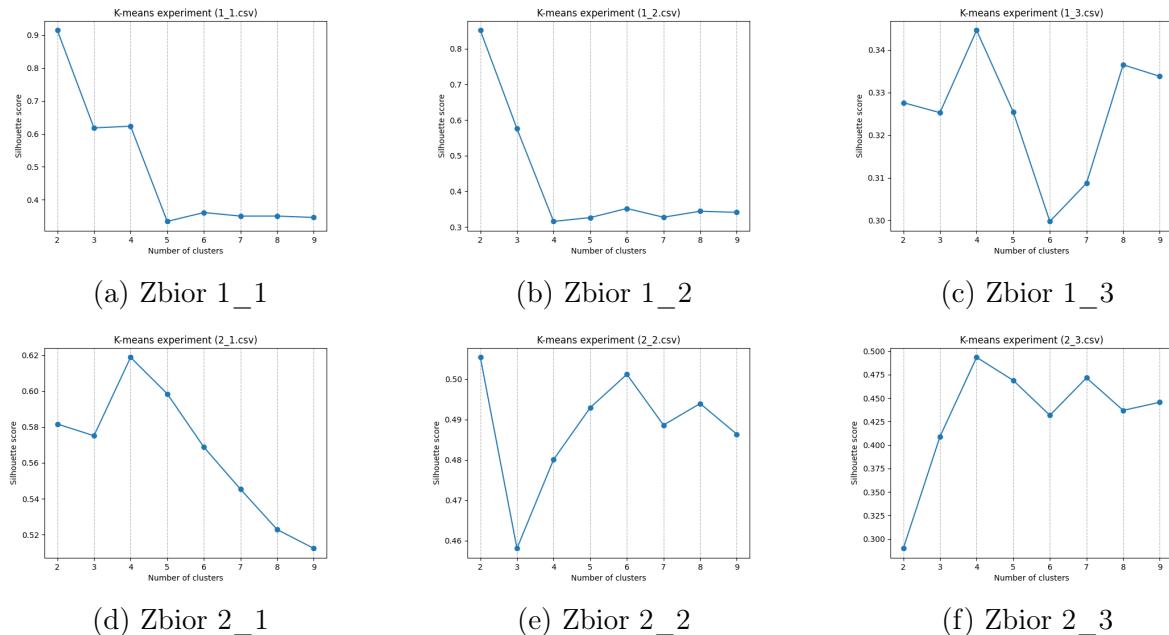
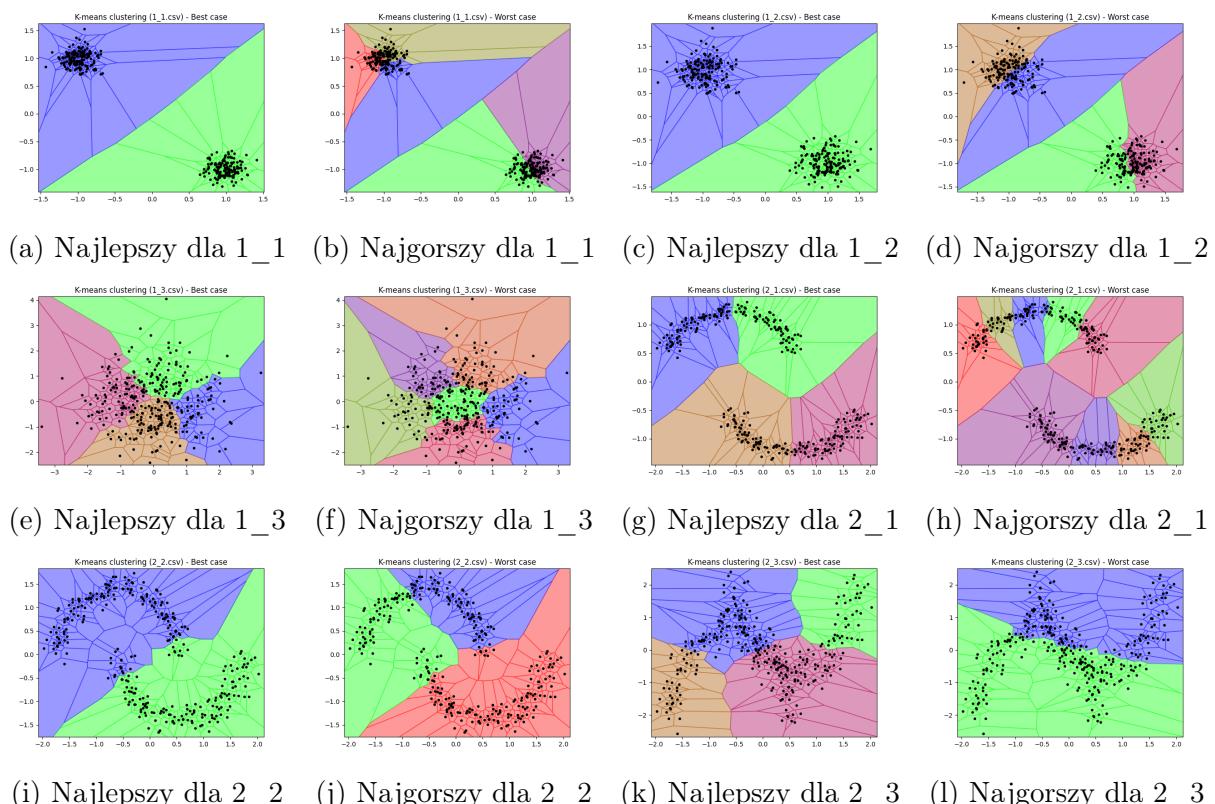


