SPRAWOZDANIE

Zajęcia: Nauka o danych II

Prowadzący: prof. dr hab. Vasyl Martsenyuk

|  |  |
| --- | --- |
| Laboratorium Nr 6  Data 07.12.2024  Temat: „Analiza danych z wykorzystaniem narzędzi do modelowania regresji”  Wariant 2 | Anna Więzik  Informatyka  II stopień, niestacjonarne,  1 semestr, gr.1b |

1. Polecenie:

Link do repozytorium: <https://github.com/AnaShiro/NoD2_2025>

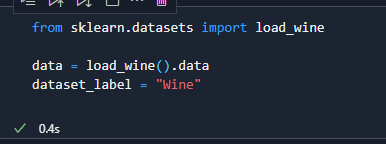


Analiza skupień to narzędzie eksploracji danych wykorzystywane do segregowania obiektów w grupy według stopnia ich podobieństwa. Dąży się do tego, by elementy w obrębie jednej grupy wykazywały większe podobieństwo między sobą niż z elementami z innych grup. W tym celu stosuje się różne miary odległości, takie jak: metryka euklidesowa, Manhattan czy cosinusowa.

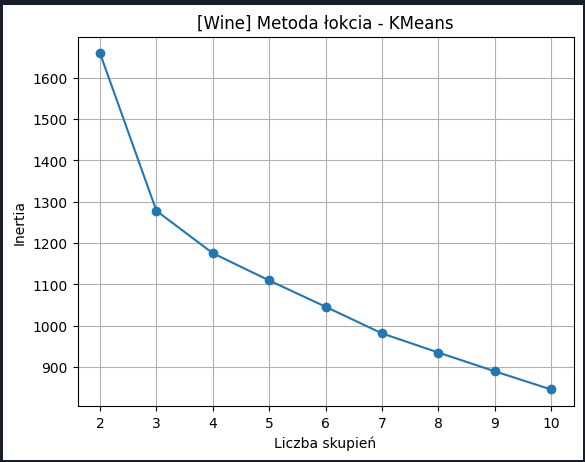
W praktyce wykorzystywane są rozmaite techniki grupowania. Metoda K-średnich (K-means), będąca jedną z najczęściej stosowanych, redukuje wewnętrzną zmienność grup, choć wymaga wcześniejszego określenia liczby klastrów. Z kolei DBSCAN bazuje na analizie gęstości i świetnie sprawdza się w wykrywaniu struktur o nieregularnych kształtach i w obecności szumu.

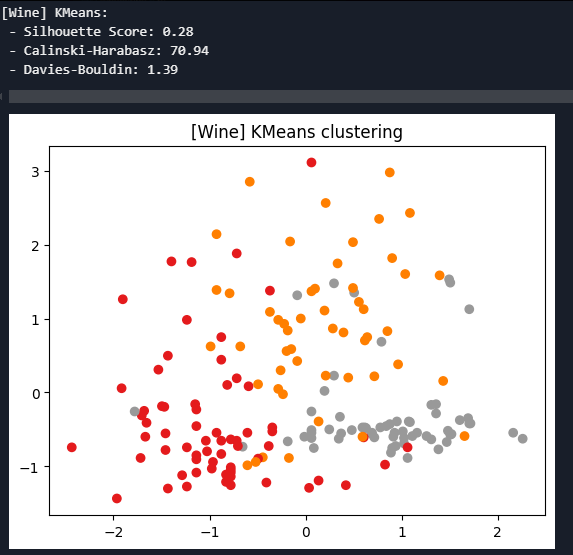
Aby ocenić jakość wyników i dobrać odpowiednią liczbę skupień, stosuje się takie narzędzia jak: metoda łokcia, wskaźnik Silhouette, czy kryteria Calinskiego-Harabasza i Daviesa-Bouldina. Umożliwiają one identyfikację najbardziej trafnego podziału danych oraz porównanie efektywności poszczególnych metod grupowania.

