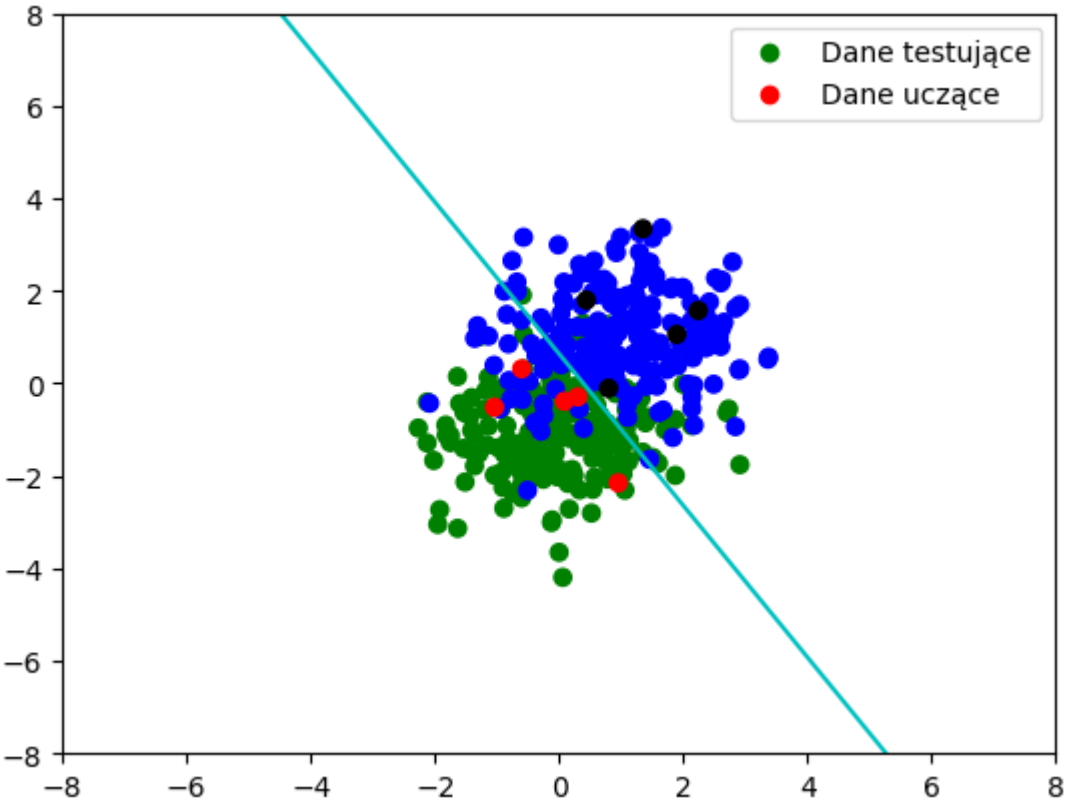
Metody Inteligencji Obliczeniowej
Laboratorium 2