1 Eksperyment 1: Metoda K-Means

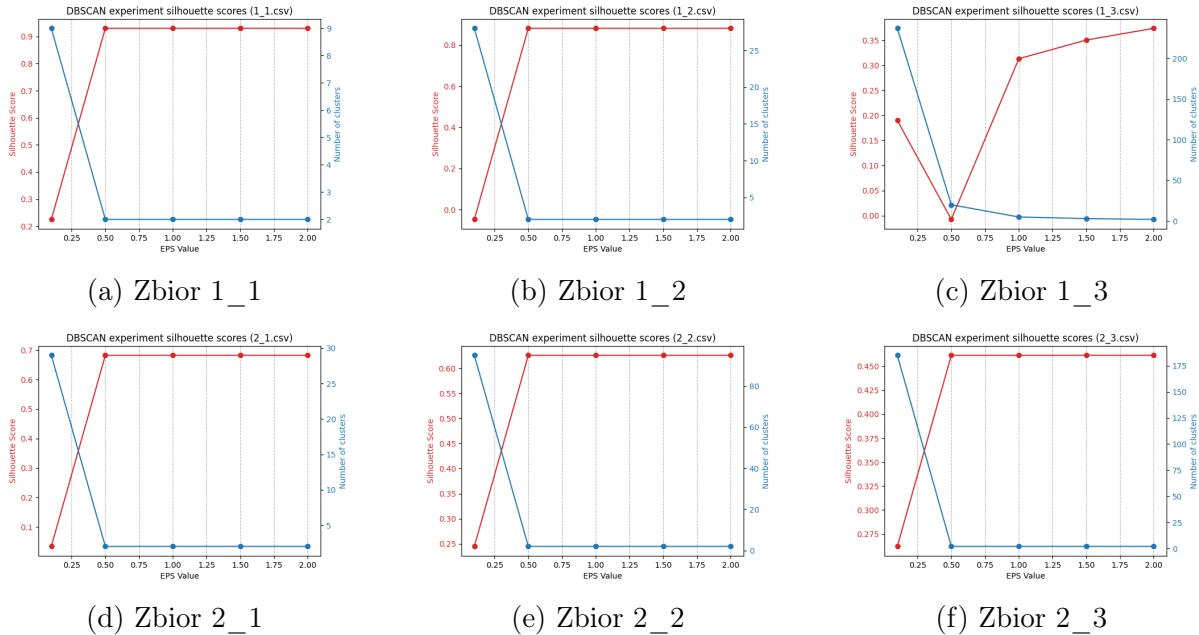


Rysunek 1: Zmiana wartości silhouette score dla wszystkich zbiorów w zależności od parametru n_clusters w metodzie K-means

