ZACHODNIOPOMORSKI UNIWERSYTET TECHNOLOGICZNY
W SZCZECINIE
INSTYTUT TECHNOLOGII MECHANICZNEJ

Metody Sztucznej Inteligencji

Sztuczne Sieci Neuronowe- Statistica

Wstęp

Zainteresowanie sieciami neuronowymi (SSN lub ANNs – artificial neural networks) - systematycznie rośnie. Są one z powodzeniem stosowane w bardzo wielu, bardzo różnych dziedzinach jak finanse, medycyna, technika, geologia czy fizyka. Sieci neuronowe mogą być zastosowane wszędzie tam, gdzie pojawiają się zadania związane z predykcją, klasyfikacją czy sterowaniem. Bardzo ważne znaczenie w przypadku sieci neuronowych ma przygotowanie zmiennych, które muszą charakteryzować się następującymi cechami: mają wpływ na modelowane zjawisko; zmienne liczbowe i nominalne używamy bezpośrednio, inne zmienne przekształcamy do jednej z tych postaci lub rezygnujemy z nich, liczba niezbędnych przypadków może być rzędu setek lub tysięcy; im więcej zmiennych tym więcej trzeba przypadków, jeżeli jest to konieczne, to można użyć niekompletnych przypadków (z "brakami danych" dla niektórych zmiennych), jednak wartości nietypowe mogą być przyczyną problemów, dlatego należy je raczej usunąć, jeżeli mamy dużo przypadków, to należy także odrzucić przypadki z brakami w danych, jeżeli mamy mało danych, spróbujmy użyć zespołów sieci i próbkowania.
Istnieje wiele typów i rodzajów sieci neuronowych, różniących się między sobą strukturą i zasadami działania, ale chyba najpopularniejsza obecnie architektura sieciowa związana jest z koncepcją wielowarstwowego perceptronu (MLP). Liczba neuronów wejściowych i wyjściowych jest zdeterminowana przez rozwiązywany problem. Wśród typów sieci mających zastosowanie w Statistice wyróżnić można:

- Perceptron wielowarstwowy (MLP - Multilayer perceptron)
- Sieć o radialnych funkcjach bazowych (RBF - Radial Basis Function)
- Probabilistyczne sieci neuronowych (PNN - probabilistic neural networks) -
- Sieci neuronowych realizujących regresję uogólnioną (GRNN - generalized regression neural networks) – zastosowanie w zagadnieniach regresyjnych

Sieci neuronowe mogą być stosowane w praktycznie każdej sytuacji, gdzie pomiędzy zmiennymi zależnymi i niezależnymi istnieje rzeczywista zależność lub zespół zależności, nawet jeśli są one bardzo skomplikowane i niewyrażalne w klasyczny sposób, poprzez korelacje czy różnice pomiędzy grupami przypadków. Mogą być wykorzystywane do: rozpoznawania jednostek chorobowych, prognozowania finansowych szeregów czasowych, oceny wiarygodności kredytowej, monitorowania stanu maszyny czy też sterowania pracą silnika. Wśród najczęściej rozwiązywanych zadań za pomocą SSN wyróżniamy:

- Zadania klasyfikacyjne – sieć ma za zadnie przydzielić rozpatrywany przypadek do jednej ze zdefiniowanych wcześniej klas., np. rozpatrywanie wniosków kredytowych (udzielić kredytu danej osobie, czy też nie), wykrywanie nowotworów (wykrycie lub wykluczenie nowotworu), rozpoznawanie podpisów (fałszywy bądź autentyczny) itp. W każdym z powyższych przypadków na wyjściu wymagana jest pojedyncza zmienna nominalna.
- Zadania regresyjne – celem jest prognozowanie wartości (zwykle ciągłej) określonej zmiennej: na przykład jutrzejszej ceny akcji, zużycia paliwa przez samochód, wysokości przyszłorocznych zysków itp. W tym przypadku na wyjściu sieci wymagana jest pojedyncza zmienna numeryczna.