Klasyfikacja z użyciem sztucznych neuronów

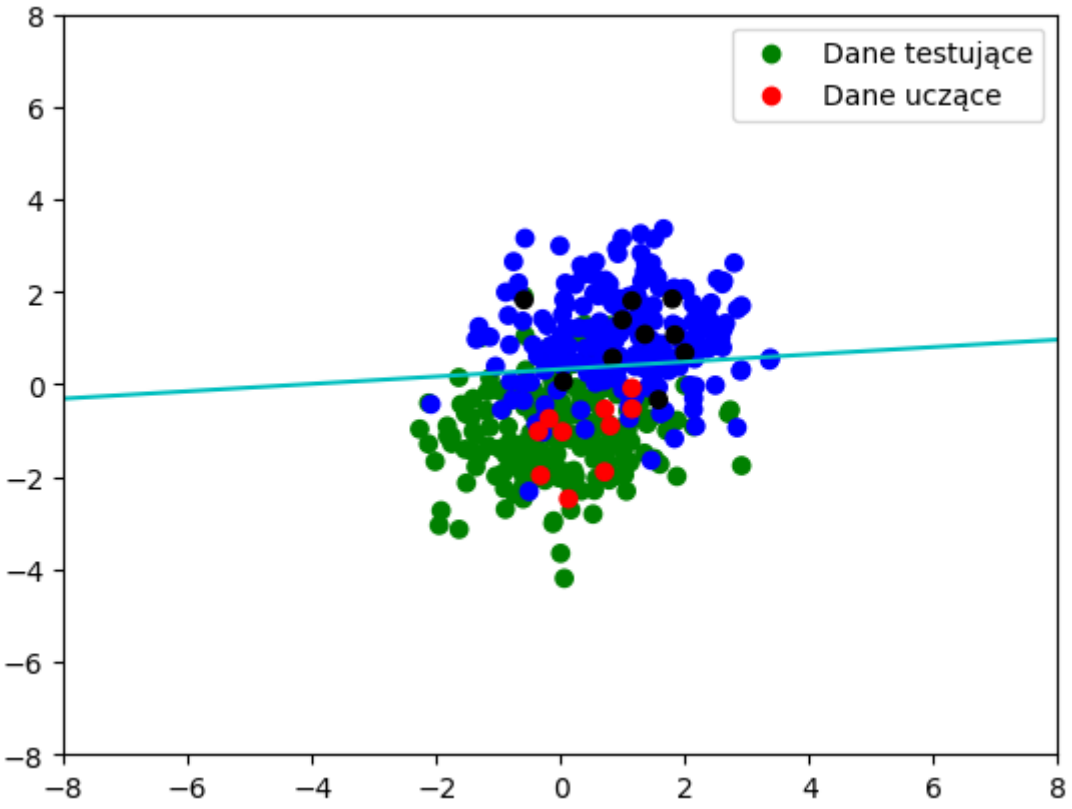
Kamil Pyla

Celem pierwszego zadania było zbadanie wpływu ilości danych uczących na poprawne zaklasyfikowanie danych testowych. Dane klasyfikowane do odpowiednich klas pochodziły z rozkładów normalnych o różnej wartości oczekiwanej.

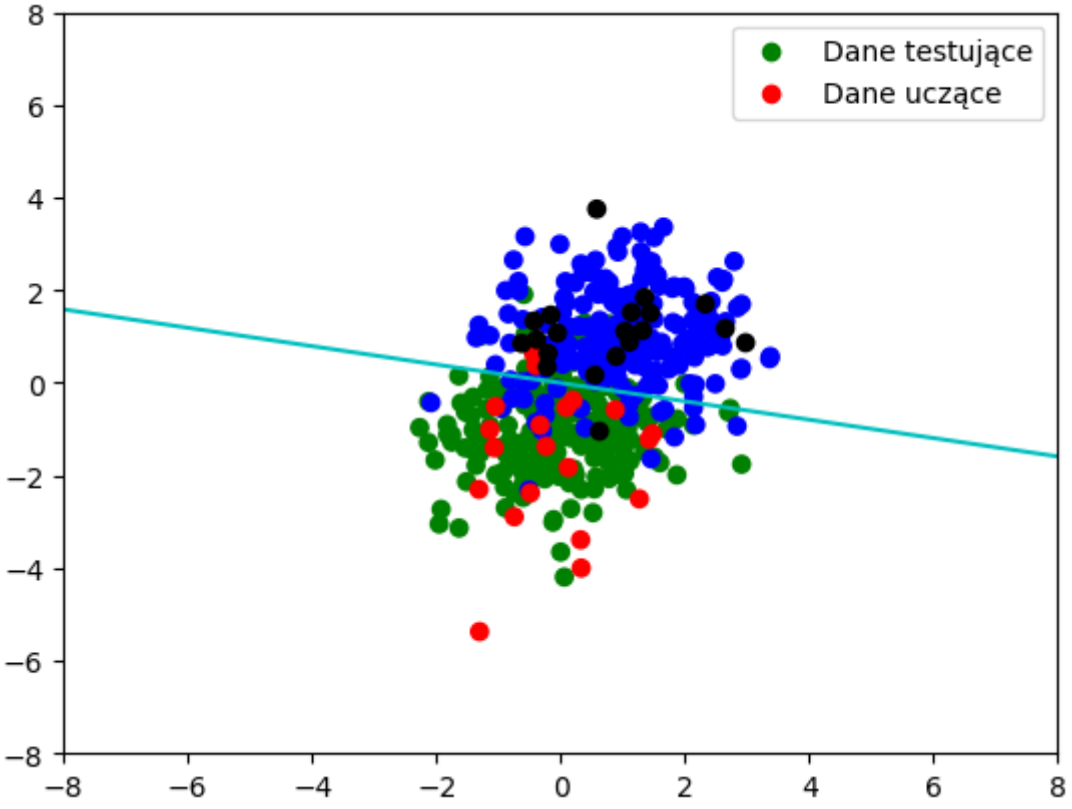
Wyniki:



