Trabalho de Clusterização – K-means e preparação para aplicação da técnica KNN.

O objetivo deste trabalho é implementar um processo de clusterização utilizando a técnica K-means, explorando cada uma das etapas envolvidas e preparando a base de dados para uma futura implementação do algoritmo KNN (K-Nearest Neighbors). O trabalho será realizado em seis etapas, detalhadas a seguir:

1. **Estrutura de dados inicial e manipulação**
   Desenvolver uma estrutura de dados capaz de armazenar registros (elementos), com as seguintes funcionalidades:
   - Inserção de novos registros.
   - Remoção e alteração de registros.
   - Indicação, para cada elemento, se ele é ou não o centróide de um cluster. Inicialmente, cada estrutura de dados deve conter apenas um elemento, formando assim dois clusters distintos (cada um com um único elemento). A estrutura deve permitir a inclusão de novos registros, que representarão a formação e o crescimento de cada cluster.
2. **Atribuição de elementos e cálculo de distâncias**
   Sempre que um novo elemento for incluído, deve ser realizada:
   - A verificação de qual cluster está mais próximo desse novo elemento, por meio do cálculo da distância euclidiana.
   - A atribuição do novo elemento ao cluster mais próximo.
   - A atualização da informação que identifica o centróide de cada cluster, mesmo que seja um centróide virtual (calculado em tempo real).
3. **Recalculo do centróide e reorganização**
   Após a atribuição do novo elemento a um cluster, deve ser:
   - Feito o recálculo do centróide desse cluster.
   - Atualizada a marcação do centróide na estrutura de dados.
   - Realizada a reorganização dos elementos de cada cluster, considerando o novo equilíbrio e a dispersão dos dados.
4. **Análise de dispersão e criação de novos clusters**
   Avaliar os elementos de cada cluster para identificar se estão se tornando muito distantes de seus centróides originais. Para isso, deve-se:
   - Definir um limiar que indique o que significa “estar distante” do centróide.
   - Localizar os “k” elementos mais distantes de seus centróides e mais próximos do outro cluster.
   - Caso sejam encontrados elementos que se encaixem nesses critérios, eles devem ser removidos de seus clusters originais e formar um novo cluster.
   - Reorganizar todos os clusters (inclusive o novo, se criado), garantindo a consistência dos centróides.
5. **Adequação de dados para uso futuro com KNN**
   Preparar a base de dados para uma futura implementação do algoritmo KNN. Para isso, deve-se:
   - Alterar ou estender a estrutura de dados existente para suportar elementos categóricos (strings).
   - Implementar um mecanismo para converter esses elementos categóricos em valores quantitativos, durante a execução do código, sem alteração direta dos dados primários.

Essa transformação permitirá que a base de dados seja utilizada em algoritmos que exigem dados numéricos.