Przykład 1 – zadanie klasyfikujące

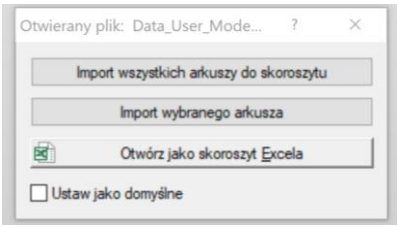
Rozpatrzymy zadanie klasyfikujące poziom wiedzy studenta w zależności od stopnia przygotowania, poświęconego czasu na naukę oraz wyniku egzaminu. Zbiór danych obejmuje poniższe wielkości:

	A	B	C	D	E	F	G	H
1	SCN	SIP	SCNP	SEPP	SEPG	PWS		SCN - stopień czasu nauki materiału gównego
2	0	0,1	0,5	0,26	0,05	Very Low		SIP - stopień ilości powtórek materiału gównego
3	0,05	0,05	0,55	0,6	0,14	Low		SCNI - stopień czasu nauki zagadnień powiązanych
4	0,08	0,18	0,63	0,6	0,85	High		SEPP - skuteczność studenta na egzaminie w zagadnieniach powiązanych
5	0,2	0,2	0,68	0,67	0,85	High		SEPG - skuteczność studeta na egzamianie z materiału głównego
6	0,22	0,22	0,9	0,3	0,9	High		PWS - poziom wiedzy studenta

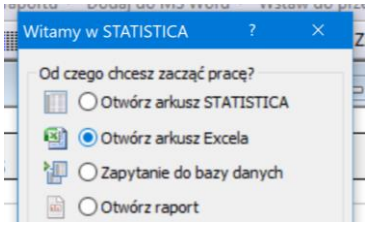
Liczy 1449 rozpatrywanych przypadków, rozkład danych w zbiorze wygląda następująco: 13% studentów wykazywało bardzo niski poziom wiedzy, 32% niski, 30% średni i 25% wysoki.

Aby rozpocząć klasyfikację należy otworzyć program Statistica i wczytać dane z pliku Example1.xls (w przypadku danych zapisanych w środowisku Excel warto zapamiętać, że Statistica rozpoznaje tylko te dane, które zapisane zostały w formacie .xls – wersja 97-2003)

Krok 1 – zamykamy okno

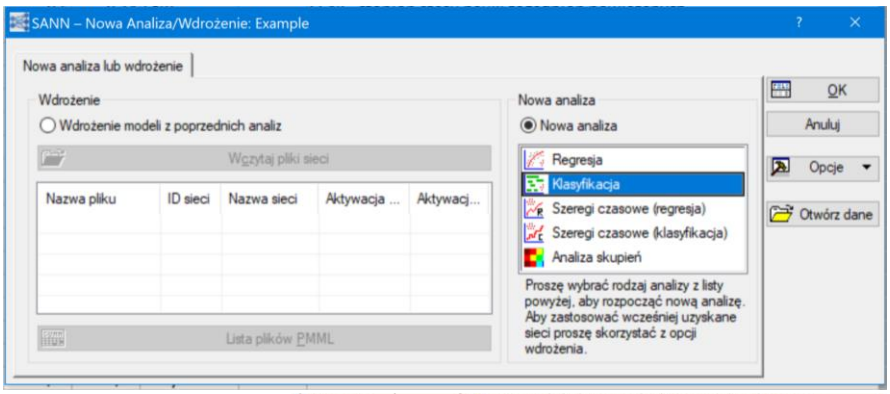


Krok 2 – zaznaczamy, a następnie wskazujemy plik **Example1.xls** i postępujemy według wskazówek wyświetlanych na ekranie.

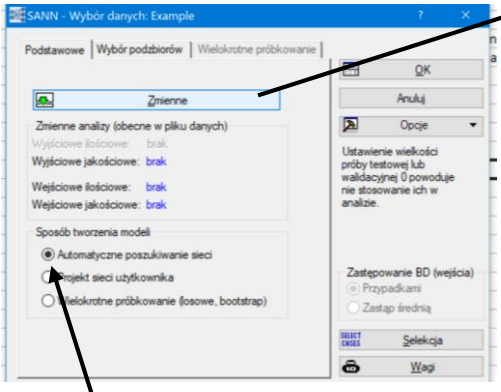


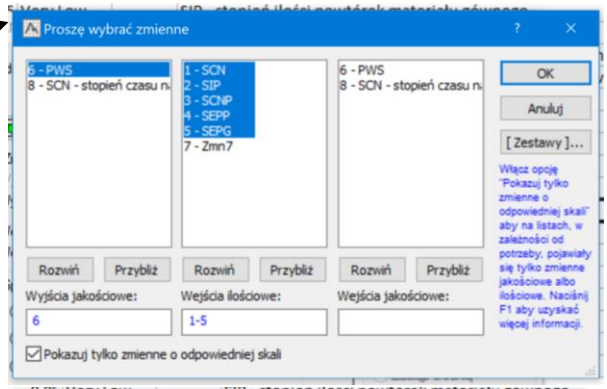
Krok 3 – otwieramy Automatyczne Sieci Neuronowe w DataMining;

