Praca domowa 4 - Wstęp do uczenia maszynowego

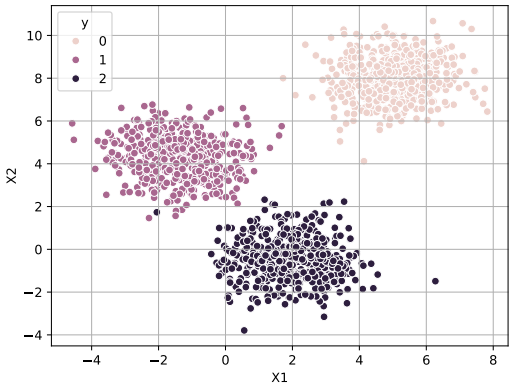
Oliwia Trzcińska

1 Wstęp

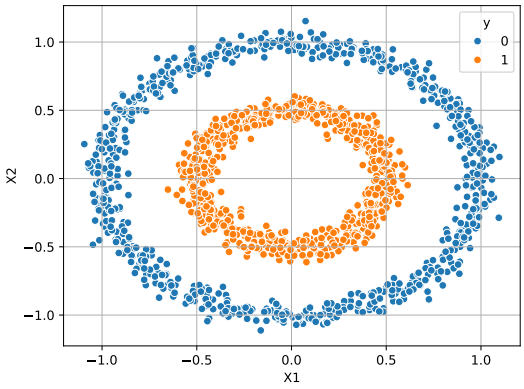
Celem pracy domowej jest odpowiedź na postawione pytania badawcze. Będziemy rozpatrywać metody analizy skupień: k-średnich oraz metodę hierarchiczną.

2 Dane

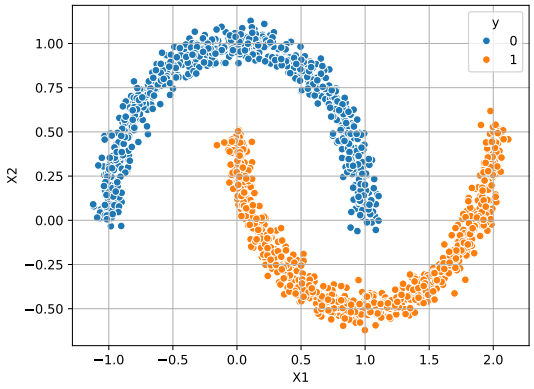
Dane, z których będziemy korzystać zostały sztucznie wygenerowane za pomocą funkcji `make_blobs`, `make_circles`, `make_moons` oraz `random.rand`. Ich graficzne reprezentacje zostały przedstawione na rysunku 1.