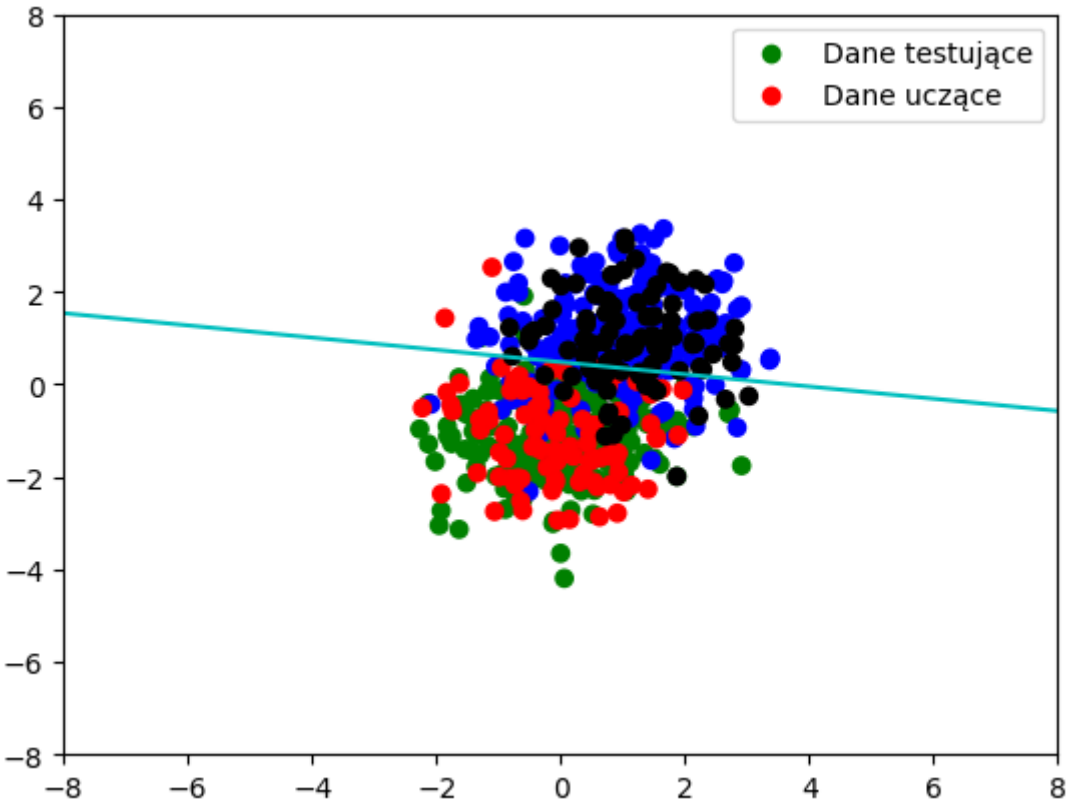
Rysunek 1. Wyniki dla 5 punktów uczących



Rysunek 2. Wyniki dla 10 punktów uczących



Rysunek 3. Wyniki dla 20 punktów uczących

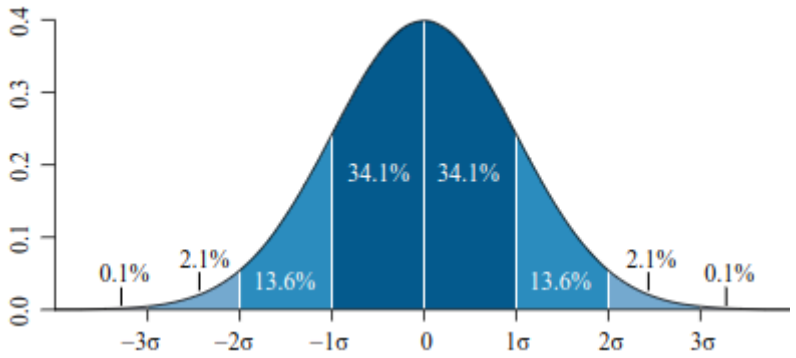


Rysunek 4. Wyniki dla 100 punktów uczących

Ilość punktów uczących	Poprawność dla danych uczących	Poprawność dla danych testowych
5	100%	82%
10	90%	83,5%
20	92,5%	87%
100	86%	85%

