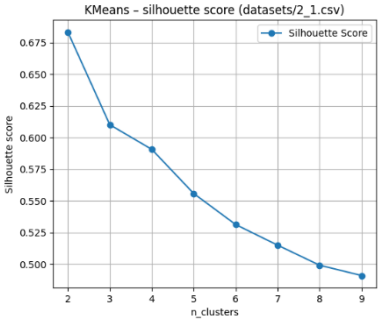
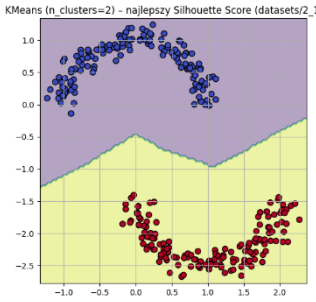
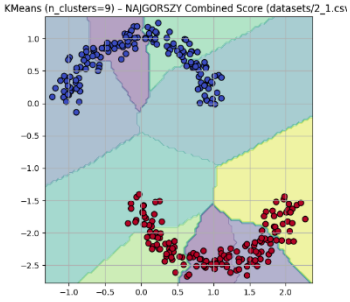
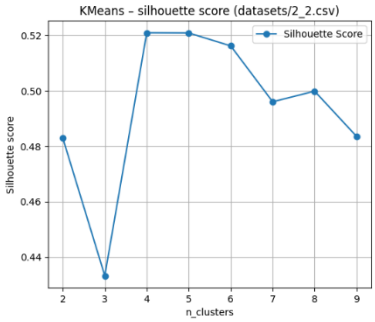
Jakub Tomaszewski