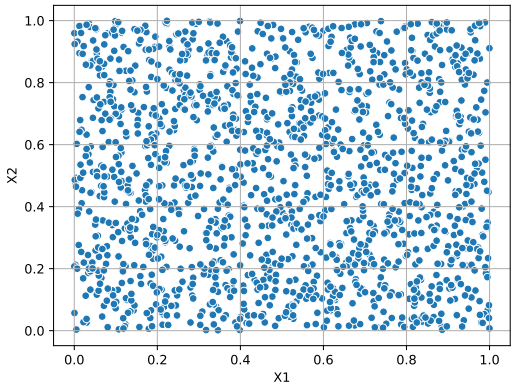
(a) Zbiór 1



(b) Zbiór 2



(c) Zbiór 3



(d) Zbiór 4

Rysunek 1: Wygenerowane zbiory danych

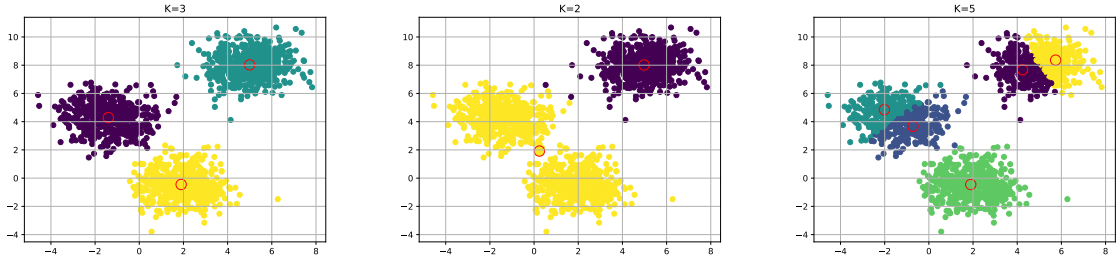
3 Pytania badawcze

3.1 Pytanie 1: Jak wybór liczby klastrów wpływa na wyniki?

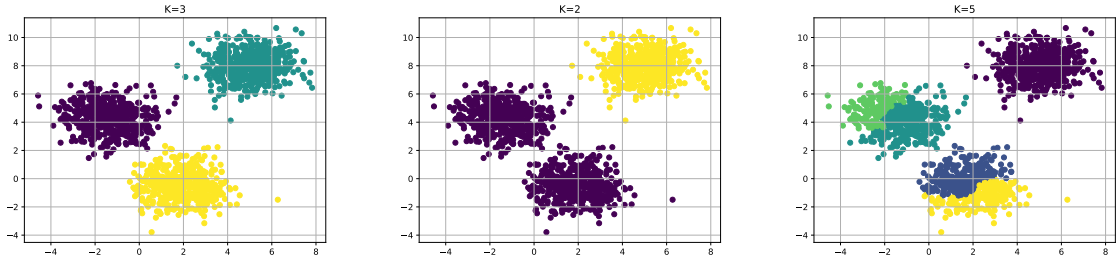
Algorytmy k-średnich oraz klasteryzacji hierarchicznej wymagają podania liczby klastrów na wejściu. Sprawdzimy, jak wybór tego parametru wpłynie na działanie metody.

3.1.1 Dane 1

Naturalnym wyborem dla tych danych jest podział na 3 klastry. Wtedy grupy są dobrze zdefiniowane, a analiza skupień jest bardziej trafna. Dane w każdym klastrze są do siebie zbliżone, a różnice między klastrami są znaczące. Wybór mniejszej liczby klastrów (2) nie uwzględnia wszystkich subtelności w danych, a wybór większej liczby (5) powoduje powstanie nieznaczących grup.



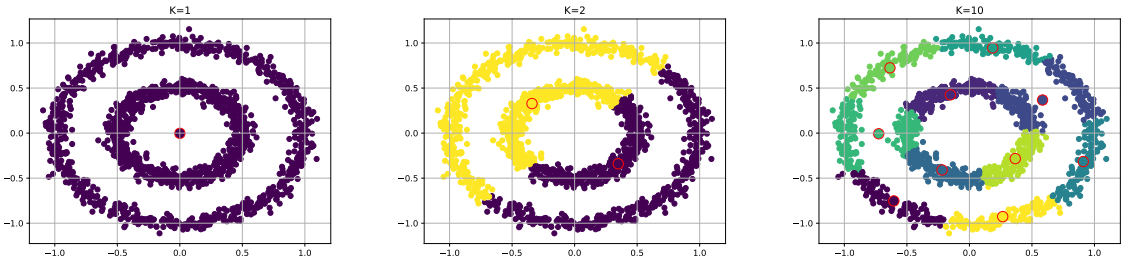
Rysunek 2: Przydział obserwacji do skupień dla różnych wartości k w metodzie k-średnich



