06

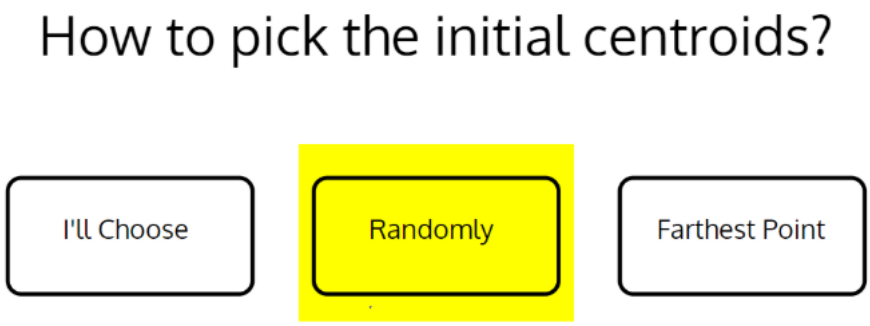
Para saber mais: K-Means

O algoritmo `kmeans` foi um dos primeiros a ser desenvolvido para executar a tarefa de cluster. Rápido e eficiente, o funcionamento dele consiste basicamente em:

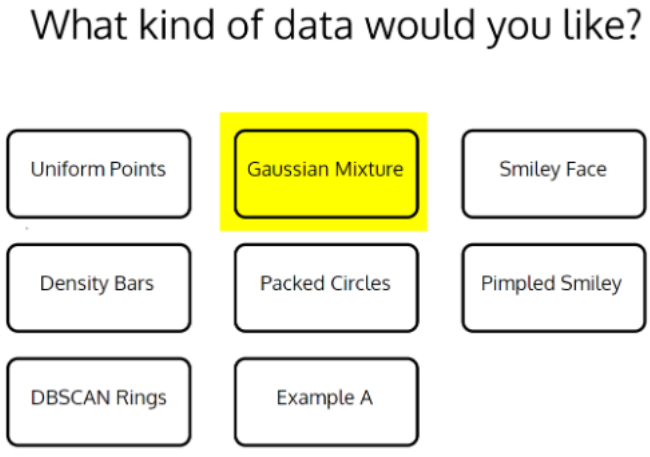
- Escolha do usuário da quantidade de cluster, geralmente representado pela letra K.
- Então, são escolhidos os centróides de cada cluster, de forma aleatória.
- A partir dos centróides, são calculadas as distâncias dos pontos pelos centros, o que define em qual cluster cada registro vai pertencer.
- Os centróides são atualizados a cada iteração.
- Finalização quando os centróides não alteram mais.

Para complementar e entender ainda melhor os principais conceitos do algoritmo `kmeans`, você pode praticar a aplicação na página de Visualização de K-means Clustering, da Naftali Harris (https://www.naftaliharris.com/blog/visualizing-k-means-clustering/).

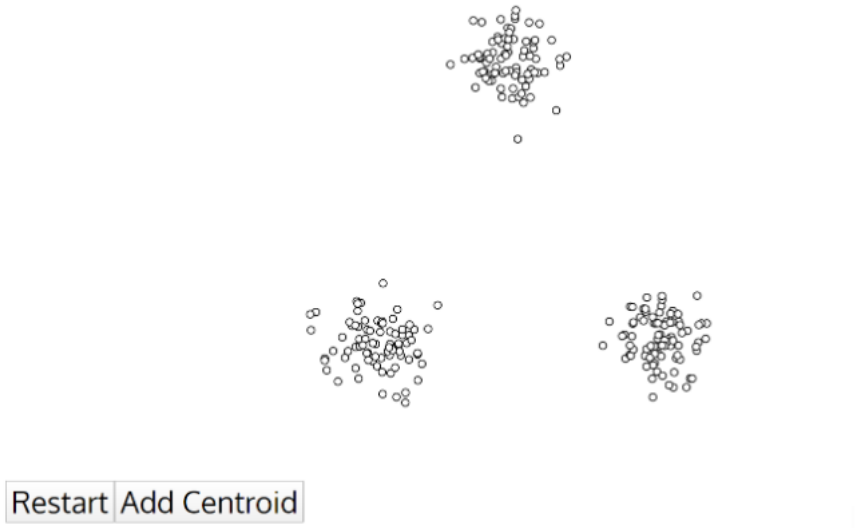
Começando com a seleção da opção “Randomly” (em português, aleatoriamente), nas alternativas de “How to pick the initial centroids?” (“Como selecionar os centróides iniciais?”, em português).