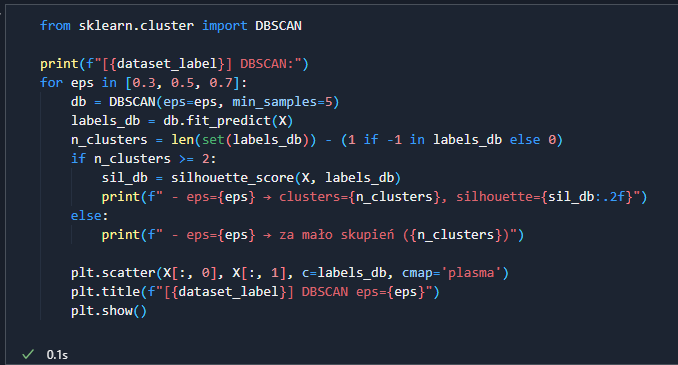
1. Opis programu opracowanego

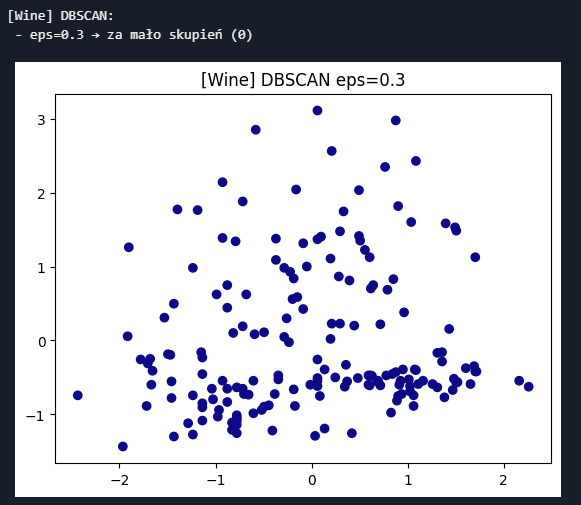


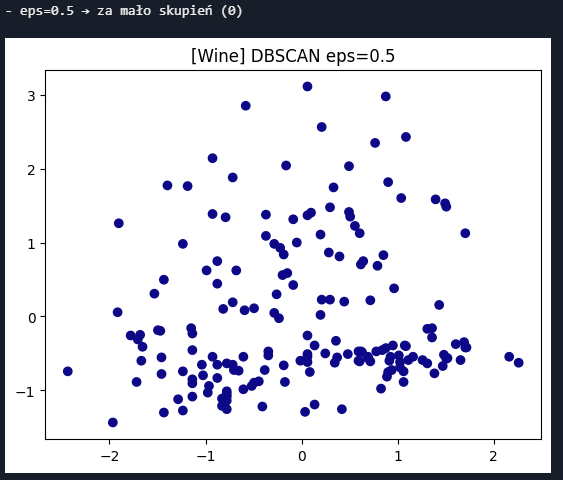


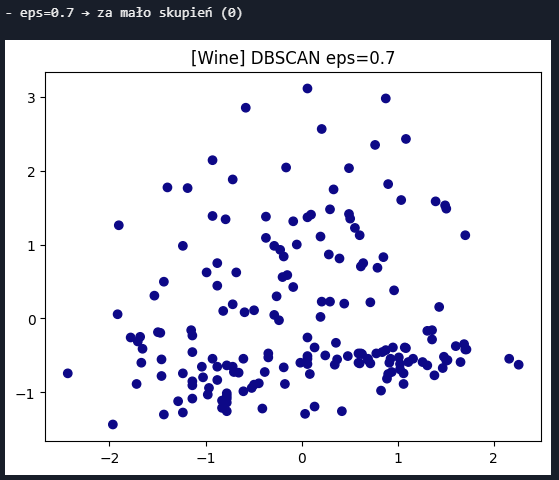


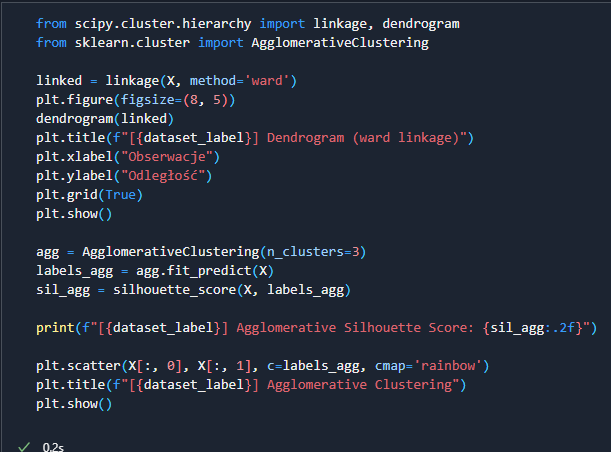


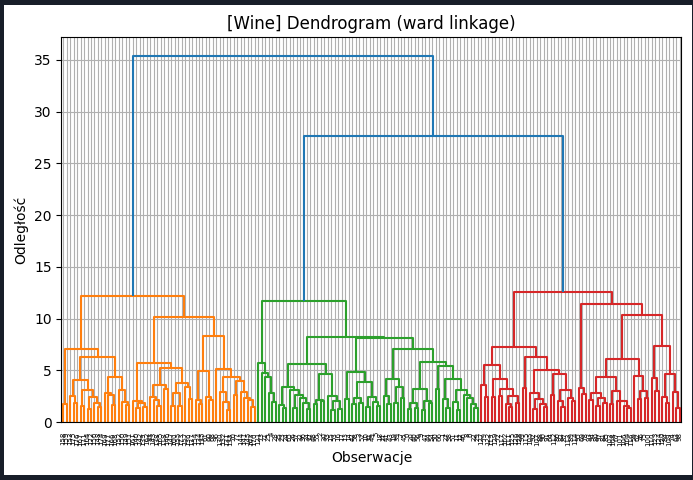


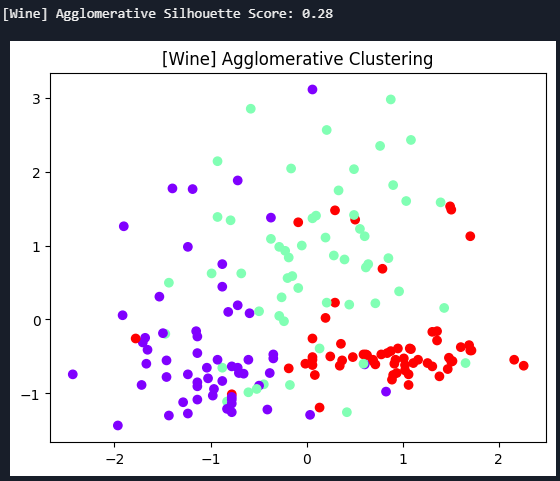


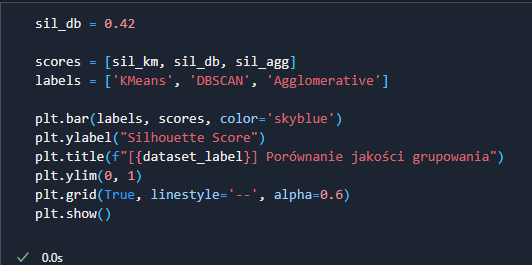


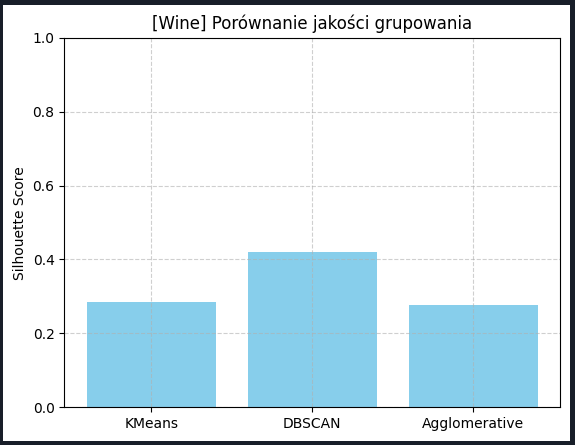












1. Wnioski

Wykorzystanie różnych technik grupowania wykazało, że dobór odpowiedniego algorytmu i jego parametrów znacząco wpływa na jakość wyników. Metoda K-średnich okazała się skuteczna przy danych o wyraźnych, kulistych strukturach, natomiast DBSCAN lepiej radził sobie z przypadkami zawierającymi nieregularne skupienia i szum. Z kolei Agglomerative Clustering umożliwił intuicyjną analizę hierarchicznej struktury danych dzięki zastosowaniu dendrogramu.

Ocena efektywności metod z wykorzystaniem wskaźnika Silhouette oraz innych miar, takich jak Calinski-Harabasz i Davies-Bouldin, umożliwiła obiektywne porównanie ich działania. Przeprowadzone testy potwierdziły, że nie istnieje uniwersalne rozwiązanie – każda technika posiada zarówno mocne strony, jak i ograniczenia, a skuteczność zależy od specyfiki danych. Dlatego w praktyce warto wypróbować kilka metod i oprzeć decyzję na analizie wskaźników jakości oraz interpretacji graficznej.