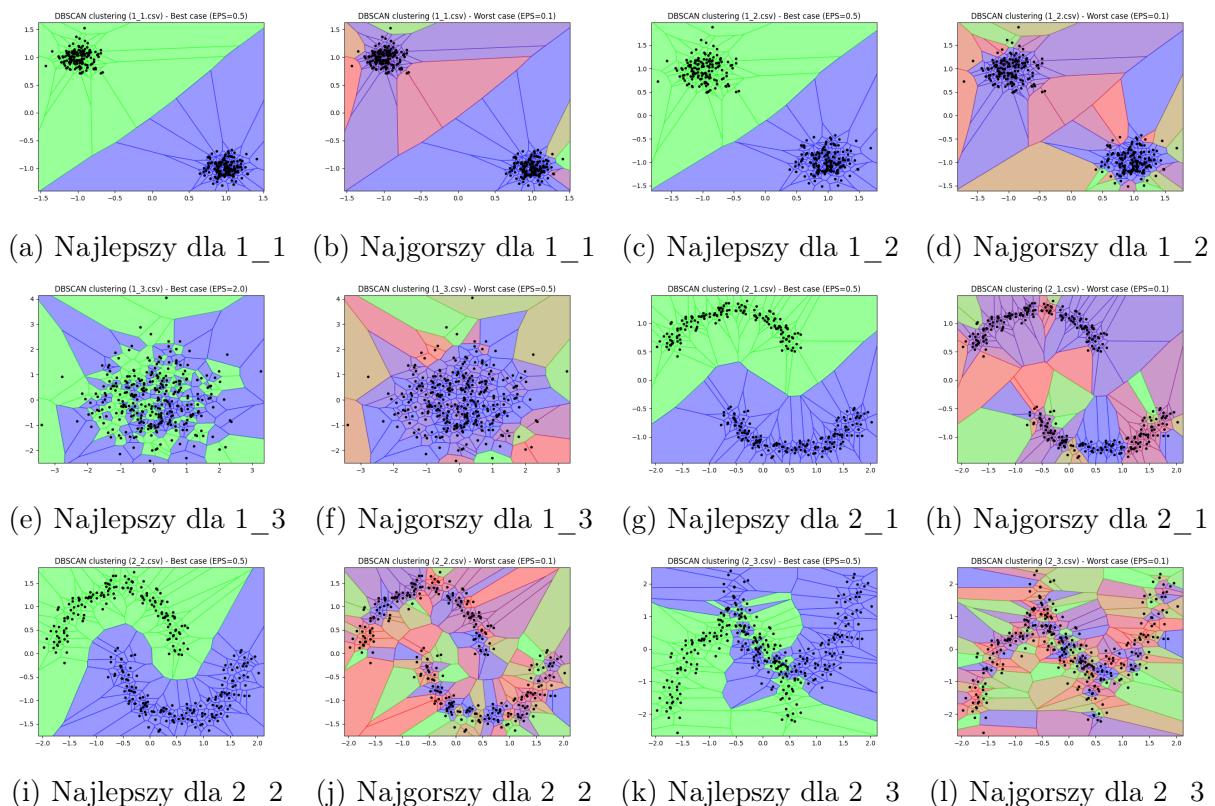


Rysunek 2: Wizualizacja klastrów dla wszystkich zbiorów na diagramie Voronoia dla najlepszego i najgorszego przypadku w metodzie K-means