Wnioski: Korzystając z danych pochodzących z rozkładu normalnego, nie jest możliwym osiągnięcie dokładności klasyfikacji na poziomie 100% z powodu rozrzucenia danych. Osiągnięcie 100% poprawności klasyfikacji danych uczących w pierwszym wypadku wynika z małej ilości punktów - im mniej punktów tym większe prawdopodobieństwo możliwości poprowadzenia prostej pomiędzy nimi. Poprawność danych testowych na poziomie około 80% można uznać za bardzo dobry wynik, ponieważ z własności rozkładu normalnego prawdopodobieństwo, że wylosowane dane będą w odległości odchylenia standardowego od wartości oczekiwanej wynosi około 80%. Można zatem

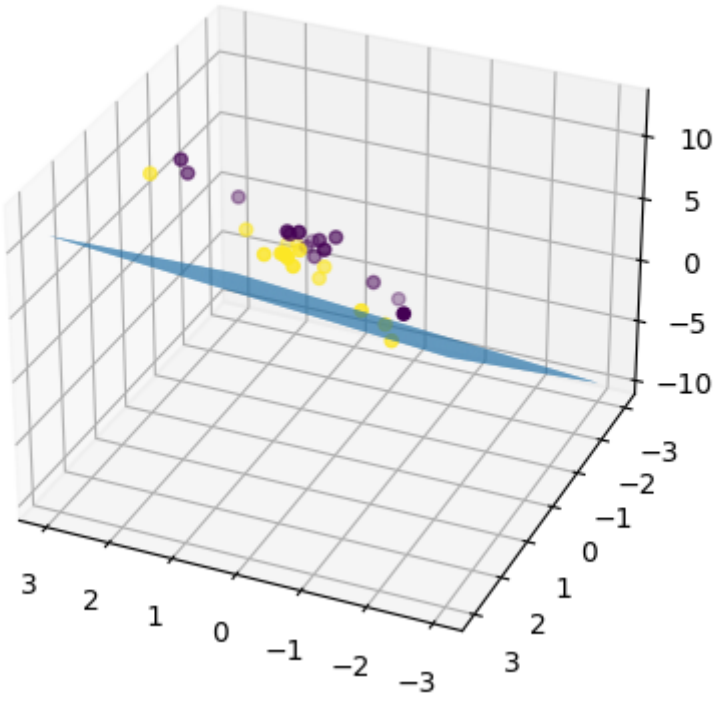
wyciągnąć wnioski, że im więcej danych uczących, tym ich dokładność będzie oscylowała w okolicy 80% i wynika to z prawdopodobieństwa rozrzucenia danych o rozkładzie normalnym.