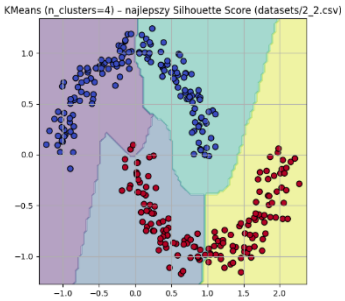
Projekt I zadanie I

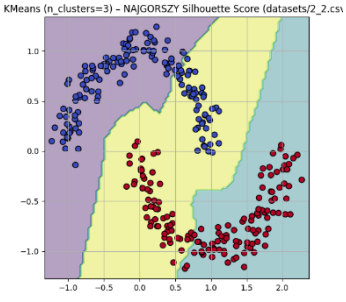


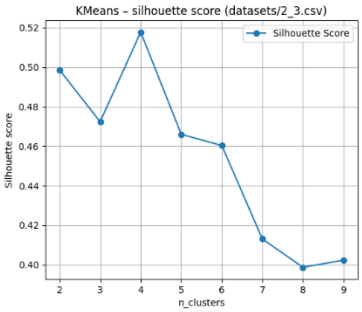


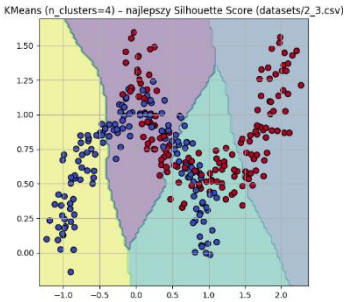


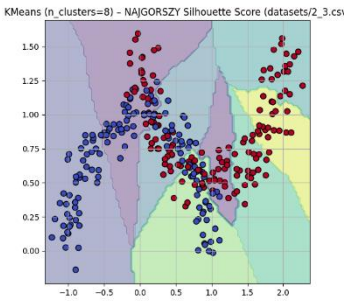


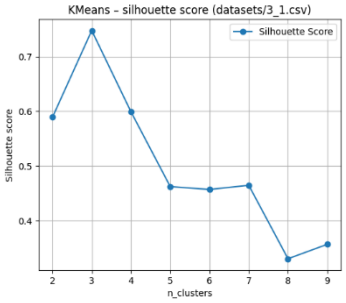


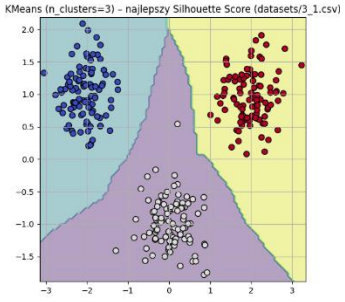


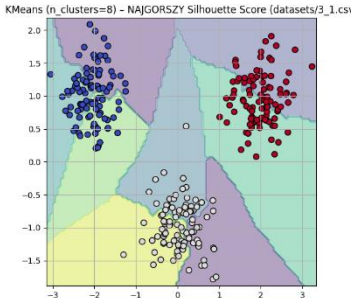


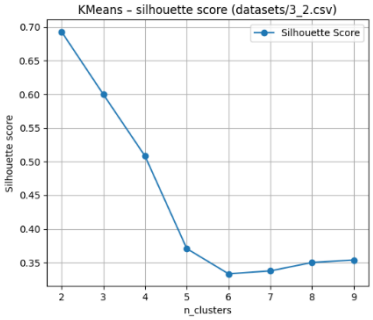


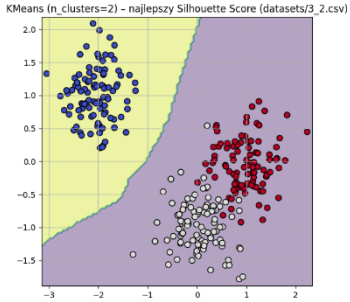


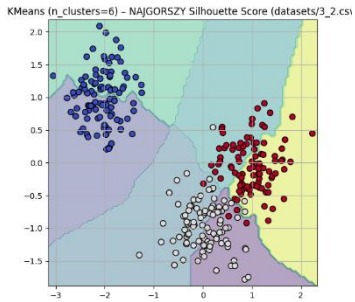


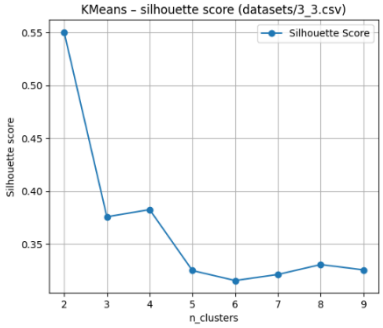


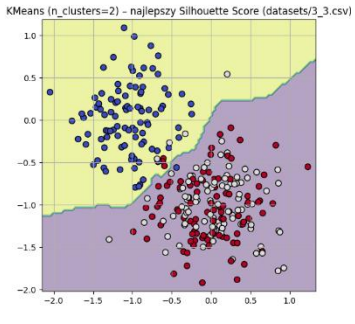


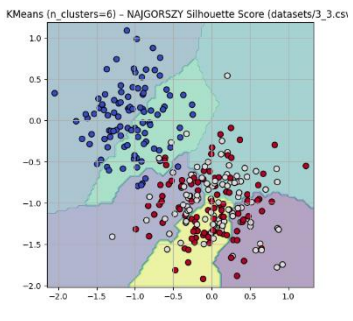


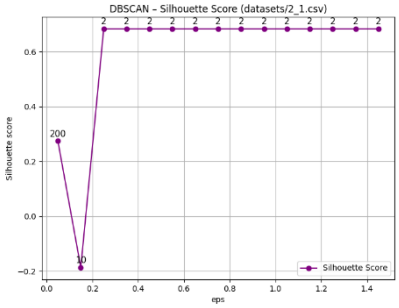


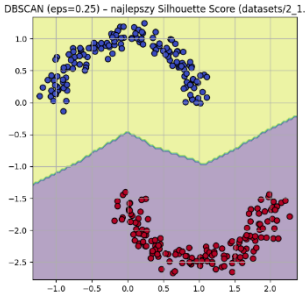


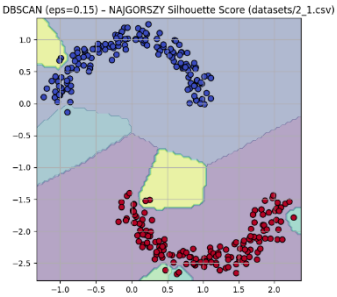


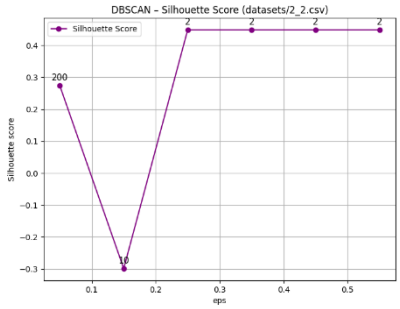


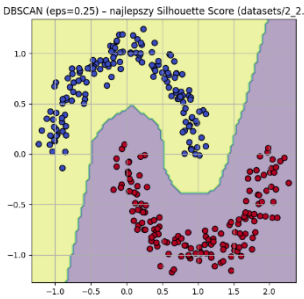


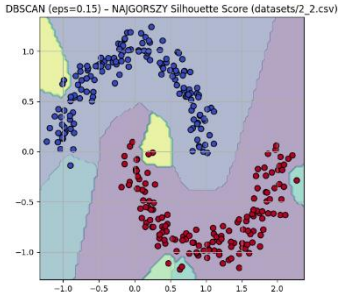

DBSCAN – Silhouette Score (datasets/2_1.csv)
Silhouette Score
Silhouette score
eps

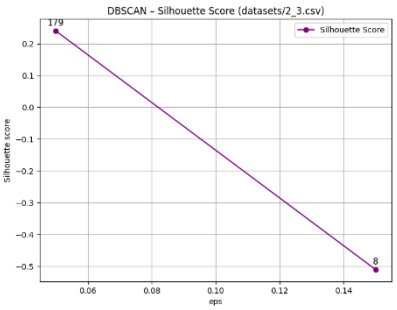

DBSCAN (eps=0.25) – najlepszy Silhouette Score (datasets/2_1.csv)

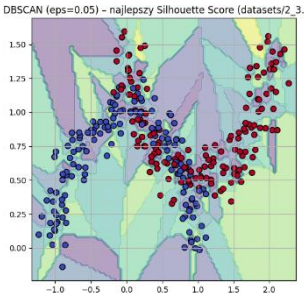

DBSCAN (eps=0.15) – NAJGORSZY Silhouette Score (datasets/2_1.csv)