2 Eksperyment 1: Metoda DBSCAN

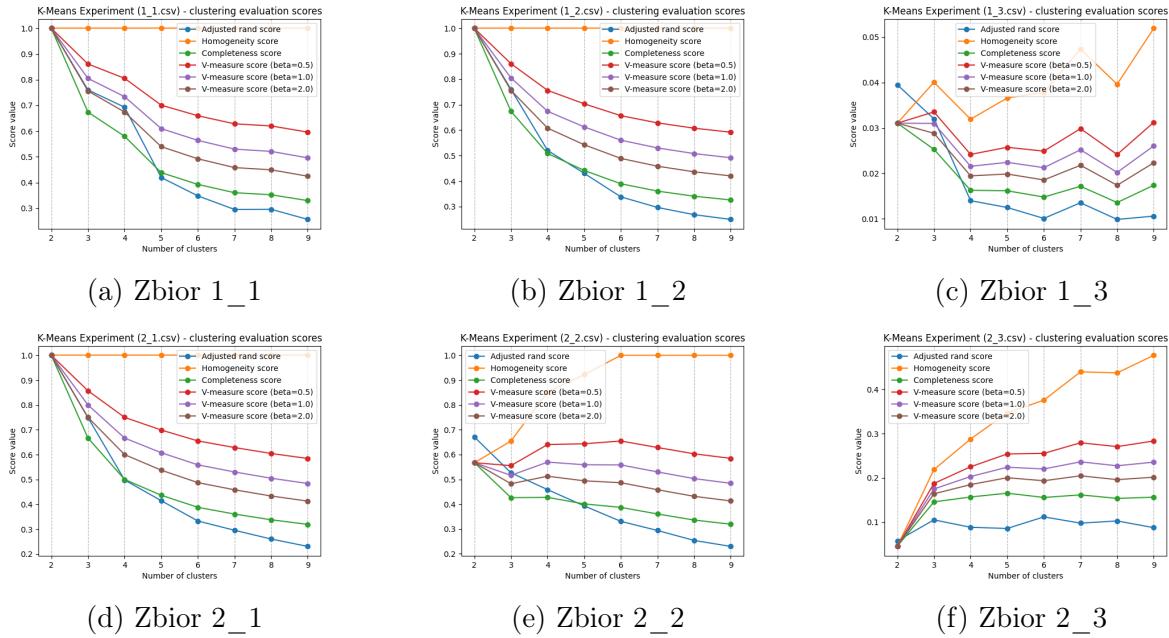


Rysunek 3: Zmiana wartości silhouette score oraz n_clusters dla wszystkich zbiorów w zależności od zmieniającego się parametru eps w metodzie DBSCAN

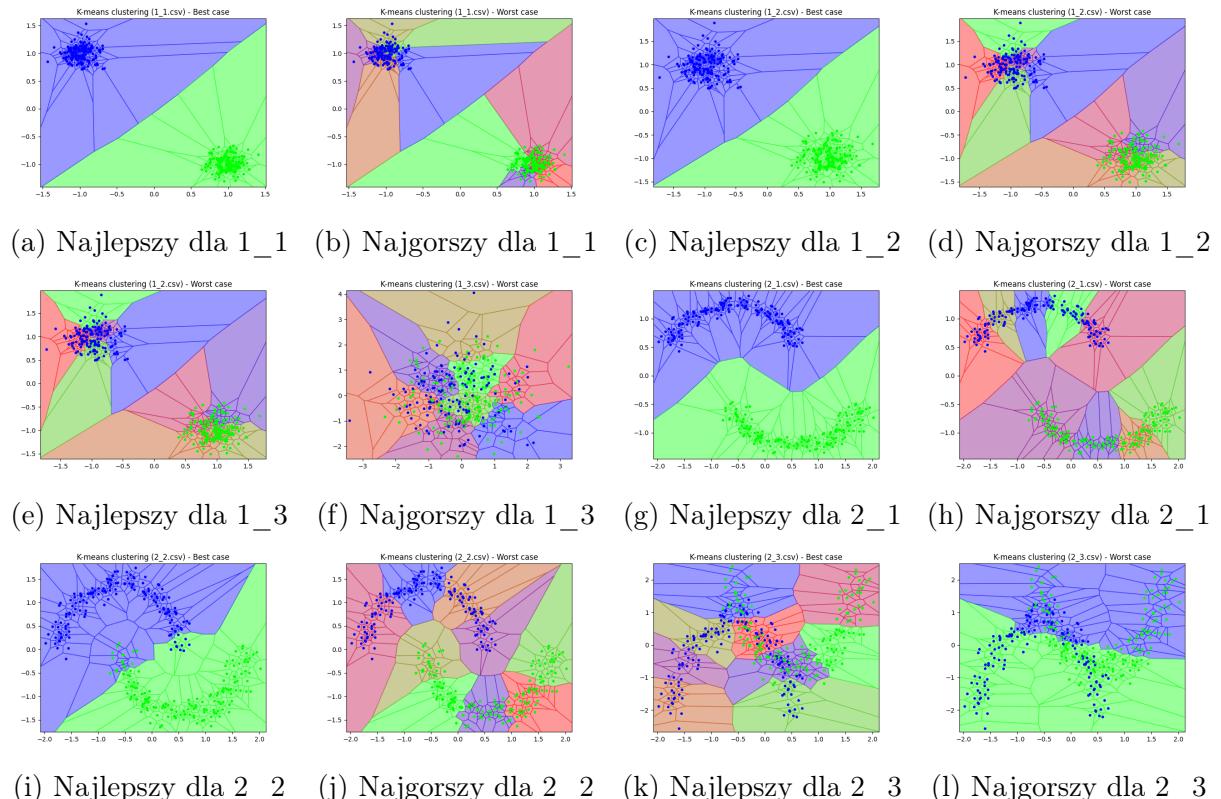


Rysunek 4: Wizualizacja klastrów dla wszystkich zbiorów na diagramie Voronoia dla najlepszego i najgorszego przypadku w metodzie DBSCAN