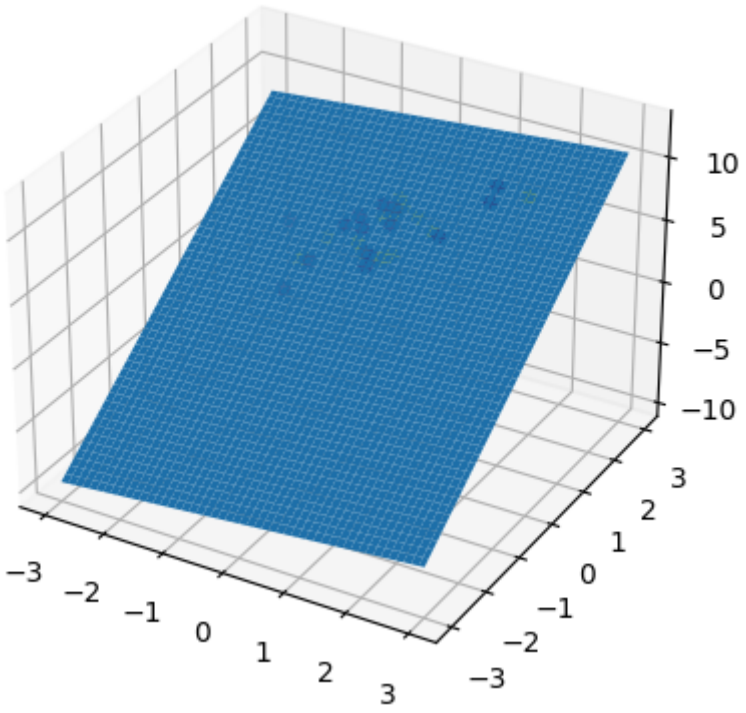
Rysunek 5. Wykres rozkładu normalnego

W drugim zadaniu należało porównać wyniki klasyfikacji w pięciu iteracjach dokonanych przed perceptron, na podstawie sklasyfikowanych danych klas benzyny

Numer iteracji	Dokładność	Macierz pomyłek
1	53.(3)%	[[15 0] [14 1]]
2	53.(3)%	[[15 0] [14 1]]
3	53.(3)%	[[15 0] [14 1]]
4	53.(3)%	[[15 0] [14 1]]
5	53.(3)%	[[15 0] [14 1]]



Rysunek 6. Wykres klasyfikacji punktów razem z hiperpłaszczyzną



Rysunek 7. Wykres z innej perspektywy

Wnioski: Przy tak dobranych danych, wpływ ilości iteracji na wynik nie ma żadnego wpływu, wynika to z jednoznacznego dopasowania hiperpłaszczyzny

względem danych uczących. Dopasowanie na poziomie 50% jest dopasowaniem bardzo słabym mając do dyspozycji jedynie dwie klasy. Klasyfikując wszystkie dane do jednej klasy osiągamy 50%. Jednak analizując wykres hiperpłaszczyzny można zauważyć, że kąt nachylenia płaszczyzny jest zbliżony do kąta nachylenia punktów klasyfikacji.

W trzecim zadaniu należało podzielić dane pochodzące ze zbioru irysów na dane uczące oraz testujące w stosunku 80% do 20% w celu szerszego spojrzenia na problem dokonałem 5 - krotnej randomizacji podziału danych uczących i testowych. Dla każdego podziału uzyskałem dokładnie te same wyniki dla każdej iteracji.

Numer randomizacji	Dokładność	Macierz pomyłek
1	86,(6)%	[[10 0 0] [0 10 0] [0 4 6]]
2	93,(3)%	[[10 0 0] [0 8 2] [0 0 10]]
3	86,(6)%	[[10 0 0] [4 6 0] [0 0 10]]
4	70%	[[10 0 0] [0 10 0] [0 9 1]]
5	70%	[[10 0 0] [0 1 9] [0 0 10]]

