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 10	Anna Więzik Informatyka II stopień, niestacjonarne, 1 semestr, gr.1b

1. Polecenie:

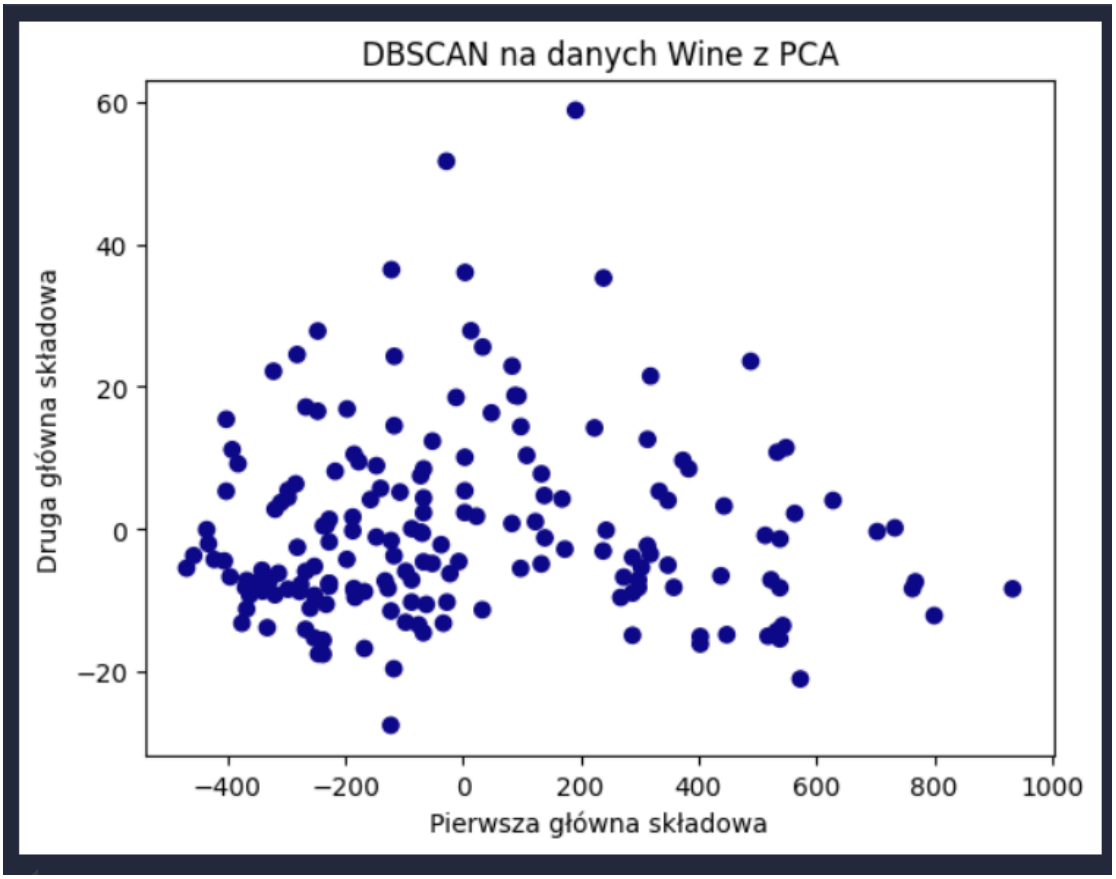
   W KNIME u˙zyj hierarchicznej analizy skupie´n na zbiorze Digits. Przedstaw dendrogram i wybierz optymalna liczbe klastr´ow.

   Link do repozytorium: https://github.com/AnaShiro/NoD1_2024

2. Opis programu opracowanego

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import load_wine
from sklearn.decomposition import PCA
from sklearn.cluster import DBSCAN

# Załaduj dane Wine
data = load_wine()
X = data.data

# Zastosuj PCA do redukcji wymiarów
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

# Zastosuj DBSCAN
dbscan = DBSCAN(eps=3, min_samples=5).fit(X_pca)

# Wizualizacja wyników
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=dbscan.labels_, cmap='plasma')
plt.xlabel('Pierwsza główna składowa')
plt.ylabel('Druga główna składowa')
plt.title('DBSCAN na danych Wine z PCA')
plt.show()
```

19.0s



3. Wnioski

Analiza skupień to technika uczenia nienadzorowanego, której celem jest podział danych na grupy (klastry) na podstawie ich podobieństwa. Jest często wykorzystywana jako etap wstępny w uczeniu maszynowym, szczególnie do redukcji wymiarowości i uproszczenia złożoności danych, tworzenia cech na podstawie klastrów (clustering-based features), wstępnej analizy struktury danych (eksploracja danych) oraz detekcji anomalii i punktów odstających. K-Means jest jednym z najczęściej stosowanych algorytmów analizy skupień. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) grupuje dane na podstawie gęstości punktów, jednocześnie identyfikując anomalie. Hierarchical Clustering tworzy strukturę drzewa (dendrogram), co pozwala analizować dane na różnych poziomach szczegółowości.