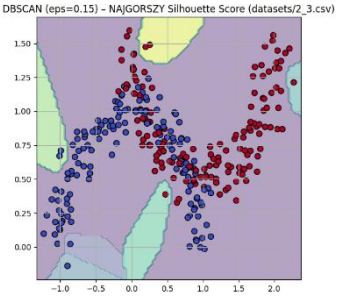

DBSCAN – Silhouette Score (datasets/2_2.csv)
Silhouette Score
Silhouette score
eps

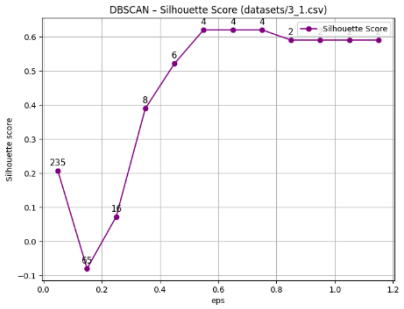

DBSCAN (eps=0.25) – najlepszy Silhouette Score (datasets/2_2.csv)

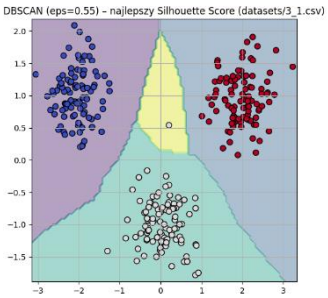

DBSCAN (eps=0.15) – NAJGORSZY Silhouette Score (datasets/2_2.csv)

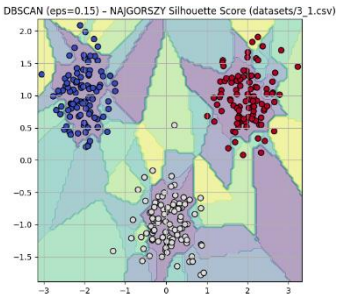

DBSCAN – Silhouette Score (datasets/2_3.csv)
Silhouette Score
Silhouette score
eps

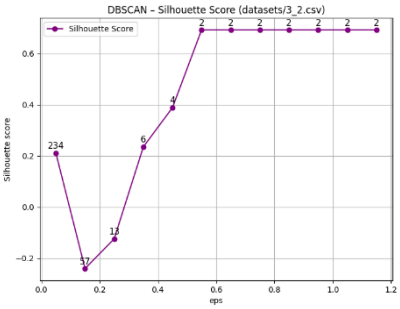

DBSCAN (eps=0.05) – najlepszy Silhouette Score (datasets/2_3.csv)

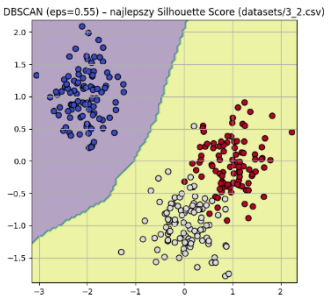

DBSCAN (eps=0.15) – NAJGORSZY Silhouette Score (datasets/2_3.csv)

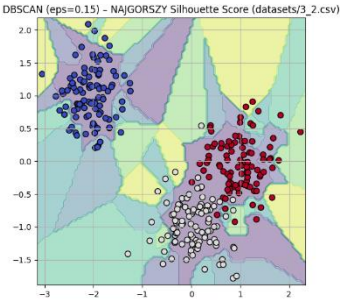

DBSCAN – Silhouette Score (datasets/3_1.csv)
Silhouette Score
Silhouette score
eps

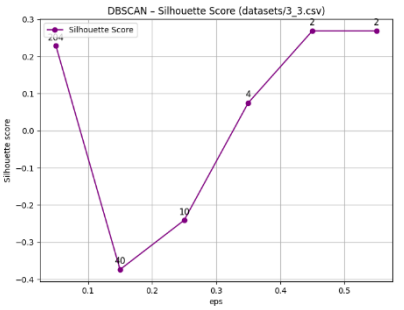

DBSCAN (eps=0.55) – najlepszy Silhouette Score (datasets/3_1.csv)

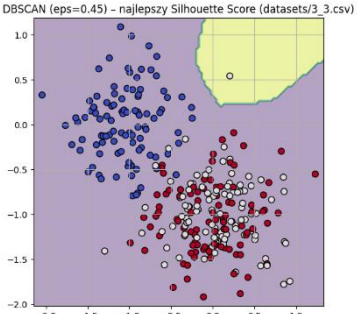

DBSCAN (eps=0.15) – NAJGORSZY Silhouette Score (datasets/3_1.csv)

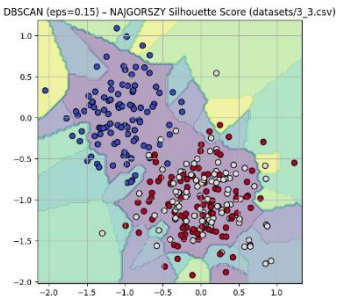

DBSCAN – Silhouette Score (datasets/3_2.csv)
Silhouette Score
Silhouette score
eps