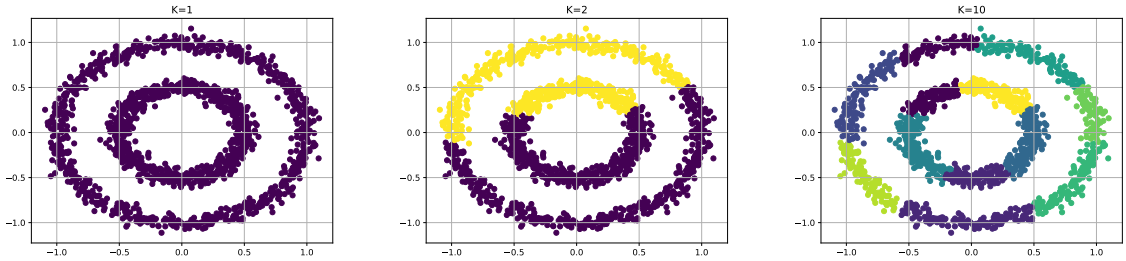
Rysunek 3: Przydział obserwacji do skupień dla różnych wartości k w metodzie hierarchicznej

3.1.2 Dane 2

Liczba klastrów równa 1 jest bezużyteczna, ponieważ nie mamy żadnego podziału obserwacji. Dla podziału na 2 klastry mamy lepsze przyporządkowanie tych danych do skupień. Wybór zbyt dużej liczby (10) ponownie prowadzi do powstania wielu grup nieznacznie różniących się między sobą.



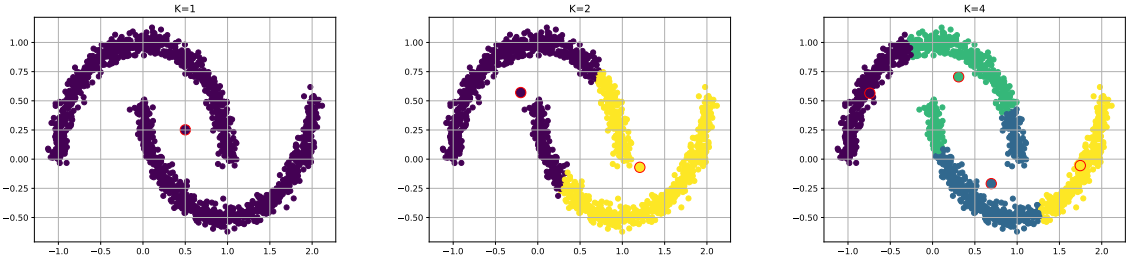
Rysunek 4: Przydział obserwacji do skupień dla różnych wartości k w metodzie k-średnich



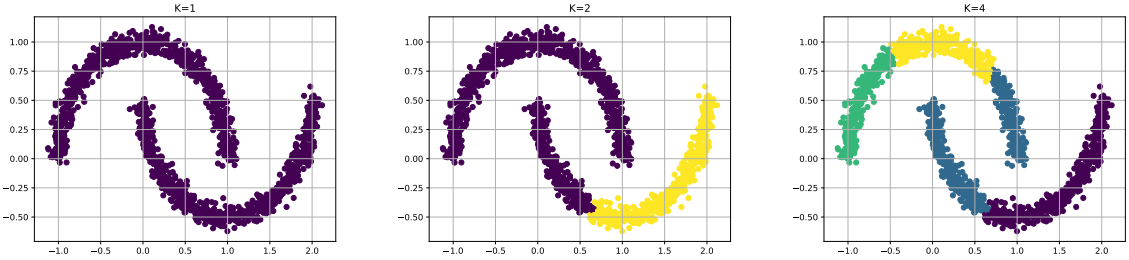
Rysunek 5: Przydział obserwacji do skupień dla różnych wartości k w metodzie hierarchicznej

3.1.3 Dane 3

W przypadku danych 3 obserwacje są analogiczne jak wyżej.



