1 Cel ćwiczenia

Celem ćwiczenia jest przeprowadzenie analizy segmentacyjnej klientów na podstawie transakcji sklepu internetowego. Zadanie polega na utworzeniu segmentów klientów w oparciu o ich zachowania zakupowe, co umożliwi wsparcie przyszłych działań marketingowych i analiz biznesowych.

2 Zestaw danych

Dołączony w pliku dataset.xlsx zbiór danych zawiera następujące kolumny:

- **InvoiceNo** – numer faktury dla każdej transakcji,
- **StockCode** – kod produktu,
- **Description** – opis produktu,
- **Quantity** – liczba zamówionych produktów,
- **InvoiceDate** – data wystawienia faktury,
- **UnitPrice** – cena jednostkowa produktu,
- **CustomerID** – identyfikator klienta,
- **Country** – kraj klienta.

3 Zadania

3.1 Przygotowanie danych

1. Zaimportuj dane do środowiska (np. Jupyter Notebook, Google Colab).
2. Sprawdź jakość danych, w tym:
   - zidentyfikuj brakujące wartości i wybierz metodę ich obsługi (np. usunięcie lub uzupełnienie braków),
   - usuń duplikaty, jeśli występują.
3. Zachowaj tylko wiersze z przypisanym CustomerID, ponieważ są one niezbędne do identyfikacji klientów.

3.2 Tworzenie cech na potrzeby segmentacji

1. **Recency (świeżość transakcji):**
   - ustal najnowszą datę transakcji w całym zbiorze danych (np. ostatnia data w InvoiceDate),
   - dla każdego CustomerID znajdź datę ostatniego zakupu (InvoiceDate),
   - oblicz liczbę dni od daty ostatniego zakupu każdego klienta do najnowszej daty transakcji.
2. **Frequency (częstotliwość zakupów):** dla każdego `CustomerID` policz ilość zamówień (unikalnych `InvoiceNo`).
3. **Wartość transakcji:** dodaj kolumnę `TotalPurchase`, która będzie iloczynem `Quantity` i `UnitPrice`.
4. **Monetary (całkowita wartość zakupów):** dla każdego unikalnego klienta (`CustomerID`) oblicz sumę kolumny `TotalPurchase`.

3.3 Normalizacja

1. Wybór cech: w analizie klasteryzacyjnej zostaną wykorzystane trzy cechy: Recency, Frequency oraz Monetary.
2. Normalizacja cech: wartości cech RFM różnią się skalą, dlatego przed użyciem algorytmu klasteryzacyjnego należy przeskalować dane używając normalizacji w celu wyrównania ich zakresu (np. Min-Max lub Z-score).

3.4 Wykonanie klasteryzacji

Przeprowadź klasteryzację z wykorzystaniem algorytmu K-means:

1. Zdefiniuj zakres liczby klastrów k, np. od 1 do 10.
2. Dla każdej wartości k:
   - wykonaj klasteryzację na zbiorze danych za pomocą algorytmu K-means.
   - oblicz sumę kwadratów odległości od środka klastra (*inertia*), która informuje o jakości klasteryzacji.
3. Stwórz wykres *elbow method*, gdzie na osi x znajduje się liczba klastrów k, a na osi y wartość *inertia*.
4. Wybierz optymalną liczbę klastrów na podstawie charakterystycznego "załamania" (łokcia) na wykresie.

4 Interpretacja i wizualizacja wyników

1. Po wyznaczeniu optymalnej liczby klastrów dokonaj analizy każdego klastra pod kątem jego charakterystyki.
2. Dla każdego klastra oblicz średnie wartości cech `Recency`, `Frequency`, `Monetary`.
3. Wykonaj wizualizacje, takie jak wykresy rozproszenia 2D (z użyciem PCA lub t-SNE do redukcji wymiarowości).
4. Zinterpretuj wyniki.
5. Opisz potencjalne działania marketingowe na podstawie wyników klasteryzacji, np. programy lojalnościowe dla klastrów wysokiej wartości.

5 Forma przekazania ćwiczenia

Do oceny proszę przekazać plik **nr_indeksu_imię_nazwisko.zip** czyli np. **123456_Jan_Kowalski.zip**. Plik ten musi zawierać sprawozdanie z ćwiczenia w formacie pdf oraz plik z kodem źródłowym (notebook).