3 Eksperyment 2: Metoda K-Means z etykietami

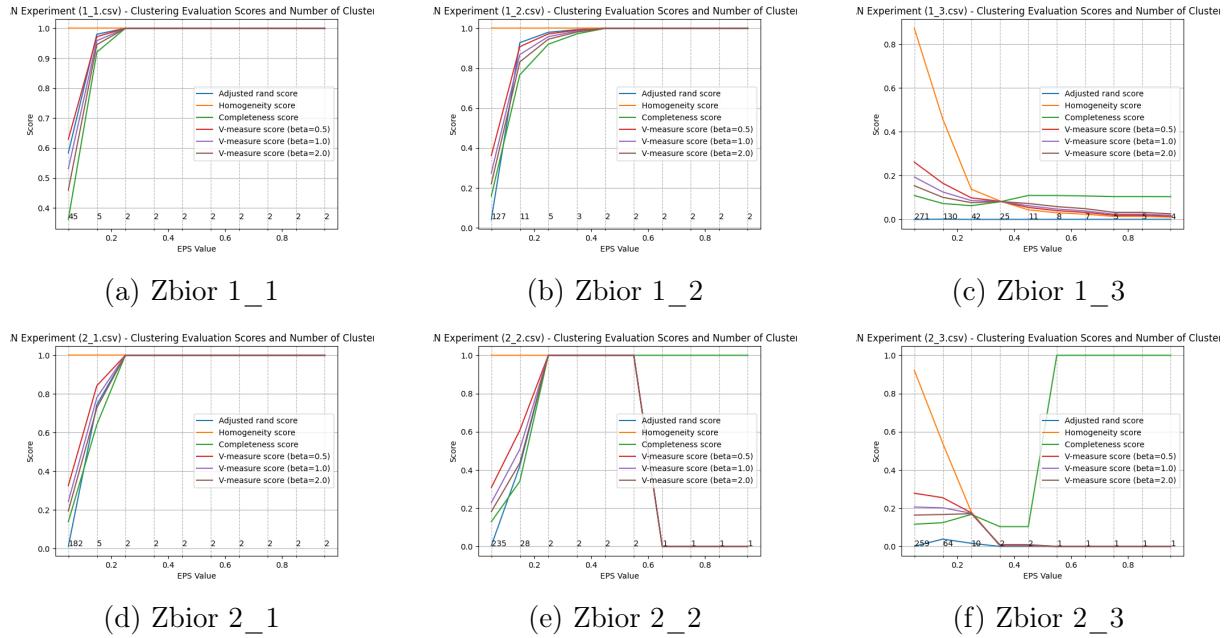


Rysunek 5: Zmiana wartości miar jakości dla wszystkich zbiorów w zależności od liczby klastrów w metodzie K-means