Krok 4 - wybieramy typ zadania dla sieci neuronowej i zatwierdzamy – OK (w naszym przypadku jest to „klasyfikacja")



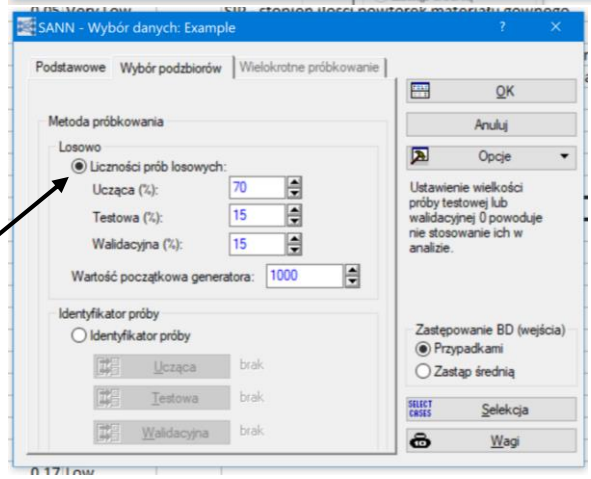
Krok 5 – wskazujemy zmienne





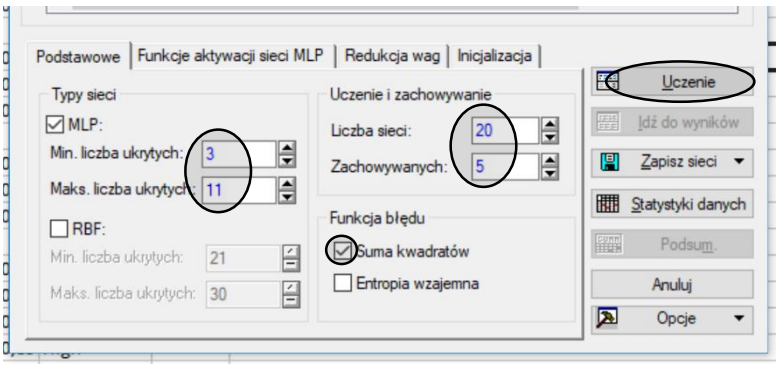
Pozostajemy przy automatycznym wyszukiwaniu sieci (ta opcja powinna być zaznaczona domyślnie).

W zakładce „Wybór podzbiorów" możemy zmienić proporcje podziału zbioru danych na próbę: uczącą, testową i walidacyjną (domyślnie program dzieli zbiór w proporcach 70:15:15 [%]).



Krok 6 – Automatyczne tworzenie sieci neuronowej

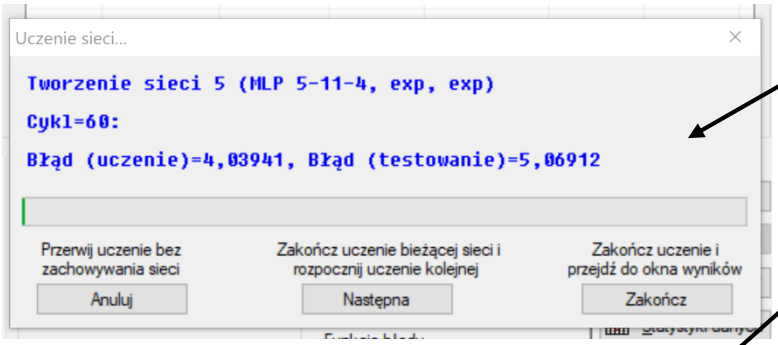
Wybieramy neurony liniowe MLP (domyślne) – możemy ustawić minimalną i maksymalną ilość neuronów w warstwach ukrytych, wybrać ilość sieci do uczenia się oraz ilość sieci jaka ma zastać zapamiętana oraz możemy ustawić funkcję błędu.