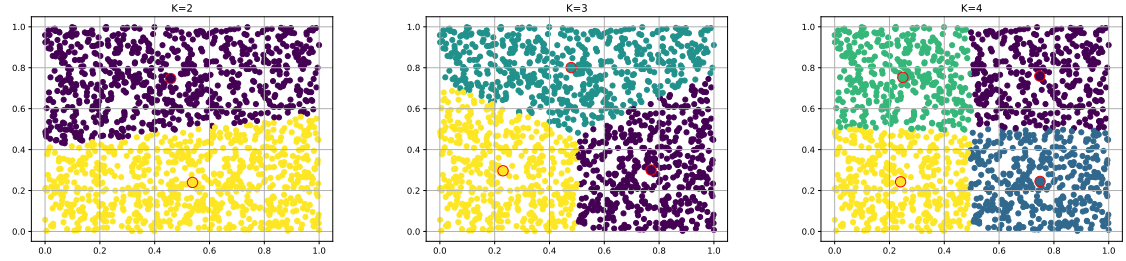
Rysunek 6: Przydział obserwacji do skupień dla różnych wartości k w metodzie k-średnich



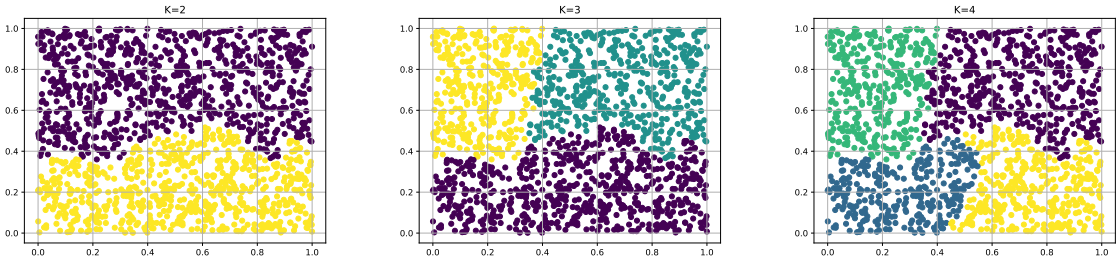
Rysunek 7: Przydział obserwacji do skupień dla różnych wartości k w metodzie hierarchicznej

3.1.4 Dane 4

W przypadku losowych danych trudno określić optymalną liczbę klastrów.



Rysunek 8: Przydział obserwacji do skupień dla różnych wartości k w metodzie k-średnich



Rysunek 9: Przydział obserwacji do skupień dla różnych wartości k w metodzie hierarchicznej

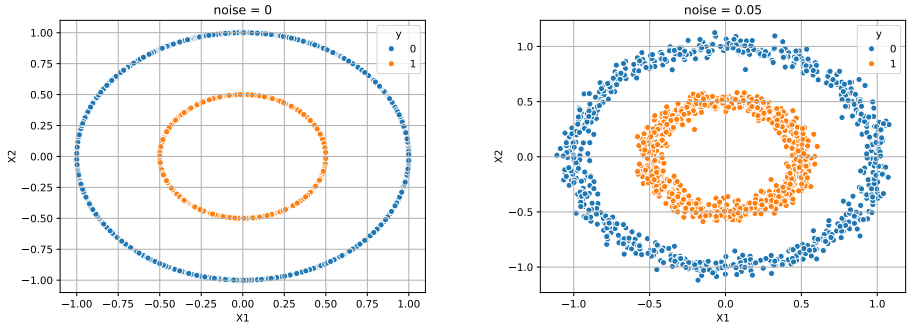
3.1.5 Wnioski

Wybór liczby klastrów ma istotny wpływ na wyniki analizy skupień. Decyzja ta prowadzi do różnych interpretacji danych. Jeśli wybierzemy zbyt małą liczbę klastrów, algorytm może nie uwzględnić wszystkich subtelności w danych. Grupy mogą być zbyt ogólne, a analiza skupień może mieć mniejsze znaczenie. Wyniki będą mniej precyzyjne. Jeśli wybierzemy z kolei zbyt dużą liczbę klastrów, to może powstać wiele małych, nieznaczących grup mniej różniących się od siebie, co utrudni interpretację wyników. Wybór optymalnej liczby klastrów może być trudny, ale prowadzi do najlepszych wyników. Wtedy dane w każdym klastrze są do siebie zbliżone, a różnice między klastrami są znaczące.

3.2 Pytanie 2: Jaki jest wpływ parametru szumu na zestawy danych 2 i 3?

3.2.1 Dane 2

Na rysunku 10 zostały przedstawione graficzne reprezentacje zbioru danych 2 dla różnych wartości parametru szumu.