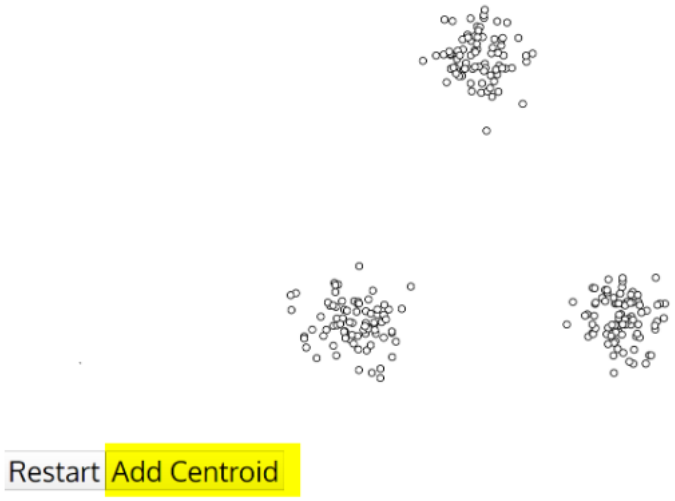
Em seguida, aparecerão outras opções. Vamos selecionar “Gaussian Mixture” (mistura de dados normais, em português).



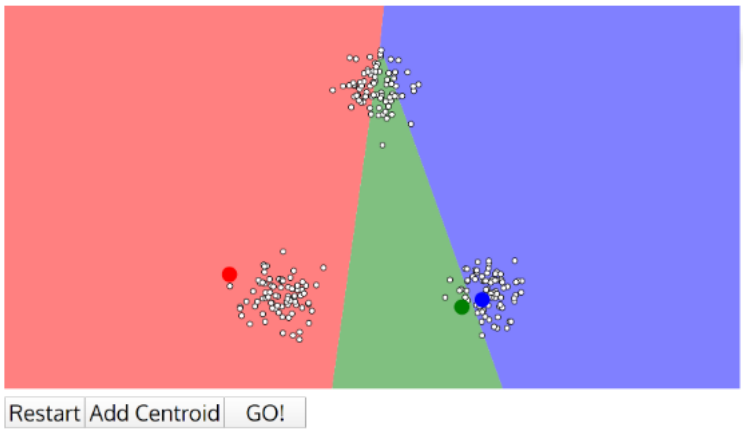
Ao clicar nessa opção, aparecerá um gráfico com vários dados.



Se analisarmos este gráfico, notamos que os dados estão meio divididos em agrupamentos, neste caso 3. Para executar o Kmeans é preciso definir a quantidade de cluster. Vamos adicionar 3, clicando em "Add Centroid".

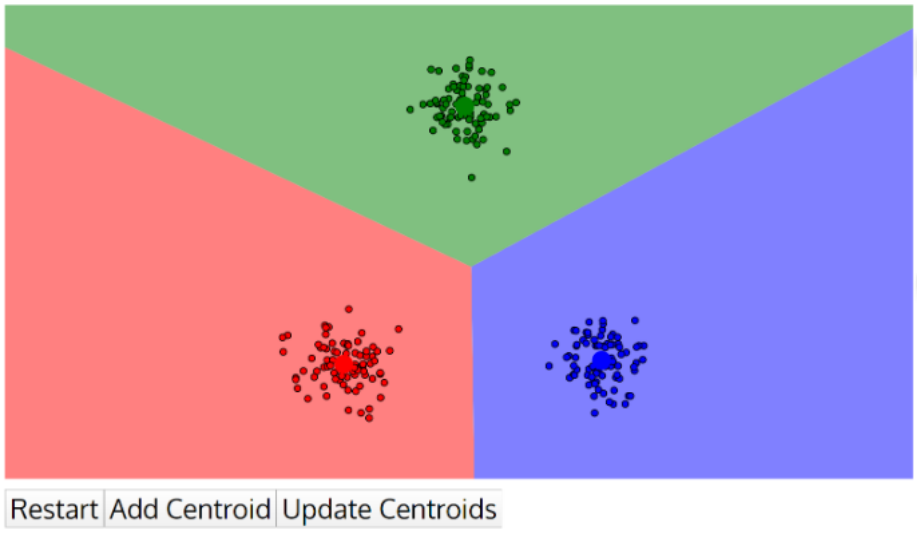


Feito isso, teremos um resultado no estilo da imagem a seguir, lembrando que esses centros(centróides) são escolhidos aleatoriamente para a clusterização, ou seja, o seu resultado provavelmente será diferente do apresentado aqui.



Agora que já definimos a quantidade de clusters, vamos executar o algoritmo clicando em "GO!".

Clicando em "Reassing Points/Update centroids" até os centroides não mudarem mais, teremos os nossos 3 clusters. Este será o resultado final



É isso que o algoritmo `kmeans` faz ao utilizarmos alguma linguagem de programação ou ferramenta de Machine Learning, por exemplo, o R, mas de forma automática.