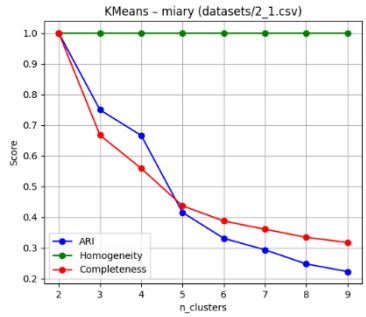

DBSCAN (eps=0.55) – najlepszy Silhouette Score (datasets/3_2.csv)

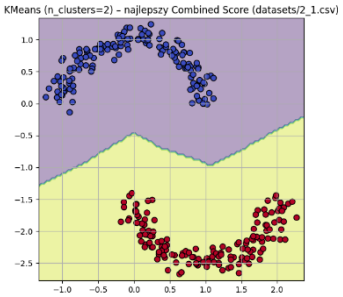

DBSCAN (eps=0.15) – NAJGORSZY Silhouette Score (datasets/3_2.csv)

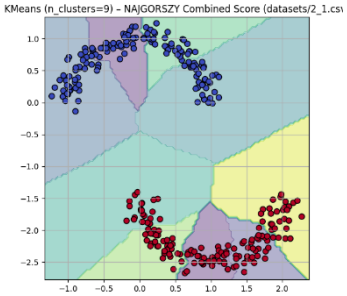

DBSCAN – Silhouette Score (datasets/3_3.csv)
Silhouette Score
Silhouette score
eps

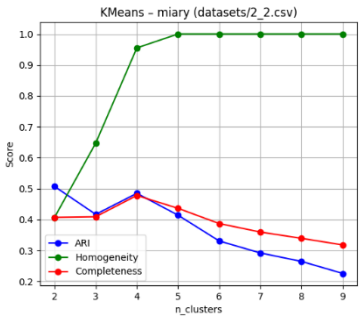

DBSCAN (eps=0.45) – najlepszy Silhouette Score (datasets/3_3.csv)

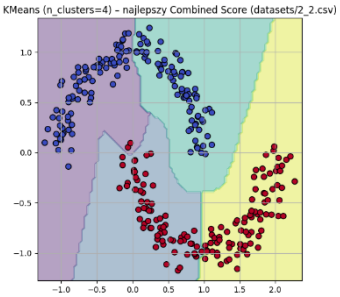

DBSCAN (eps=0.15) – NAJGORSZY Silhouette Score (datasets/3_3.csv)

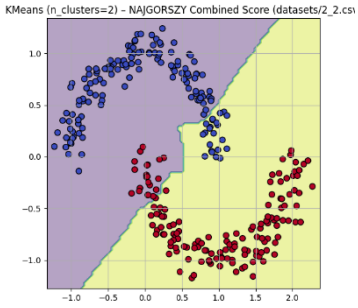

KMeans – miary (datasets/2_1.csv)
Score
n_clusters
ARI
Homogeneity
Completeness

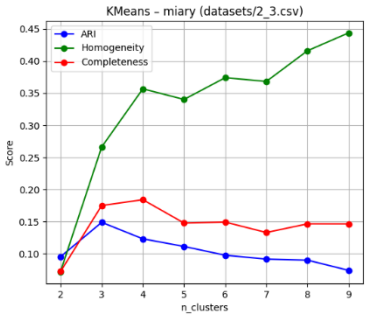

KMeans (n_clusters=2) – najlepszy Combined Score (datasets/2_1.csv)

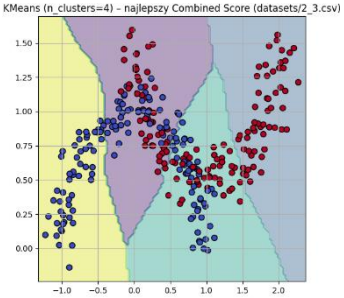

KMeans (n_clusters=9) – NAJGORSZY Combined Score (datasets/2_1.csv

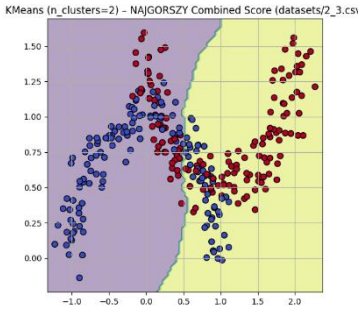

KMeans – miary (datasets/2_2.csv)
Score
n_clusters
ARI
Homogeneity
Completeness

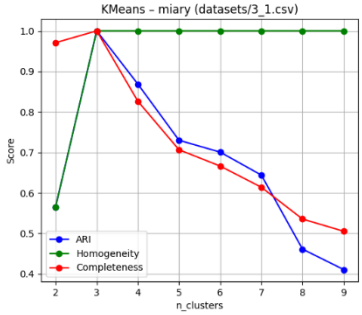

KMeans (n_clusters=4) – najlepszy Combined Score (datasets/2_2.csv)