Domyślnie program ma ustawione:

- Minimalnie (3) maksymalnie (10) neuronów w ukrytych warstwach
- 20 sieci do nauczenia, zostanie wybranych 5 najlepszych
- Funkcja błędu: suma kwadratów różnic między wartością oczekiwaną a wartością rzeczywistą

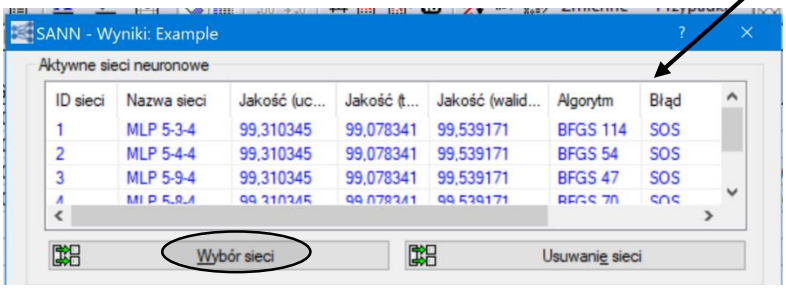
Krok 7 – uczenie sieci



Sieć w trakcie uczenia się.

Tworzenie sieci numer 5 z 20; Sieć MLP „5-11-5" – jest to sieć z 5 neuronami w warstwie wejściowej, 11 neuronami w warstwie ukrytej oraz 4 w warstwie wyjściowej.

WYNIK

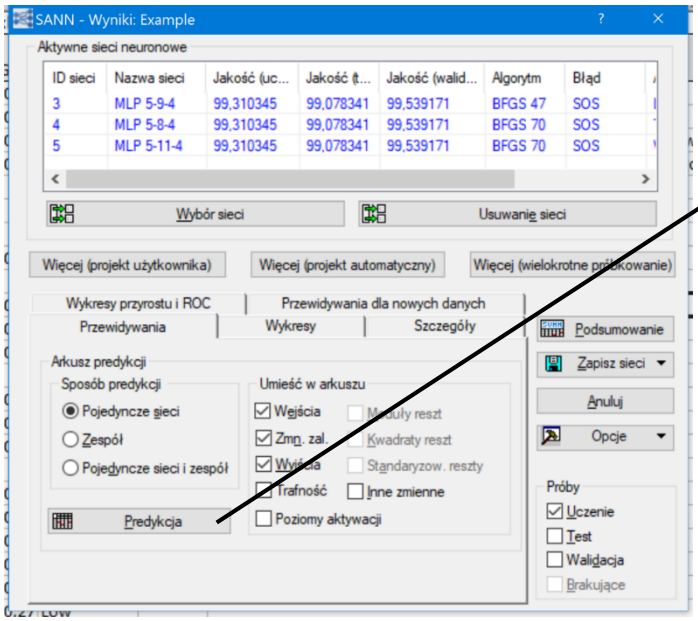


W wyniku uczenia otrzymaliśmy 5 najlepszych sieci. Można zauważyć, że sieci te różnią się między sobą ilością neuronów w warstwie ukrytej, i tak np. sieć numer 1 ma ich 3, sieć numer 2 ma ich 4 a sieć numer 3 ma ich 9.

Każda z sieci opisana jest przez nazwę (mówiącą ile neuronów znajduje się w każdej z warstw), błędy uzyskane dla grup: uczącej, testowej i walidacyjnej, zastosowany algorytm uczenia się (zazwyczaj jest to BFGS czyli Back Propagation) i informację dotyczącą funkcji błędu. W kolejnym kroku należy wybrać sieci, które najlepiej nam odpowiadają – zaznaczamy je myszką w wciśniętym klawiszem CTRL.

ANALIZA

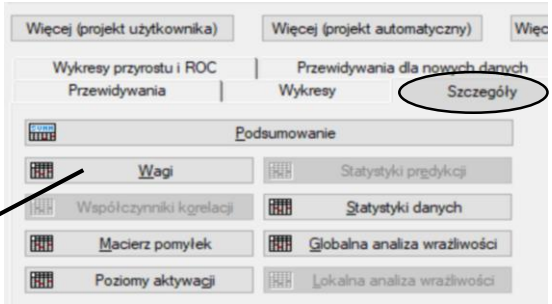
Dla 3 najlepszych sieci mamy możliwość sprawdzenia: przewidywania, wykresów, szczegółowych danych, wykresów przyrostu i ROC i przewidywania dla nowych danych.



