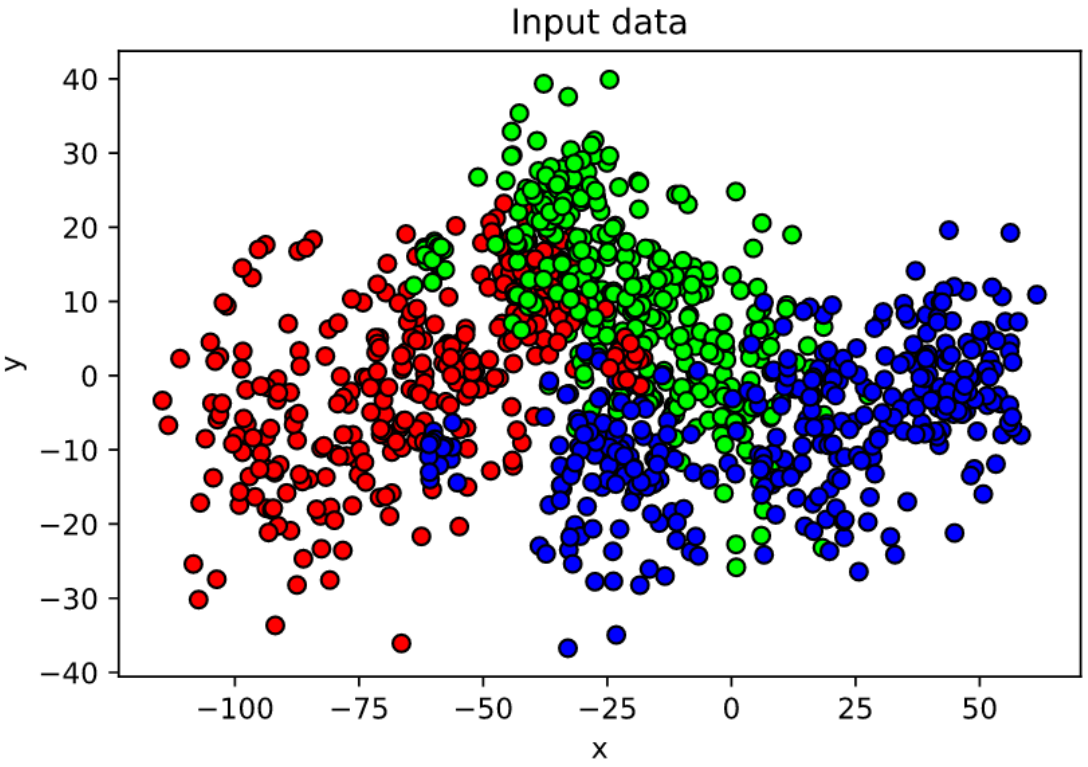
Machine Learning

lab 2

03.04.2021 Mikołaj Zatorski

Zad 1.

Z użyciem funkcji *sklearn.datasets.make_blobs* wygenerowałem zbiór danych spełniający warunki z treści zadania:



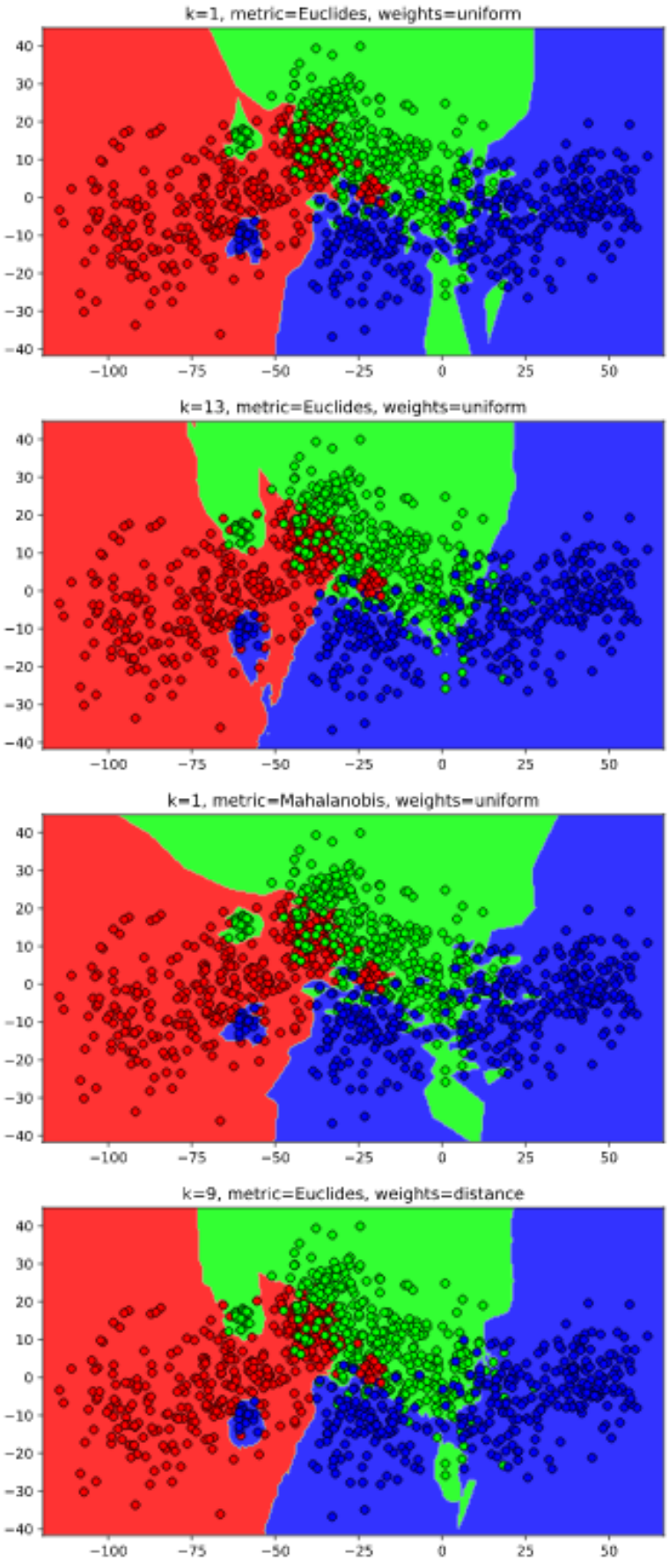
Wykres nr 1. Zbiór danych wejściowych.

Komentarz:

Jak widać na *wykresie 1* zbiór został podzielony na trzy klasy, oznaczone odpowiednio kolorem czerwonym, zielonym oraz niebieskim.

Zad 2.

Następnie przetestowałem różne warianty klasyfikatora k-NN i sprawdziłem jak kształtują się granice decyzyjne dla każdego z nich.