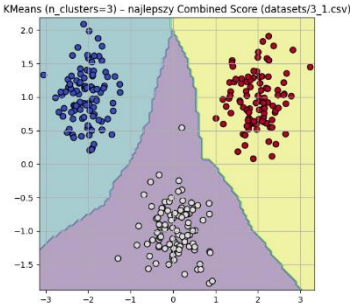

KMeans (n_clusters=2) – NAJGORSZY Combined Score (datasets/2_2.csv

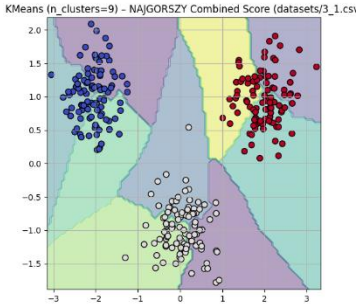

KMeans – miary (datasets/2_3.csv)
Score
n_clusters
ARI
Homogeneity
Completeness

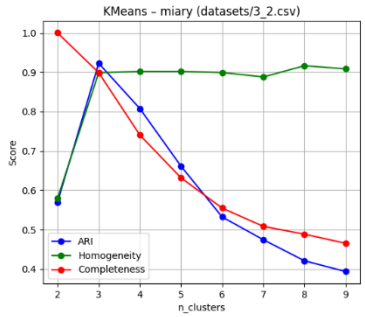

KMeans (n_clusters=4) – najlepszy Combined Score (datasets/2_3.csv)

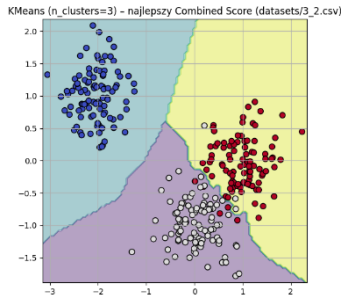

KMeans (n_clusters=2) – NAJGORSZY Combined Score (datasets/2_3.csv

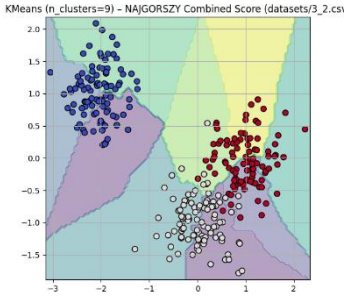

KMeans – miary (datasets/3_1.csv)
Score
n_clusters
ARI
Homogeneity
Completeness

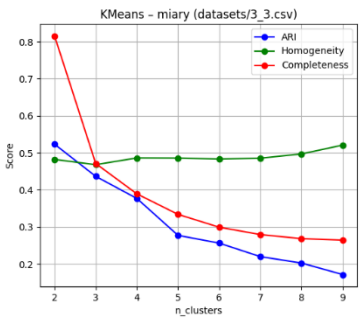

KMeans (n_clusters=3) – najlepszy Combined Score (datasets/3_1.csv)

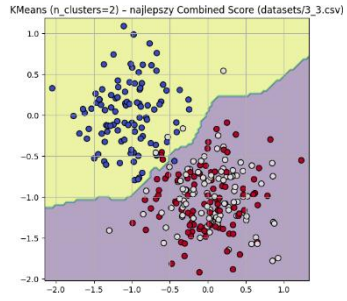

KMeans (n_clusters=9) – NAJGORSZY Combined Score (datasets/3_1.csv

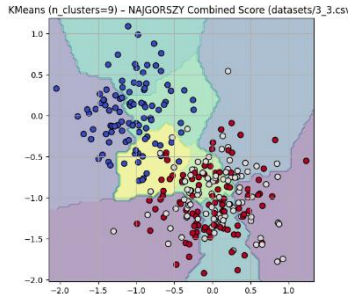

KMeans – miary (datasets/3_2.csv)
Score
n_clusters
ARI
Homogeneity
Completeness

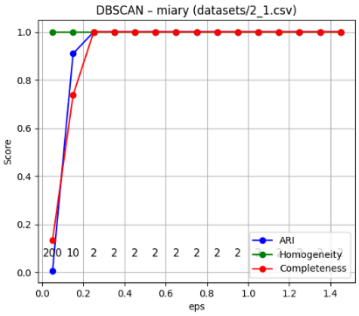

KMeans (n_clusters=3) – najlepszy Combined Score (datasets/3_2.csv)

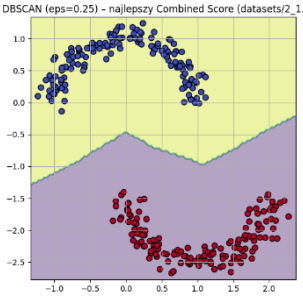

KMeans (n_clusters=9) – NAJGORSZY Combined Score (datasets/3_2.csv

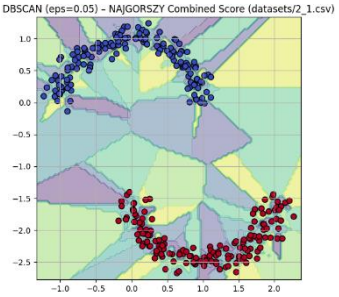

KMeans – miary (datasets/3_3.csv)
Score
n_clusters
ARI
Homogeneity
Completeness

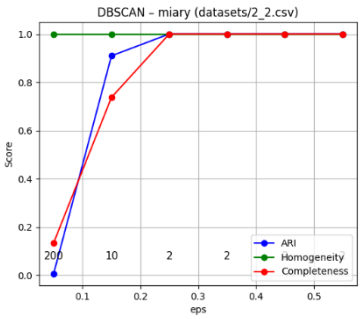

KMeans (n_clusters=2) – najlepszy Combined Score (datasets/3_3.csv)