(a) noise = 0

(b) noise = 0.05

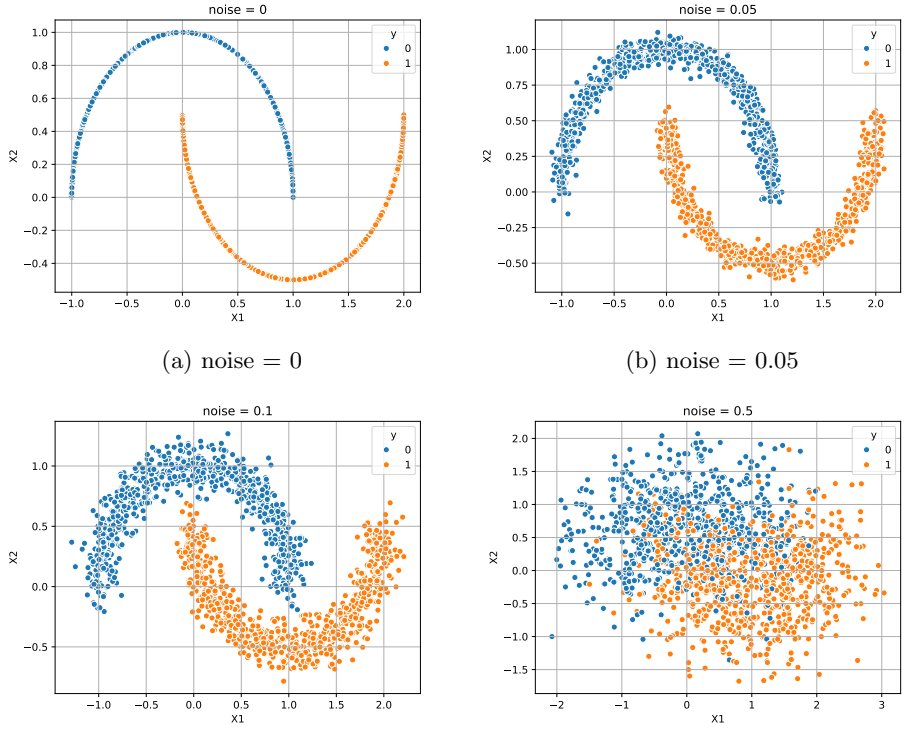
(c) noise = 0.1

(d) noise = 0.5

Rysunek 10: Reprezentacja graficzna zbioru 2 dla różnych wartości parametru noise

3.2.2 Dane 3

Na rysunku 11 zostały przedstawione graficzne reprezentacje zbioru danych 3 dla różnych wartości parametru szumu.