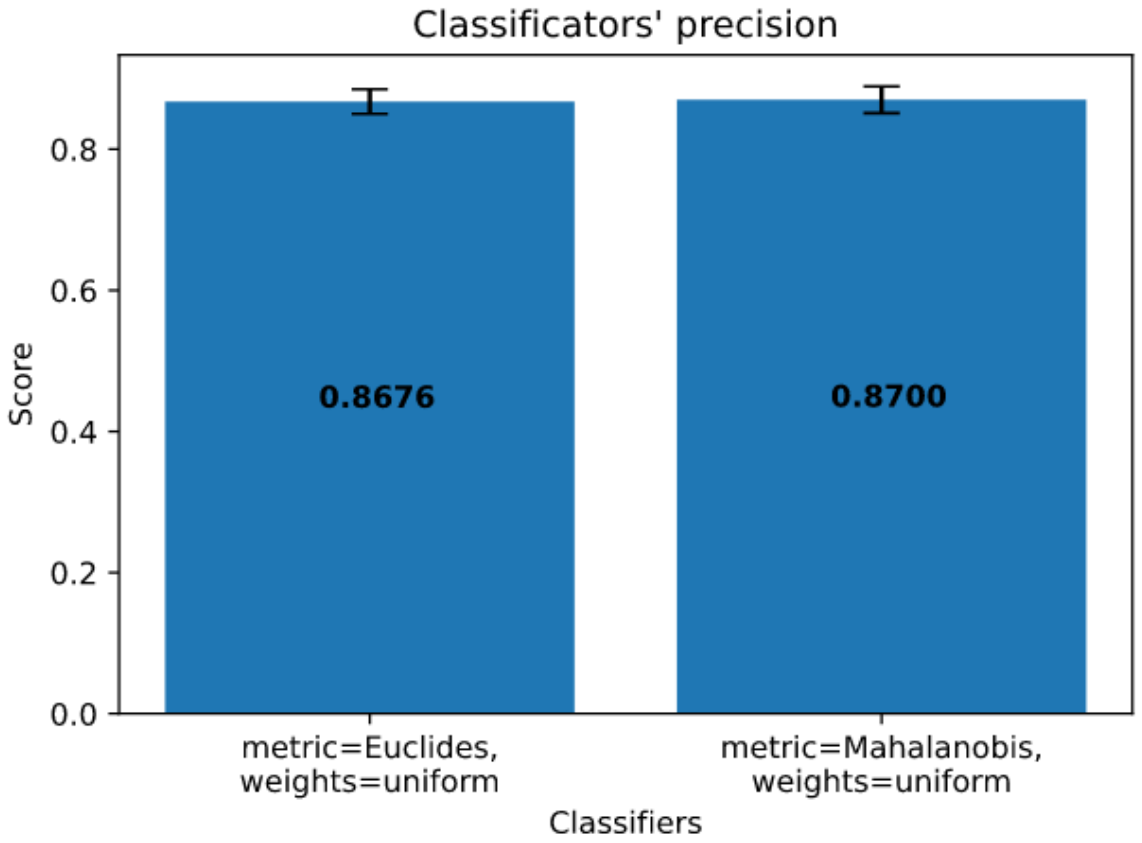


Wykres nr 2. Granice decyzyjne w zależności od parametrów klasyfikatora k-NN.

Komentarz:
Granice dla klasyfikatorów z *k=1* były dużo bardziej “czułe”, natomiast te o wyższym *k* odrzucały skrajne przypadki. Subiektywnie wybrałem metrykę Euklidesa wraz z głosowaniem większościowym jako najskuteczniejszą parę (metryka + rodzaj głosowania), a metrykę Mahalanobisa wraz z głosowaniem większościowym jako najgorszą.

Zad 3.

W tym zadaniu sprawdziłem rzeczywistą skuteczność wybranych klasyfikatorów z zadania drugiego. Wszystkie pomiary dla każdego klasyfikatora powtórzyłem **10** razy, natomiast dla pojedynczego *k* w obrębie jednego testu klasyfikatora, pomiar powtórzyłem **20** razy. Wynik porównania skuteczności klasyfikatora subiektywnie najlepszego z najgorszym prezentuje się następująco:



Wykres nr 3. Porównanie średniej skuteczności klasyfikatorów.

Komentarz:
Jak widać na *wykresie 3* klasyfikator z metryką Mahalanobisa i głosowaniem większościowym (początkowy uznany za gorszy) uzyskał minimalnie lepszą średnią skuteczność niż klasyfikatora z metryką Euklidesa. Błąd pierwotnego założenia prawdopodobnie wynika z faktu, że w zadaniu drugim *k* było równe 1, a w tym podpunkcie szukaliśmy **najlepszego k** dla danego klasyfikatora. Warto jednak zauważyć, że wynik obu klasyfikatorów różnią się minimalnie oraz mieszczą się w granicach odchyleń standardowych. Odchylenie standardowe na poziomie około 0.02 w obu przypadkach sugeruje nam, że klasyfikatory z tymi parametrami, są stabilne dla tego zbioru.

Podsumowanie

K-NN jest prostą metodą klasyfikacji danych. Potrafi być ona dość skuteczna (ok. 85% dla niełatwego zbioru danych), jednak wymaga sporego nakładu pracy w celu dostrojenia hiperparametrów (dla dużych danych może być problem z wydajnością). Działanie i w konsekwencji skuteczność tej metody można zmieniać za pomocą rodzaju metryk oraz głosowania. W naszym przypadku te wahania nie były duże, za pomocą modyfikacji metryki nie udało się znacząco zmienić skuteczności.