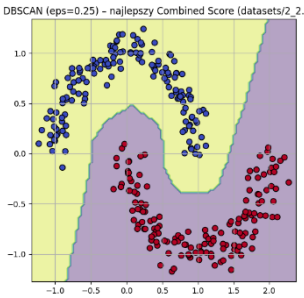

KMeans (n_clusters=9) – NAJGORSZY Combined Score (datasets/3_3.csv

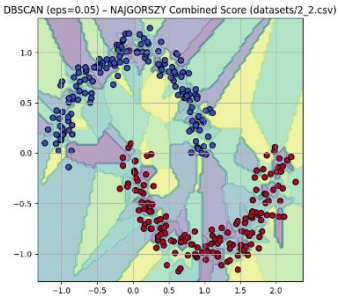

DBSCAN – miary (datasets/2_1.csv)
Score
eps
ARI
Homogeneity
Completeness
200 10 2 2 2 2 2 2 2 2

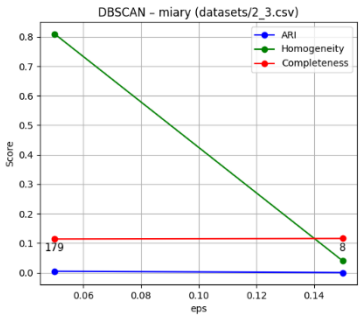

DBSCAN (eps=0.25) – najlepszy Combined Score (datasets/2_1.csv)

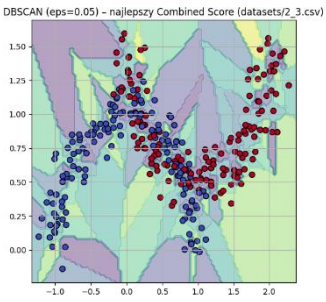

DBSCAN (eps=0.05) – NAJGORSZY Combined Score (datasets/2_1.csv)

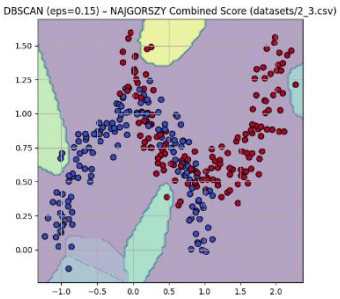

DBSCAN – miary (datasets/2_2.csv)
Score
eps
ARI
Homogeneity
Completeness
200 10 2 2

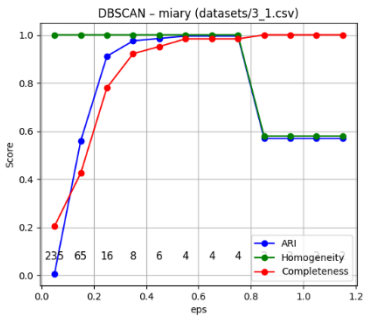

DBSCAN (eps=0.25) – najlepszy Combined Score (datasets/2_2.csv)

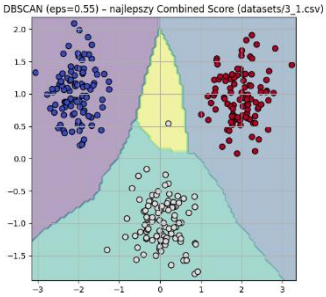

DBSCAN (eps=0.05) – NAJGORSZY Combined Score (datasets/2_2.csv)

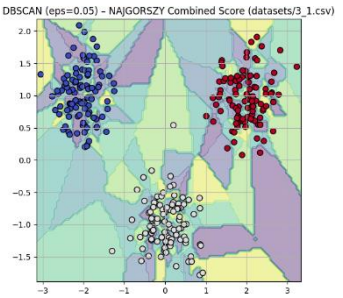

DBSCAN – miary (datasets/2_3.csv)
Score
eps
ARI
Homogeneity
Completeness
179 8

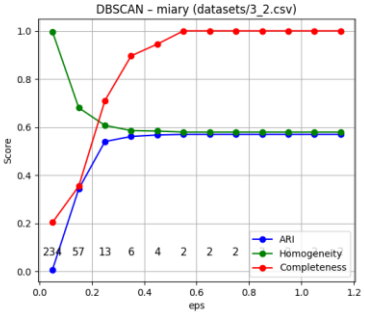

DBSCAN (eps=0.05) – najlepszy Combined Score (datasets/2_3.csv)

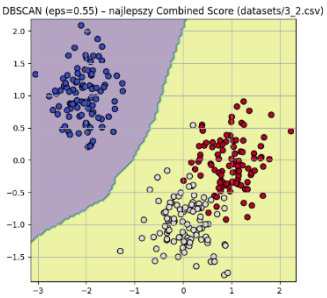

DBSCAN (eps=0.15) – NAJGORSZY Combined Score (datasets/2_3.csv)

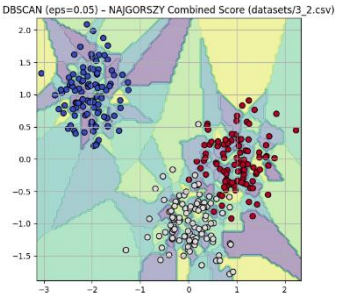

DBSCAN – miary (datasets/3_1.csv)
Score
eps
ARI
Homogeneity
Completeness
235 65 16 8 6 4 4 4