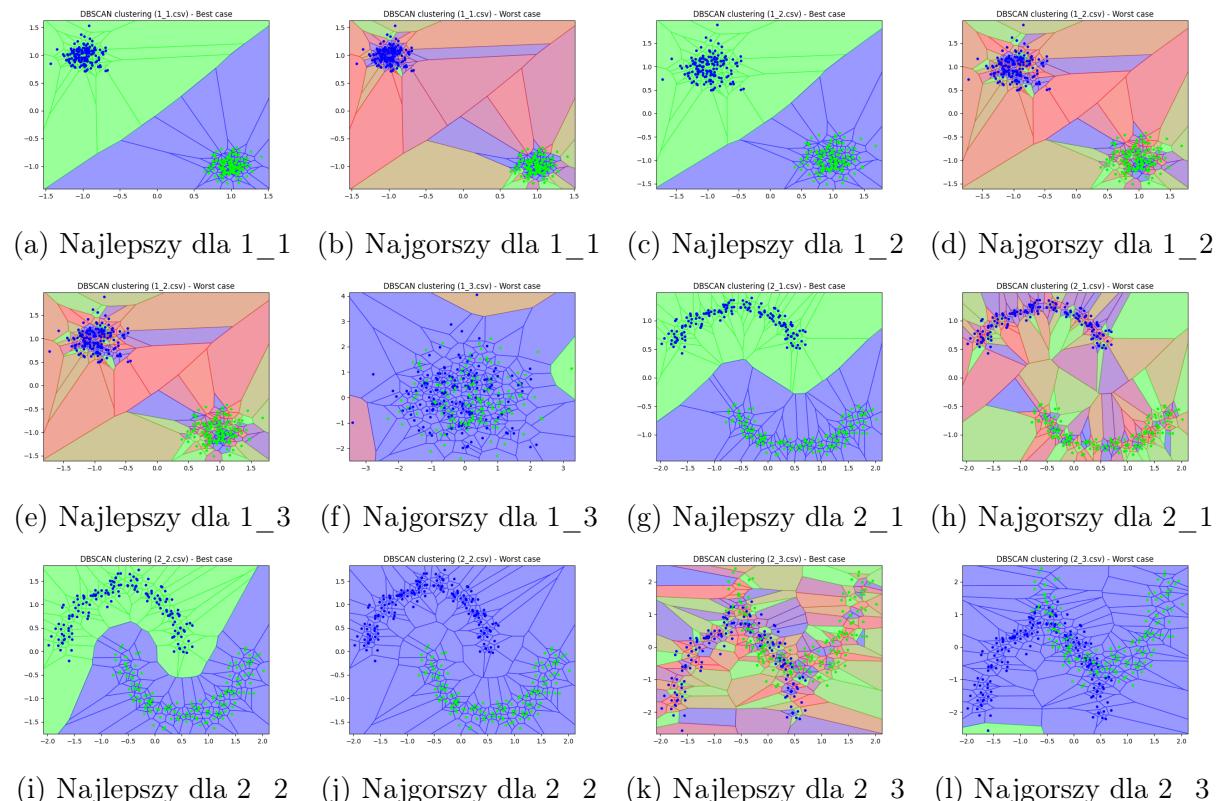


Rysunek 6: Wizualizacja klastrów wraz z prawdziwymi etykietami dla wszystkich zbiorów dla najlepszego i najgorszego przypadku w metodzie K-means

4 Eksperyment 2: Metoda DBSCAN z etykietami



Rysunek 7: Zmiana wartości miar jakości oraz liczba klastrów dla wszystkich zbiorów w zależności od wartości eps w metodzie DBSCAN



Rysunek 8: Wizualizacja klastrów wraz z prawdziwymi etykietami dla wszystkich zbiorów dla najlepszego i najgorszego przypadku w metodzie DBSCAN

5 Wnioski

Eksperyment 1

Z eksperymentów na wygenerowanych danych można wyciągnąć wnioski o częściowej użyteczności miary silhouette score przy dobieraniu parametrów klasyfikacji. Satysfakcjonujące rezultaty zaobserwowano dla wartości powyżej 0.5. Przypadki w których dla wszystkich sprawdzonych wartości parametrów, nie udało się osiągnąć wartości powyżej tej wartości, przy wizualnej inspekcji okazały się nie być łatwo separowalne z powodu przemieszania różnych klastrów.

Niestety, pewne wady miary, jaką jest silhouette score zostały zauważone w eksperymencie z metodą K-means w przypadku zbioru 2_1. Ponieważ silhouette score zależy od relacji pomiędzy odleglosciami wewnątrz klastra i dystansami do innych klastrów, najbardziej nadaje się do okrągłych skupisk, a mniej sprawdza się w przypadku innych kształtów.

Podłużny kształt widoczny na rys. 2g znacznie zwiększa odległość wewnątrz klastra, niekorzystnie wpływa na skuteczność silhouette score. Dlatego też, jak widać na rys. 1d, wartość silhouette score jest najwyższa dla 4 klastrów, choć w praktyce, co wskazano na rys. 5d oraz 6g, dla zbioru 2_1 i metody K-means, najlepszy jest podział na 2 klastry.