Po aktywowaniu funkcji „Predykcja"- zobaczymy, które sieci błędnie sklasyfikowały poziom wiedzy naszego studenta

	SEPG Wejście	PWS Zm.zal	PWS - Wyjście 3. MLP 5-9-4	PWS - Wyjście 4. MLP 5-8-4	PWS - Wyjście 5. MLP 5-11-4
)0	0,770000	High	High	High	High
)0	0,250000	Very Low	Very Low	Very Low	Very Low
)0	0,130000	Very Low	Very Low	Very Low	Very Low
)0	0,240000	Low	Low	Low	Low
)0	0,830000	High	High	High	High
)0	0,130000	Very Low	Very Low	Very Low	Very Low
)0	0,310000	Low	Low	Low	Low
)0	0,270000	Low	Low	Low	Low
)0	0,260000	Middle	Low	Low	Low
)0	0,310000	Low	Low	Low	Low

W zakładce „szczegóły" w podsumowaniu możemy ponownie odczytać dane opisujące poszczególne sieci.



W zakładce „wagi" możemy sprawdzić jakie wartości wag przyjmowały poszczególne neurony w procesie uczenia się.

Wagi sieci (Example)

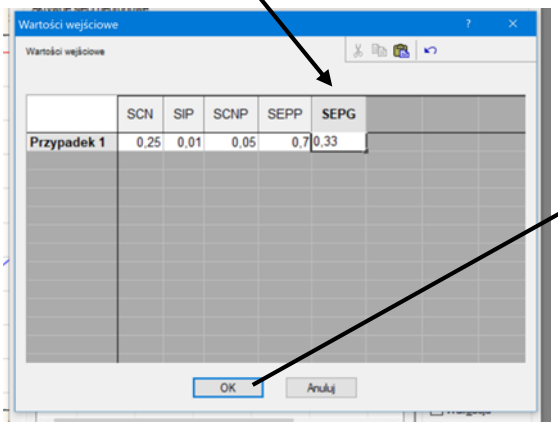
ID wagi	Połączenia 3.MLP 5-9-4	Wartości wag 3.MLP 5-9-4	Połączenia 4.MLP 5-8-4	Wartości wag 4.MLP 5-8-4	Połączenia 5.MLP 5-11-4
1	SCN --> ukryty neuron 1	2,6778	SCN --> ukryty neuron 1	0,11232	SCN --> ukryty neuron 1
2	SIP --> ukryty neuron 1	3,5121	SIP --> ukryty neuron 1	0,52905	SIP --> ukryty neuron 1
3	SCNP --> ukryty neuron 1	0,8159	SCNP --> ukryty neuron 1	0,23380	SCNP --> ukryty neuron 1
4	SEPP --> ukryty neuron 1	8,9022	SEPP --> ukryty neuron 1	0,63694	SEPP --> ukryty neuron 1
5	SEPG --> ukryty neuron 1	18,3274	SEPG --> ukryty neuron 1	-1,29373	SEPG --> ukryty neuron 1
6	SCN --> ukryty neuron 2	0,8680	SCN --> ukryty neuron 2	-0,22529	SCN --> ukryty neuron 2

Zakładka „Macierz pomyłek" informuje o poziomie błędnych wskazań dla poszczególnych sieci.