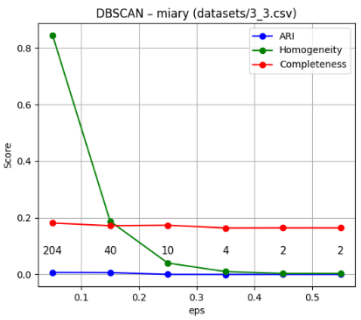

DBSCAN (eps=0.55) – najlepszy Combined Score (datasets/3_1.csv)

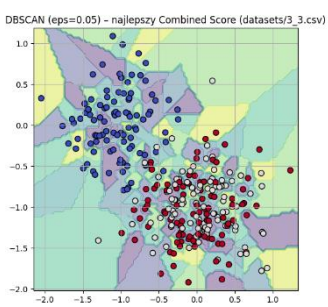

DBSCAN (eps=0.05) – NAJGORSZY Combined Score (datasets/3_1.csv)

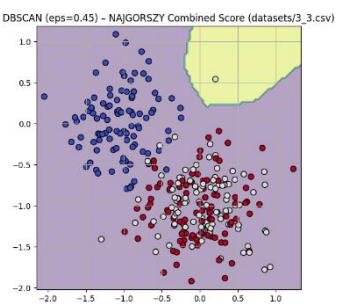

DBSCAN – miary (datasets/3_2.csv)
Score
eps
ARI
Homogeneity
Completeness
234 57 13 6 4 2 2 2


DBSCAN (eps=0.55) – najlepszy Combined Score (datasets/3_2.csv)


DBSCAN (eps=0.05) – NAJGORSZY Combined Score (datasets/3_2.csv)


DBSCAN – miary (datasets/3_3.csv)
Score
eps
ARI
Homogeneity
Completeness
204 40 10 4 2 2


DBSCAN (eps=0.05) – najlepszy Combined Score (datasets/3_3.csv)


DBSCAN (eps=0.45) – NAJGORSZY Combined Score (datasets/3_3.csv)

Wnioski z przeprowadzonych eksperymentów na zbiorach sztucznych:

Przeprowadzone eksperymenty pokazały, że zarówno zbyt mała, jak i zbyt duża liczba klastrów w metodzie KMeans, jak również nieoptymalna wartość parametru eps w DBSCAN, prowadzą do spadku wartości silhouette score, co oznacza pogorszenie jakości klasteryzacji. W zbiorach, gdzie nie obserwujemy wyraźnych maksimów silhouette score, można wnioskować, że dane nie posiadają wyraźnej struktury skupisk lub są one mocno nakładające się.

Na podstawie wyników eksperymentów można przyjąć, że wartość silhouette score powyżej około **0.6** daje szansę na wysokiej jakości klasteryzację. Przykładowo, w KMeans dla zbiorów 1_1 i 3_1 uzyskano silhouette score powyżej 0.7, co odpowiadało dokładnej klasteryzacji potwierdzonej wizualizacją na diagramach Woronoja. Jednak w przypadku DBSCAN, mimo uzyskania podobnie wysokich wartości silhouette score (powyżej 0.6 dla zbiorów 3_1 i 3_2), nie zawsze przekładało się to na dobrą separację punktów w klastrach.

Wskaźniki homogeneity score i completeness score dostarczają dodatkowej informacji o jakości klasteryzacji: homogeneity mierzy, czy każdy klaster zawiera obiekty tylko jednej klasy, natomiast completeness ocenia, czy wszystkie obiekty danej klasy zostały przypisane do tego samego klastra. Adjusted Rand Index (ARI) natomiast mierzy zgodność podziału na klastry z rzeczywistymi etykietami, uwzględniając przypadkowe dopasowania. Wysoka wartość ARI oznacza, że klasteryzacja dobrze odzwierciedla rzeczywisty podział klas.

Do oceny ogólnej jakości klasteryzacji zastosowano średnią ważoną trzech miar:

0.4 * ARI + 0.3 * homogeneity + 0.3 * completeness

Wartość równa 1.0 oznaczała idealną klasteryzację, co zaobserwowano m.in. dla zbiorów 2_1 i 3_1 w metodzie KMeans. Ciekawym przypadkiem był zbiór 3_1 w DBSCAN, gdzie średnia wyniosła około 0.99, co wynikało z utworzenia osobnego klastra dla pojedynczego punktu – co potwierdziła analiza diagramu Woronoja.

Podsumowując, uzyskanie wartości 1.0 dla każdej z analizowanych miar w eksperymencie drugim świadczy o maksymalnej efektywności klasteryzacji. Warto jednak zawsze wspierać analizę wizualizacjami klastrów i etykiet, a w przypadku danych o większej liczbie wymiarów – stosować metody redukcji wymiarowości, takie jak PCA.