(a) noise = 0

(b) noise = 0.05

(c) noise = 0.1

(d) noise = 0.5

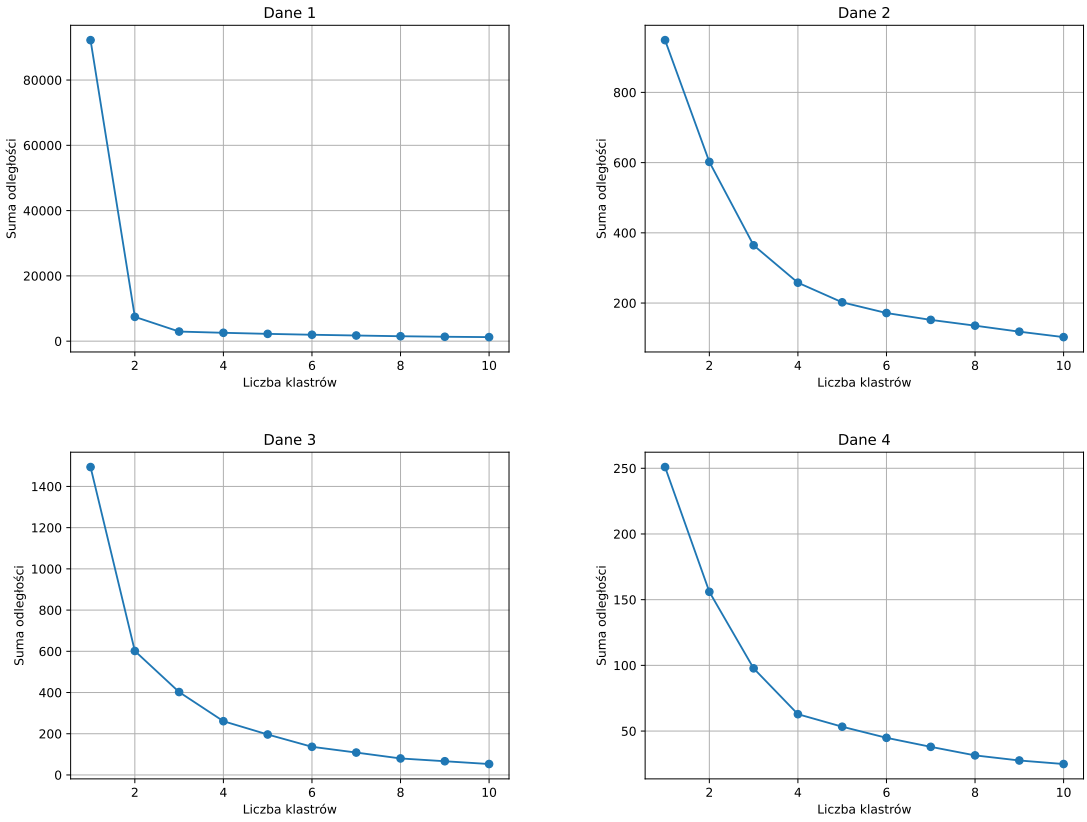
Rysunek 11: Reprezentacja graficzna zbioru 3 dla różnych wartości parametru noise

3.2.3 Wnioski

Parametr `noise` przyjmowany przez funkcje `make_circles` oraz `make_moons` ma duży wpływ na strukturę generowanych danych. Wpływa to na trudność zadania klasyfikacji. Wartość parametru szumu równa 0 powoduje wygenerowanie danych bez szumu, czyli idealnie układających się w kształt odpowiednio okręgu i księżyca. Można zaobserwować, że im większa wartość parametru `noise`, tym bardziej dane są rozproszone. Punkty zostają losowo przesunięte, co sprawia, że dane są mniej jednoznaczne. Już dla wartości 0.5 trudno zauważyć jakieś kształty w danych.

3.3 Pytanie 3: Jak całkowita suma odległości między punktami w klastrach zależy od liczby klastrów?

Całkowita suma odległości między punktami w klastrach jest miarą jakości klasteryzacji w metodzie k-średnich. Im mniejsza jest ta suma, tym lepsza klasteryzacja. Na podstawie rysunku 12 możemy zauważyć, że całkowita suma odległości między punktami w klastrach jest odwrotnie proporcjonalna do liczby klastrów - gdy zwiększamy liczbę kwadratów, ta suma się zmniejsza. Istnieje jednak pewien punkt, po którym dalsze zwiększanie liczby klastrów już tylko nieznacznie zmniejsza sumę odległości, a przy tym prowadzi do powstania bardziej skomplikowanego modelu i w praktyce wcale nie lepszej klasteryzacji. Optymalna liczba klastrów jest zazwyczaj tam, gdzie dodatkowe klastry są coraz mniej użyteczne w redukcji sumy odległości.



Rysunek 12: Zależności sumy odległości między punktami w klastrach od liczby klastrów

4 Podsumowanie

Metoda k-średnich oraz metoda klasteryzacji hierarchicznej to przykłady algorytmów stosowanych do grupowania podobnych obserwacji, co ułatwia rozumienie struktury danych. Kluczową decyzją przy stosowaniu obu tych metod jest wybór optymalnej liczby klastrów.