Metoda DBSCAN z racji bazowania na gęstości, bardziej nadaje się do podłużnych klastrów, przez co w przypadku zbioru 2_1 nie tworzy klastrów, które mogłyby dać wyższy wynik silhouette score.

Eksperyment 2

Drugi eksperyment ilustruje różne metryki porównujące przypisane etykiety do prawdziwych. Adjusted rand score ogólnie mierzy jak podobne są 2 klasteryzacje, biorąc poprawkę na losowe przypisanie do klastrów. Metryka ta sprawdza czy wybrane pary punktów należą do tego samego lub różnych klastrów, w obu klasteryzacjach. Służy ona za tem jako ogólny pogląd na podobieństwo dwóch klasteryzacji.

Homogeneity score określa jak bardzo jednolite są stworzone klastry. Mniejsza wartość oznacza, że w przypisanych klastrach znajdują się punkty z różnymi prawdziwymi etykietami. Wartość ta będzie miała generalną tendencję do wzrostu wraz ze wzrostem liczby klastrów. Maksymalna wartość 1 występuje gdy liczba klastrów jest równa ilości punktów, gdyż wtedy każdy klaster zawiera tylko jeden punkt.

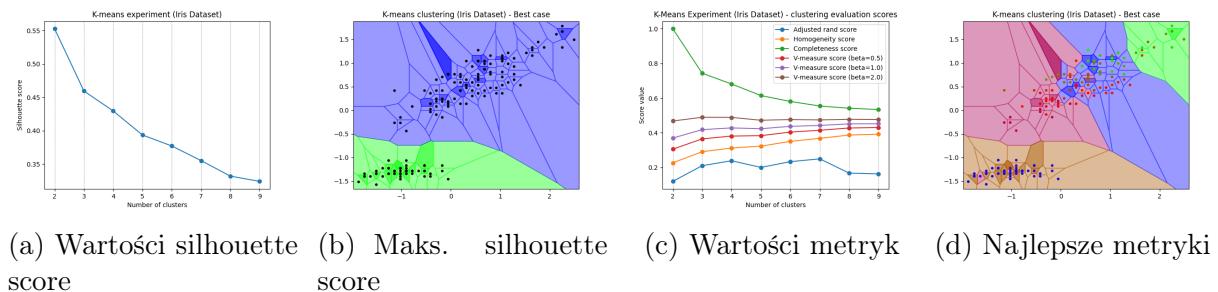
Odwrotnością homogeneity score jest completeness, określające do ilu różnych klastrów są przypisane punkty z tą samą prawdziwą etykietą. Maksymalna wartość występuje w przypadku 1 klastra, gdy dla każdej prawdziwej etykiety, wszystkie jej punkty przynależą do tego samego (jedynego klastra). Wraz ze wzrostem liczby klastrów, gdy prawdziwe grupy będą dzielone, wartość completeness score będzie miała tendencję malejącą.

V-measure score to metryka będąca średnią harmoniczną z homogeneity oraz completeness score, gdzie parametr beta kontroluje wagę tych metryk w średniej. Ponieważ wraz ze wzrostem klastrów homogeneity score będzie miał tendencję rosnącą, a completeness malejącą, może to utrudnić porównanie konkretnych przypadków. V-measure łączy homogeneity oraz completeness w jedną wartość, ułatwiającą porównywanie jakości klasteryzacji.