Wnioski z przeprowadzonych eksperymentów na zbiorach rzeczywistych:

Iris

Dla zbioru **Iris** zarówno KMeans, jak i DBSCAN osiągnęły najlepsze wyniki przy podziale na dwa klastry, mimo że rzeczywista liczba gatunków wynosi trzy. Wynika to z faktu, że dwa gatunki (Versicolor i Virginica) są do siebie bardzo podobne pod względem cech, co utrudnia ich rozdzielenie bez nadzoru.

Wartości miar (Silhouette Score: 0.582, ARI: 0.568) wskazują na umiarkowaną zgodność z rzeczywistym podziałem. Wysoka kompletność (1.0) oznacza, że wszystkie próbki jednej klasy są przypisane do tego samego klastra, jednak niska homogeniczność sugeruje mieszanie się klas w jednym klastrze.

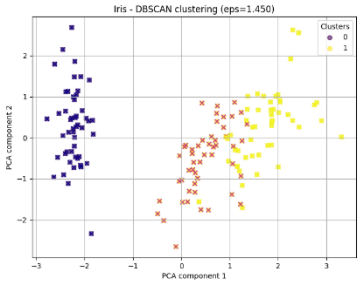
Wnioski te potwierdzają wizualizacje rzutów cech na dwuwymiarowe przestrzenie (np. długość i szerokość płatka), gdzie Setosa jest wyraźnie oddzielona, a dwa pozostałe gatunki tworzą nakładające się skupiska. Wykresy PCA oraz wykresy wartości miar (Silhouette, ARI, Homogeneity, Completeness) również pokazują, że podział na dwa klastry jest optymalny z punktu widzenia algorytmów, choć nie w pełni zgodny z rzeczywistym podziałem biologicznym.

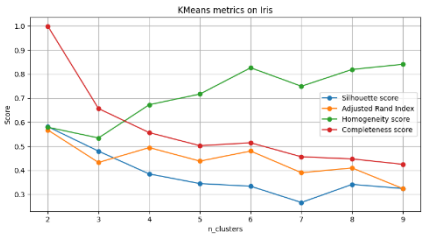
Wine

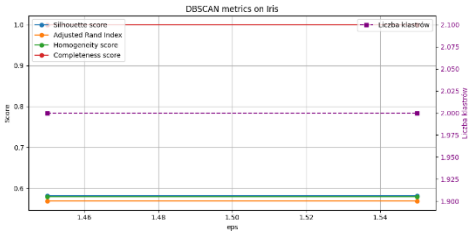
W przypadku zbioru **Wine** najlepsze wyniki uzyskano dla KMeans z trzema klastrami (n_clusters=3), co odpowiada liczbie klas w zbiorze. Wysokie wartości ARI (0.897), homogeniczności (0.879) i completeness (0.873) świadczą o bardzo dobrej zgodności klasteryzacji z rzeczywistym podziałem. DBSCAN nie znalazł sensownych klastrów dla testowanych parametrów, co sugeruje brak wyraźnych struktur gęstościowych w danych lub konieczność innego doboru parametrów.

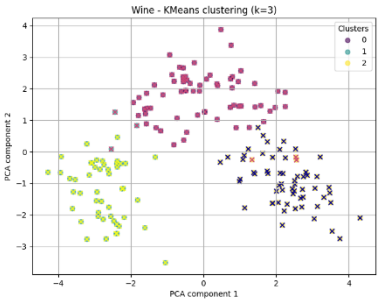
Breast Cancer Wisconsin

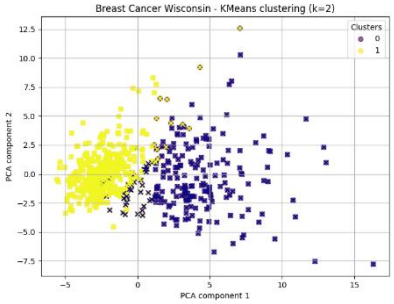
Dla zbioru **Breast Cancer Wisconsin** KMeans najlepiej dzieli dane na dwa klastry, co odpowiada rzeczywistemu podziałowi na przypadki łagodne i złośliwe. Wartości miar (Silhouette Score: 0.345, ARI: 0.677) są przyzwoite, choć niższe niż dla zbioru Wine, co może wynikać z większego nakładania się cech między klasami. DBSCAN ponownie nie znalazł sensownych klastrów.

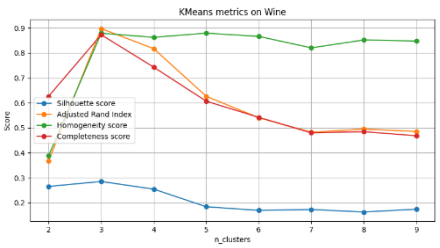












Zbiór danych	Algorytm	Najlepszy parametr	Silhouette Score	ARI	Homogeneity	Completeness	Combined Score
Iris	KMeans	2	0.582	0.568	0.579	1.000	0.701
Iris	DBSCAN	1.45	0.582	0.568	0.579	1.000	0.701
Wine	KMeans	3	0.285	0.897	0.879	0.873	0.885
Wine	DBSCAN	brak	brak	brak	brak	brak	brak
Breast Cancer Wisconsin	KMeans	2	0.345	0.677	0.551	0.573	0.608
Breast Cancer Wisconsin	DBSCAN	brak	brak	brak	brak	brak	brak