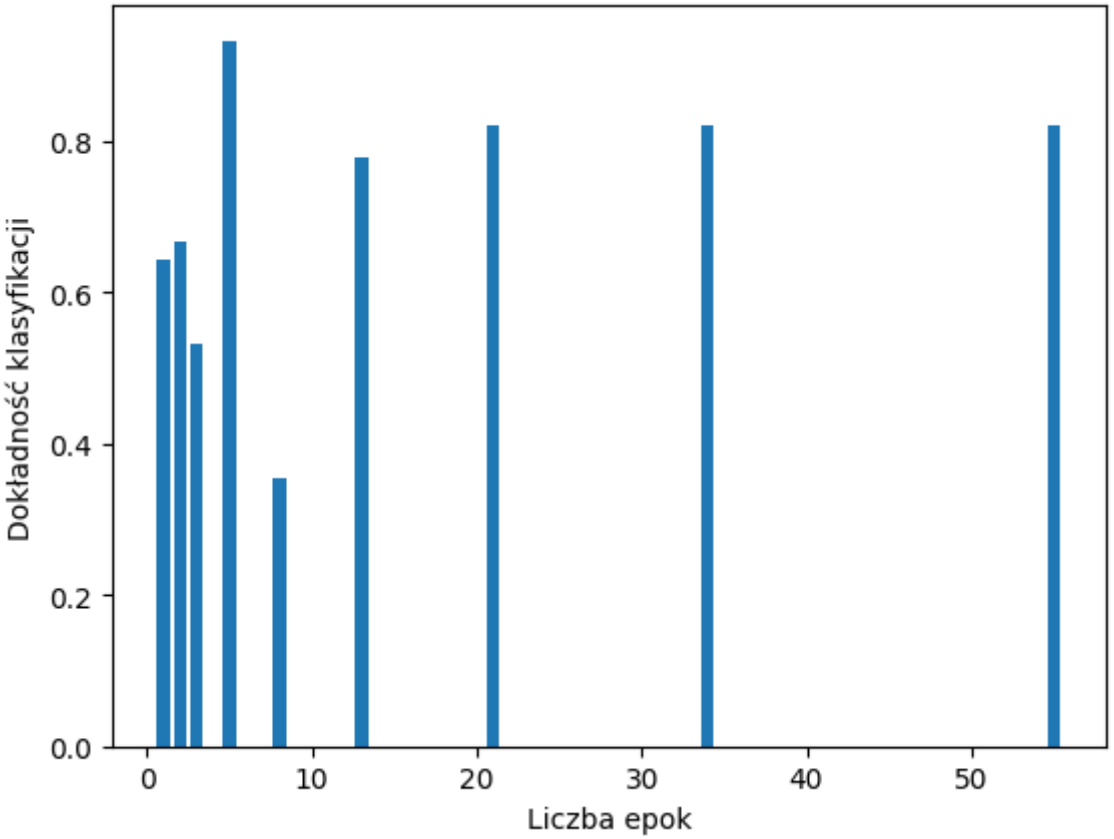
W kolejnym ćwiczeniu należało zbadać wpływ stosunku podziału na poprawność klasyfikacji. Wykonałem dwa warianty tego ćwiczenia, w pierwszym zmieniałem stosunek podziału, bez randomizacji danych, w drugim, przed kolejnymi podziałami dokonywałem randomizacji danych.
Wyniki:

% Danych uczących	Dokładność klasyfikacji	Macierz pomyłek
20%	66.(6)%	[[40 0 0] [28 0 12] [0 0 40]]
30%	66.(6)%	[[35 0 0] [1 0 34] [0 0 35]]
40%	87.(7)%	[[30 0 0] [0 19 11] [0 0 30]]
50%	66.(6)%	[[25 0 0] [21 0 4] [0 0 25]]
60%	66.(6)%	[[20 0 0] [17 0 3] [0 0 20]]
70%	66.(6)%	[[15 0 0] [10 0 5] [0 0 15]]
80%	66.(6)%	[[10 0 0] [10 0 0] [0 0 10]]
90%	66.(6)%	[[5 0 0] [5 0 0] [0 0 5]]

% Danych uczących	Dokładność klasyfikacji	Macierz pomyłek
20%	66.(6)%	[[40 0 0] [7 0 33] [0 0 40]]
30%	67.62%	[[35 0 0] [0 35 0] [0 34 1]]
40%	76.(6)%	[[30 0 0] [2 9 19] [0 0 30]]
50%	66.(6)%	[[25 0 0] [20 0 5] [0 0 25]]
60%	66.(6)%	[[20 0 0] [3 0 17] [0 0 20]]
70%	66.(6)%	[[15 0 0] [0 0 15] [0 0 15]]
80%	66.(6)%	[[10 0 0] [0 0 10] [0 0 10]]
90%	66.(6)%	[[5 0 0] [5 0 0] [0 0 5]]

Wnioski: Ilość danych użytych do uczenia ma mniejszy wpływ na wynik poprawności klasyfikacji, niż odpowiedni ich dobór. W trakcie wykonywania doświadczenia, zdarzyła się sytuacja, gdy przy użyciu 10% danych do uczenia perceptron osiągnął dokładność na poziomie 90%, oraz taka sama dokładność przy zastosowaniu 70% danych do uczenia. Jednocześnie nie jest możliwe osiągnięcie klasyfikacji na poziomie 100%, ze względu na liniową nie separowalność danych.

Ostatnim ćwiczeniem było zbadanie ilości epok na dokładność klasyfikacji.



Rysunek 8. wykres opisywanych zależności

Z powyższego wykresu można zauważyć, że największe różnice dokładności występują dla niewielkiej ilości epok. Większa ilość epok wpływa na dokładność, do pewnego punktu optymalnego. Na początku różnice są duże, później stabilizują się.

Link do repozytorium:
https://github.com/KamilPyla/MIO_2023/tree/master/lab_01