Opis Implementacji Algorytmów Klasteryzacji

Weronika Kłujszo (223599), Michał Korzeniewski (223399),
Miłosz Malinowski (223391), Piotr Misiejuk (223302)

April 15, 2025

Plik data.py

- Główny moduł pomocniczy zawierający wspólne funkcje dla wszystkich implementacji metod analizy skupień
- Dostarcza narzędzia do:
  - Wczytywania i preprocessingu danych
  - Transformacji i skalowania (StandardScaler, MinMaxScaler, RobustScaler)
  - Usuwania wartości odstających
  - Redukcji wymiarowości i wizualizacji (PCA, t-SNE)
  - Przekształcenia zmiennych kategorycznych na numeryczne
  - Wizualizacji wyników
  - Oceny jakości klastrów

Implementacje Gaussian Mixture Models

gmm.py

- Implementacja rozwiązania GMM wykorzystująca standaryzację
- Wizualizuje wyniki klastrowania dla 3 i 4 klastrów

gmm_robust.py

- Wersja GMM z RobustScaler dla 3 i 4 klastrów
- Specjalizacja dla danych z outlierami, odporne skalowanie

gmm_minmax.py

- Wariant z MinMaxScaler dla 3 i 4 klastrów

Pliki implementujące K-Means

kmeans.py

- Podstawowa implementacja algorytmu K-Means dla 3 i 4 klastrów
- Wykorzystuje standaryzację danych

kmeans_robust.py

- Wersja K-Means z RobustScaler dla 4 klastrów
- Przetwarzanie danych odpornych na outliery

kmeans_minmax.py

- Implementacja K-Means z MinMaxScaler dla 3 klastrów

Pozostałe pliki pomocnicze

standaryzacja.py

- Narzędzia do analizy skutków standaryzacji
- Generuje:
  - Statystyki opisowe danych
  - Wykresy rozkładu przed/po standaryzacji
  - Analizę wartości odstających

wizualizacje.py

- Biblioteka funkcji wizualizacyjnych
- Zawiera implementacje:
  - Histogramów dla zmiennych ciągłych
  - Wykresów kołowych dla danych kategorycznych
  - Wykresów pudełkowych

matrixes.py

- Skrypt generujący macierze korelacji zmiennych po różnych typach skalowania
- Uwzględnia zarówno dane ciągłe, jak i kategoryczne (z One-Hot Encoding)
- Tworzy i zapisuje wizualizacje macierzy korelacji dla:

– MinMaxScaler,
– StandardScaler,
– RobustScaler,
– zmiennych kategorycznych po One-Hot Encodingu.