SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

Laboratorium Nr 8 Data 11.01.2025 Temat: Praktyczne zastosowanie analizy skupień(clustering) do zbiorów danych Wariant 9	Imię Nazwisko Hubert Mentel Informatyka II stopień, niestacjonarne, 1 semestr, gr.1a

1. Zadanie:

Praktyczne zastosowanie analizy skupień (clustering) jako wstępnej analizy danych do dalszych etapów uczenia maszynowego.

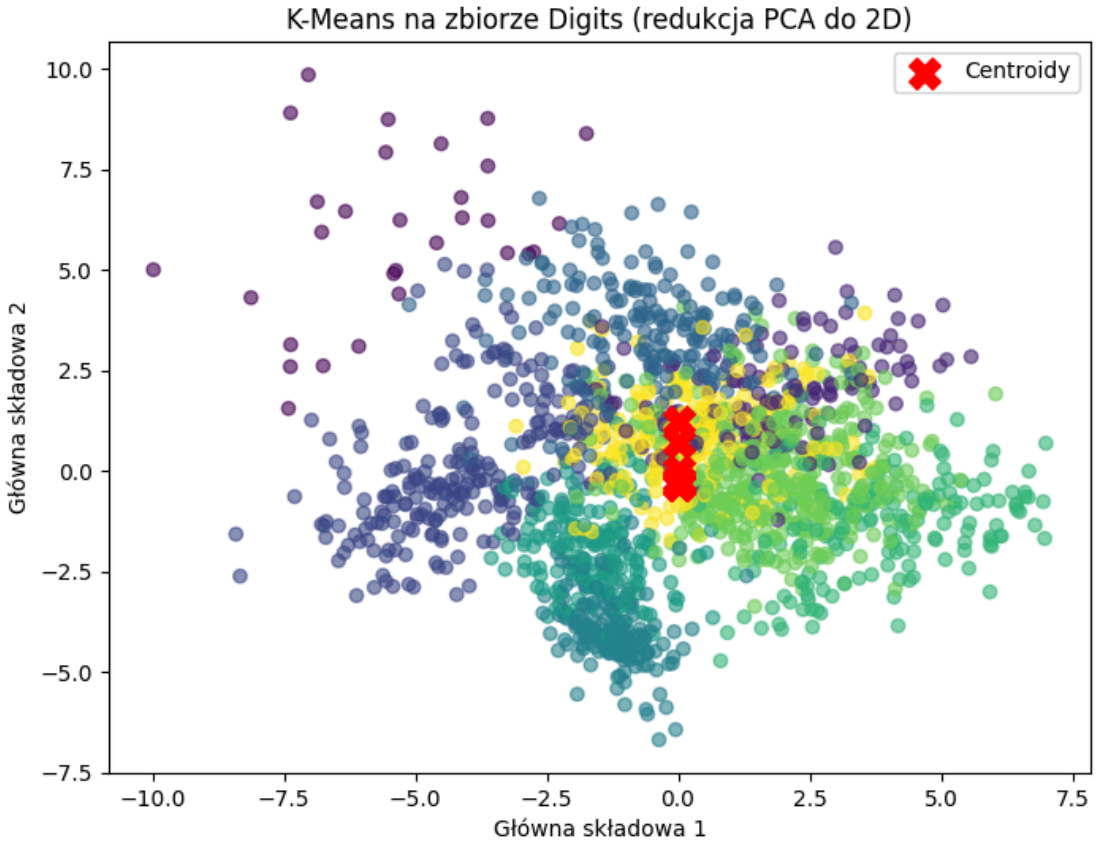
W Pythonie zastosuj k-means na zbiorze Digits dostępnych w sklearn.datasets. Zwizualizuj wyniki klasteryzacji na wykresie 2D po redukcji wymiarowości za pomocą PCA.

Pliki dostępne są na GitHubie pod linkiem:
https://github.com/HubiPX/NOD/tree/master/Zadanie%208

2. Opis programu opracowanego (kody źródłowe, zrzuty ekranu)

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import load_digits
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler

# 1. Wczytanie danych Digits
digits = load_digits()
X = digits.data  # wektory 64-pikselowe
y = digits.target  # Prawdziwe etykiety cyfr

# 2. Normalizacja danych
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 3. Klasteryzacja K-Means
kmeans = KMeans(n_clusters=10, random_state=42)
clusters = kmeans.fit_predict(X_scaled)
```

```
# 4. Redukcja wymiarów do 2D za pomocą PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

# 5. Wizualizacja klastrów na wykresie 2D
plt.figure(figsize=(8, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=clusters, cmap='viridis', alpha=0.6)
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],
            s=200, c='red', marker='X', label='Centroidy')

plt.title("K-Means na zbiorze Digits (redukcja PCA do 2D)")
plt.xlabel("Główna składowa 1")
plt.ylabel("Główna składowa 2")
plt.legend()
plt.show()
```



Wykres przedstawia wyniki klasteryzacji zbioru **Digits** za pomocą algorytmu **K-Means**, z wizualizacją w przestrzeni 2D po redukcji wymiarowości metodą **PCA**.

- **Kolorowe punkty** reprezentują próbki danych (obrazy cyfr) przypisane do różnych klastrów przez K-Means.
- **Czerwone "X"** to centroidy klastrów – średnie wartości punktów w danym klastrze.
- **Redukcja PCA** została zastosowana, aby sprowadzić dane z 64 wymiarów (piksele obrazów) do 2D, umożliwiając wizualizację struktury klastrów.

Widać, że centroidy znajdują się w centralnym obszarze, co sugeruje pewne nakładanie się grup, co może wynikać z podobieństwa niektórych cyfr w zbiorze.

3. Wnioski

W przeprowadzonej analizie skupień zastosowano algorytm **K-Means** do podziału danych zbioru **Digits** na 10 klastrów. Przed klasteryzacją dane zostały poddane **normalizacji** za pomocą **StandardScaler**, co pozwoliło na bardziej równomierne rozłożenie wartości cech i lepszą separację klastrów. Następnie wynik klasteryzacji został zwizualizowany w przestrzeni **2D** po redukcji wymiarowości przy użyciu **PCA**, co umożliwiło interpretację podziału danych. Na wykresie widać centroidy klastrów, które zostały oznaczone czerwonymi znacznikami.

Wyniki pokazują, że metoda **K-Means** była w stanie wyodrębnić pewne struktury w danych, jednak częściowe nakładanie się klastrów sugeruje, że niektóre cyfry mogą być trudne do odróżnienia w tej przestrzeni. Dalsza analiza mogłaby obejmować ocenę jakości klasteryzacji, np. za pomocą wskaźnika **Silhouette Score**, oraz eksperymenty z innymi metodami redukcji wymiarowości. Mimo to, przeprowadzona analiza pozwala lepiej zrozumieć strukturę danych i potencjalne zależności między klastrami.