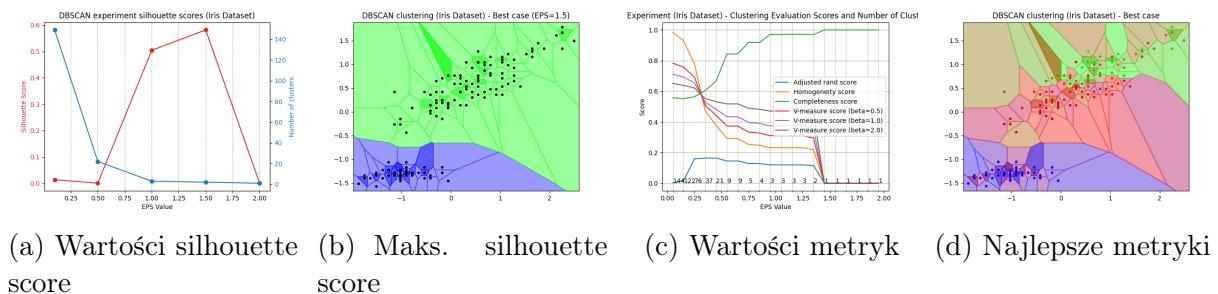
6 Analiza pozostałych zbiorów danych

Rysunki 9 i 10 przedstawiają wyniki eksperymentów z użyciem K-means oraz DBSCAN na zbiorze Iris wraz z wizualizacją dla 2 wybranych wymiarów. W przypadku wyboru parametrów w oparciu o silhouette score, obie metody doszły do identycznej klasteryzacji, która odpowiada klastrom dostrzegalnym przy wizualnej inspekcji.

Po dołączeniu metryk biorących pod uwagę prawdziwe etykiety, najlepsza klasteryzacja uległa zmianie. Obie metody poprawnie zaklasyfikowały grupę w lewym dolnym rogu diagramu Voronoi (niebieskie punkty), jednak przemieszanie dwóch pozostałych grup (zielone i czerwone punkty) sprawiło algorytmowi problem. Algorytm DBSCAN w miarę poprawnie wyodrębnił 3 klastry, jako najlepsze rozwiązanie. Zadziwiająco, algorytmowi K-means udało się uzyskać lepsze wartości metryk dla 4 klastrów, co jest większą liczbą niż ilość faktycznych etykiet.



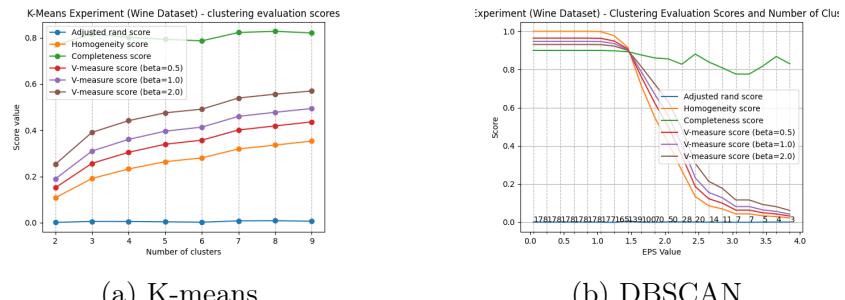
Rysunek 9: Wyniki eksperymentu K-means na zbiorze Iris



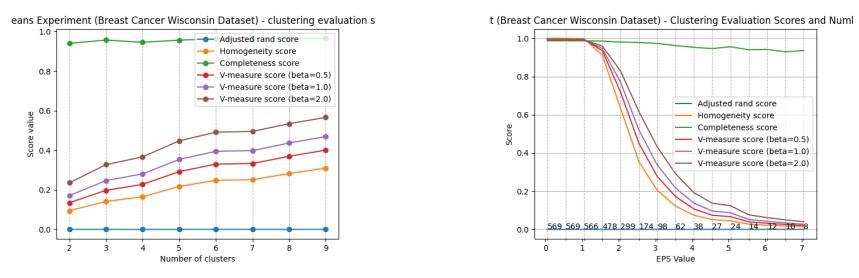
Rysunek 10: Wyniki eksperymentu DBSCAN na zbiorze Iris

W przypadku pozostałych zbiorach, z uwagi na dużą ilość wymiarów, nie udało się skonstruować satysfakcjonującej wizualizacji. Obu metodom nie udało się uzyskać satysfakcjonujących wartości silhouette score niezależnie od parametrów, dlatego owych wykresów nie przedstawiono.

Na rys. 11 i 12 znajdują się wykresy pozostałych metryk dla zbiorów Wine oraz Breast Cancer Wisconsin z użyciem metod K-means oraz DBSCAN. Niestety, w żadnym przypadku nie udało się uzyskać satysfakcjonujących wyników, co widać po bliskiej zeru wartości adjuster rand score, świadczącej o bardzo niskim podobieństwie przypisanej klasteryzacji do właściwych etykiet. Zadziwiająca jest, zwłaszcza w przypadku metody DBSCAN, bardzo wysoka wartość completeness score nawet przy dużej ilości klastrów, co może sugerować bardzo dużą ilość prawdziwych etykiet.



Rysunek 11: Wyniki eksperymentów na zbiorze Wine



Rysunek 12: Wyniki eksperymentów na zbiorze Breast Cancer Wisconsin