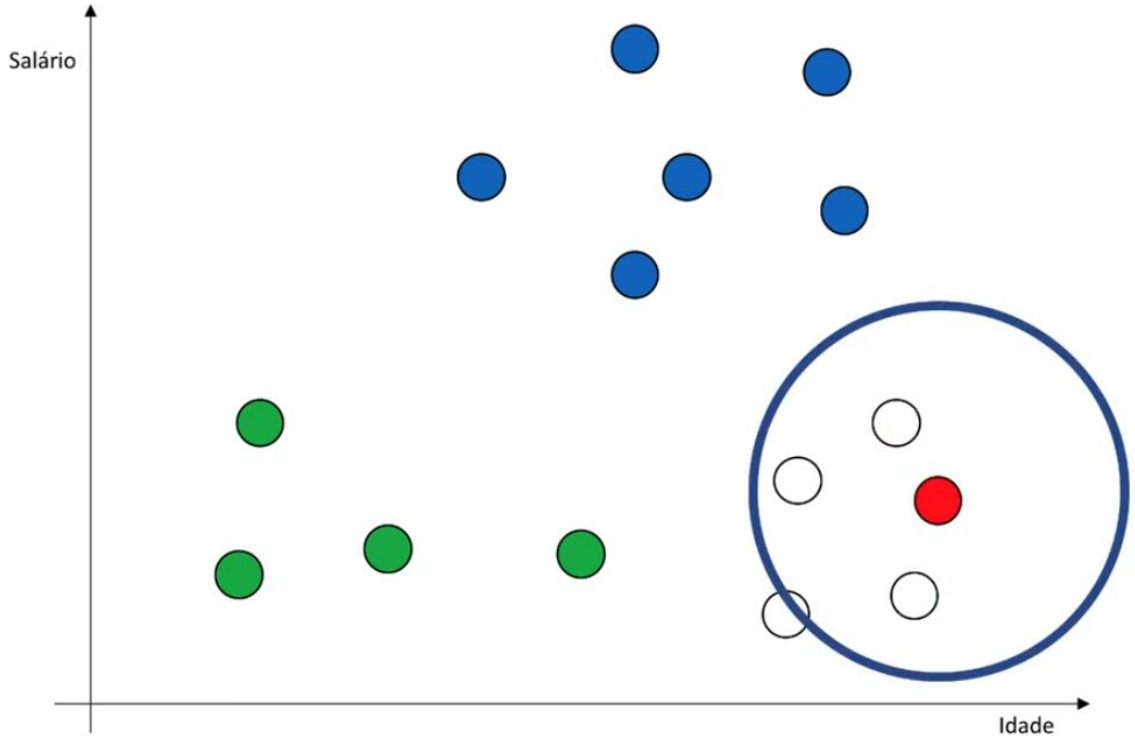
O algoritmo DBSCAN

Exemplo 1: minPts = 1



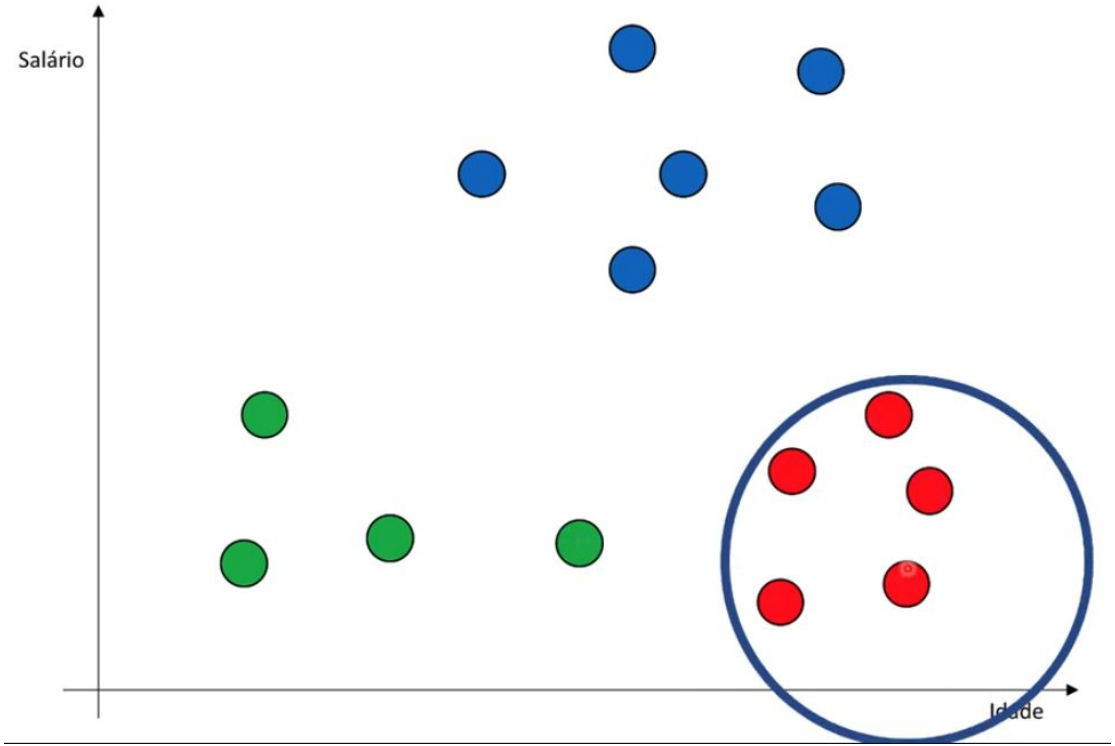
O algoritmo DBSCAN

Exemplo 1: minPts = 1



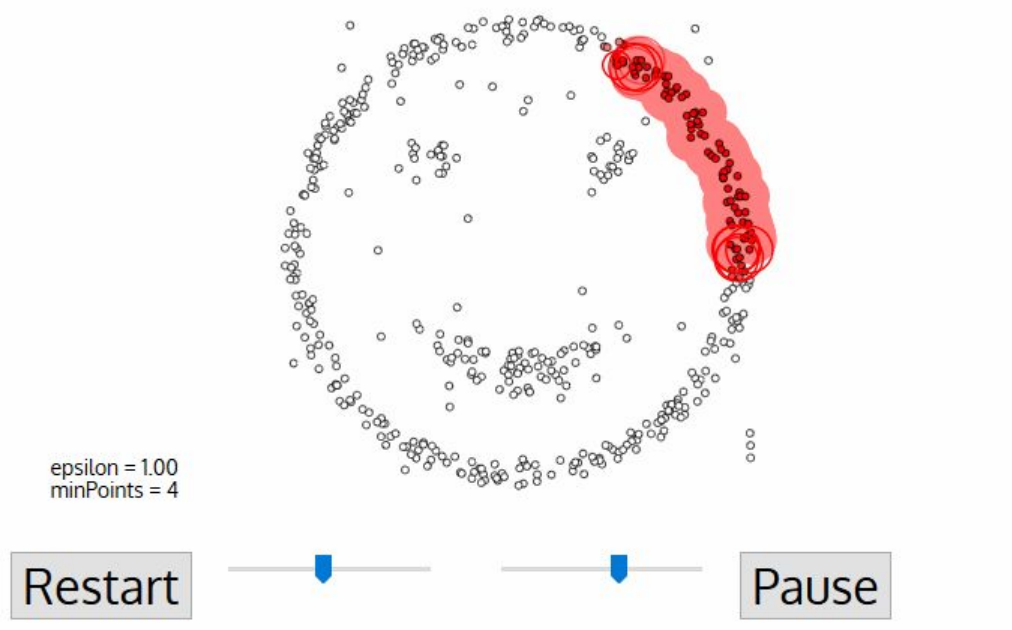
O algoritmo DBSCAN

Exemplo 1: minPts = 1



O algoritmo DBSCAN

Exemplo 2



O algoritmo DBSCAN
Pontos Fracos

- **Sensibilidade a Parâmetros:** o desempenho do DBSCAN depende da definição correta de parâmetros, como o *eps* e o número mínimo de pontos *minPts*. Escolher valores apropriados pode ser desafiador e pode exigir conhecimento do domínio.
    - Heurísticas comuns são $minPts \geq D + 1$, onde D representa o número de dimensões. O parâmetro *eps* pode ser definido a partir da curva do cotovelo de um gráfico de *k* pontos vizinhos.
- **Escalabilidade:** o DBSCAN pode se tornar computacionalmente custoso para conjuntos de dados maiores, pois precisa calcular distâncias entre pares de pontos, resultando em uma complexidade temporal de O(n^2);
- **Ambiguidade de Pontos de Borda:** A atribuição de pontos de borda pode ser às vezes ambígua, levando a diferenças nos resultados do clustering dependendo da ordem de processamento. Isso pode tornar o clustering menos determinístico.

Exercício

Utilize o algoritmo DBSCAN para agrupar os registros de vendas de jogos de videogame ao redor do mundo. Utilize o link https://shorturl.at/crFNW para acessar o *notebook* e https://shorturl.at/bpFI4 para acessar os dados.

DATA SCIENCE

AULA 6 - Aprendizagem Não-Supervisionada II

Dúvidas e/ou perguntas?

gabrielmachado@unifeso.edu.com

https://www.linkedin.com/in/machadogabriel

https://github.com/UNIFESO-Gabriel/data-science