PWS (Podsumowanie klasyfikacji) (Example)
Próby: Uczenie

		PWS-High	PWS-Low	PWS-Middle	PWS-Very Low	PWS-Wszystkie
3.MLP 5-9-4	Razem	275,0000	326,0000	248,0000	166,0000	1015,000
	Poprawne	275,0000	326,0000	241,0000	166,0000	1008,000
	Niepoprawne	0,0000	0,0000	7,0000	0,0000	7,000
	Poprawne (%)	100,0000	100,0000	97,1774	100,0000	99,310
	Niepoprawne (%)	0,0000	0,0000	2,8226	0,0000	0,690
4.MLP 5-8-4	Razem	275,0000	326,0000	248,0000	166,0000	1015,000
	Poprawne	275,0000	326,0000	241,0000	166,0000	1008,000
	Niepoprawne	0,0000	0,0000	7,0000	0,0000	7,000
	Poprawne (%)	100,0000	100,0000	97,1774	100,0000	99,310
	Niepoprawne (%)	0,0000	0,0000	2,8226	0,0000	0,690
5.MLP 5-11-4	Razem	275,0000	326,0000	248,0000	166,0000	1015,000
	Poprawne	275,0000	326,0000	241,0000	166,0000	1008,000
	Niepoprawne	0,0000	0,0000	7,0000	0,0000	7,000
	Poprawne (%)	100,0000	100,0000	97,1774	100,0000	99,310
	Niepoprawne (%)	0,0000	0,0000	2,8226	0,0000	0,690

W zakładce „przewidywania dla nowych danych" możemy wprowadzić dane („*Wartości na wejściu*") dla których chcemy aby sieci nam sklasyfikowały poziom wiedzy danego studenta, np. (dane wpisane losowo):



#	3.PWS	4.PWS	5.PWS	SCN	SIP	SCNP
1	Middle	Low	Low	0,250000	0,010000	0,050(

W wyniku przeprowadzonej analizy, możemy zauważyć, że sieci 4 i 5 sklasyfikowały poziom studenta jako niski, zaś sieć numer 3 jako średni.

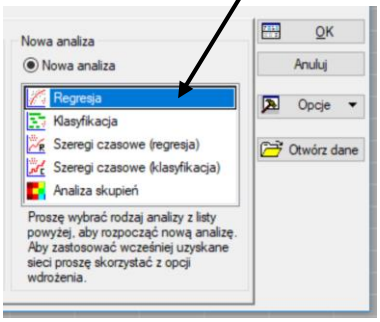
Zadanie do rozwiązania: Sprawdź działanie wybranych sieci dla 5 różnych wymyślonych przypadków, wyciągnij wnioski.

Przykład 2 – zadanie regresyjne

Rozpatrzmy zadanie regresyjne, które ma na celu przewidywanie zapotrzebowania na energię elektryczną instalacji elektrociepłowni pracującej z pełnym obciążeniem. Zbiór danych przedstawia się następująco:

	A	B	C	D	E	F	G	H
1	T	C	W	P	E		T - temperatura	
2	14,96	41,76	1024,07	73,17	463,26		C - ciśnienie	
3	25,18	62,96	1020,04	59,08	444,37		W - wilgotność	
4	5,11	39,4	1012,16	92,14	488,56		P - próżnia przelotowa	
5	20,86	57,32	1010,24	76,64	446,48		E- energia	

Uruchamiamy program Statistica, wczytujemy dane - plik *Example2.xls* i postępujemy podobnie jak we wcześniejszym przykładzie wybierając dla sieci zadanie regresyjne:



Nie mamy już do wyboru typu funkcji błędu gdyż przy zadaniach regresyjnych zawsze jest stosowana funkcja sumy kwadratów różnic.

Po wyborze najlepszych sieci dokonujemy analizy otrzymanych wyników (predykcja, itd.) W analizie predykcji tym razem przedstawione są wartości zmiennej wyjściowej określonej przez poszczególne sieci.
Przy zadaniach regresyjnych warto zrobić w zakładce Wykresy, wykresy rozproszenia. Warto zauważyć, że nie mamy już do dyspozycji macierzy pomyłek.

Zadanie do rozwiązania: Sprawdź działanie sieci dla poniższych przypadków:

	T	C	W	P	E
2	9,44	40	1015,62	81,16	471,32
3	23,49	49,3	1003,35	77,96	442,76
4	4,99	39,04	1020,45	78,89	472,52
5	18,24	58,46	1017,38	86,92	449,63
6	27,49	63,78	1015,43	47,45	445,66
7	13,61	41,16	1020,49	75,09	462,67
8	29,75	73,5	1011,13	67,31	433,63
9	27,38	77,24	1008,25	82,49	435,81
10	18,28	60,1	1009,72	85,79	452,93
11	23,74	62,96	1019,65	63,39	449,48

Wartość energii E została przedstawiona celem weryfikacji działania sieci.

SPRAWOZDANIE

Na podstawie przeprowadzonego ćwiczenia przygotować sprawozdanie, szczególną uwagę zwracając na działanie sieci po procesie uczenia się.