Wydział Fizyki i Informatyki Stosowanej

Sztuczne Sieci Neuronowe

Sprawozdanie z projektu

Aleksandra Poręba
Grzegorz Podsiadło

17 czerwca 2020

Spis treści

1 Wstęp **3**
1.1 Sieć perceptron wielowarstwowy 3

2 Zbiór danych **5**
2.1 Analiza zbioru danych 5
2.2 Korelacja pomiędzy danymi 5

3 Poszukiwanie najlepszej konfiguracji sieci neuronowej **10**
3.1 Wykresy pudełkowe 10
3.2 Otrzymane błędy uczenia i testu dla różnych konfiguracji 11

4 Badanie wpływu ilości kolumn na jakość sieci **18**
4.1 Uzyskane błędy w przypadku usunięcia pierwszej kolumny 18
4.2 Uzyskane błędy w przypadku usunięcia drugiej kolumny 18
4.3 Uzyskane błędy w przypadku usunięcia trzeciej kolumny 19
4.4 Uzyskane błędy w przypadku usunięcia czwartej kolumny 20
4.5 Uzyskane błędy w przypadku usunięcia piątej kolumny 21
4.6 Wnioski 22

5 Przewidywanie wyniku egzaminu bazując na rezultatach pozostałych egzaminów **23**

6 Podsumowanie **26**

1 Wstęp

Tematem projektu było stworzenie sieci neuronowej, za której pomocą mielibyśmy możliwość przewidzenia wyniku egzaminu SAT na podstawie różnych czynników środowiskowych. Rozwiązanie tego problemu można znaleźć na wiele sposobów - w tej pracy zbadane zostały takie aspekty jak:

- Jaka jest najlepsza konfiguracja sieci neuronowej dla tego zbioru danych?
- Czy usunięcie niektórych kolumn wpłynie znacząco na jakość działania sieci?
- Czy da się przewidzieć wynik egzaminu na podstawie znajomości rezultatów pozostałych z zadowalającą jakością?

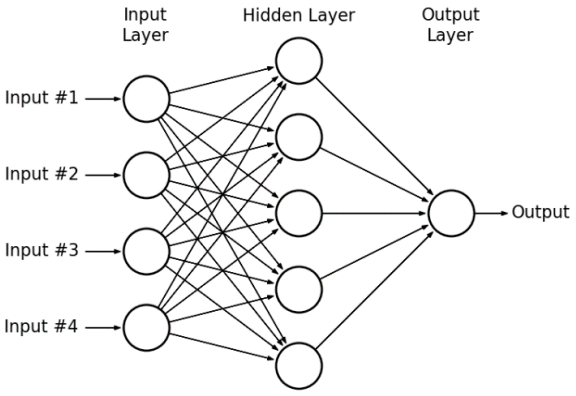
Przed przystąpieniem do implementacji sieci zbiór danych został dokładnie przeanalizowany pod kątem statystyk oraz korelacji parametrów, dzięki czemu łatwiej było określić poprawność uzyskiwanych wyników.

Projekt został zrealizowany w oparciu o środowisko Matlab.

1.1 Sieć perceptron wielowarstwowy

Perceptron wielowarstwowy (*MLP*, ang. *Multilayer Perceptron*) jest rodzajem sztucznej sieci neuronowej. W odróżnieniu od sieci jednowartswowej, takiej jak perceptron, składa się on minimalnie z dwóch warstw: wejściowej i wyjściowej, oraz może zawierać jedną lub więcej warstw ukrytych. Każda warstwa może mieć inną liczbę neuronów, czy funkcję aktywacji.

Schemat sieci został przedstawiony na rysunku poniżej.



Rysunek 1: Schemat sieci *MLP* [4]. Ilość neuronów w warstwie wejściowej zazwyczaj odpowiada ilości parametrów wejściowych, a w warstwie wyjściowej ilości możliwych klas.

Wynik działania perceptronu wielowarstwowego możemy zapisać jako złożenie funkcji[3]:

$$y = f^{(n)}(...f^{(3)}(f^{(2)}(f^{(1)}(x, \theta_1), \theta_2), \theta_3), \theta_n) \tag{1}$$

gdzie:

y - wyjście sieci,

x - wyjście sieci

θ - wagi.

W przeciwieństwie do perceptronu jednowarstwowego sieć ta może być wykorzystywana dla zbiorów, które nie są liniowo separowalne.

2 Zbiór danych

W projekcie został wykorzystany zbiór danych *Students Performance in Exams* [1] pochodzący z repozytorium *Kaggle* [2]. Zawiera on 8 kolumn, zawierających informacje o studentach, takie jak:

- płeć,
- rasa,
- wykształcenie rodzica,
- dieta dostarczana przez szkołę,
- przystąpienie do kursu przygotowawczego,
- wynik egzaminu SAT z części pisemnej,
- wynik egzaminu SAT z części czytania,
- wynik egzaminu SAT z części matematycznej.

W zbiorze wyróżnione zostały dwie płcie, pięć różnych ras, sześć poziomów wykształcenia rodzica (uczęszczanie do szkoły średniej bez jej ukończenia, ukończenie szkoły średniej, uczęszczanie na uczelnię wyższą bez jej ukończenia, ukończenie *community college*, ukończenie studiów licencjackich, ukończenie studiów magisterskich). Dieta określona została przez dwie wartości: standardowa i pomniejszona, a przygotowania do egzaminu jako brak lub kurs ukończony. Wyniki egzaminów są liczbami całkowitymi z zakresu od 0 do 100.

Zbiór danych zawiera informacje dotyczące 1000 różnych studentów, dla wszystkich dane są kompletne. Jako zbiór uczący wybrane zostały kolumny 1-5, a wynikiem działania sieci będą rezultaty kolejnych egzaminów.

Statystyki dotyczące zbioru zostały przedstawione w rozdziale poniżej.

2.1 Analiza zbioru danych

2.2 Korelacja pomiędzy danymi

W ramach projektu został obliczony współczynnik korelacji *Pearsona* pomiędzy danymi z repozytorium. Dzięki niemu możemy sprawdzić w jakim stopniu zależne są od siebie dane, a dzięki temu dostosować czynniki użyte do uczenia sieci.

Do obliczenia współczynnika korelacji została użyta funkcja pakietu MATLAB `corr`.

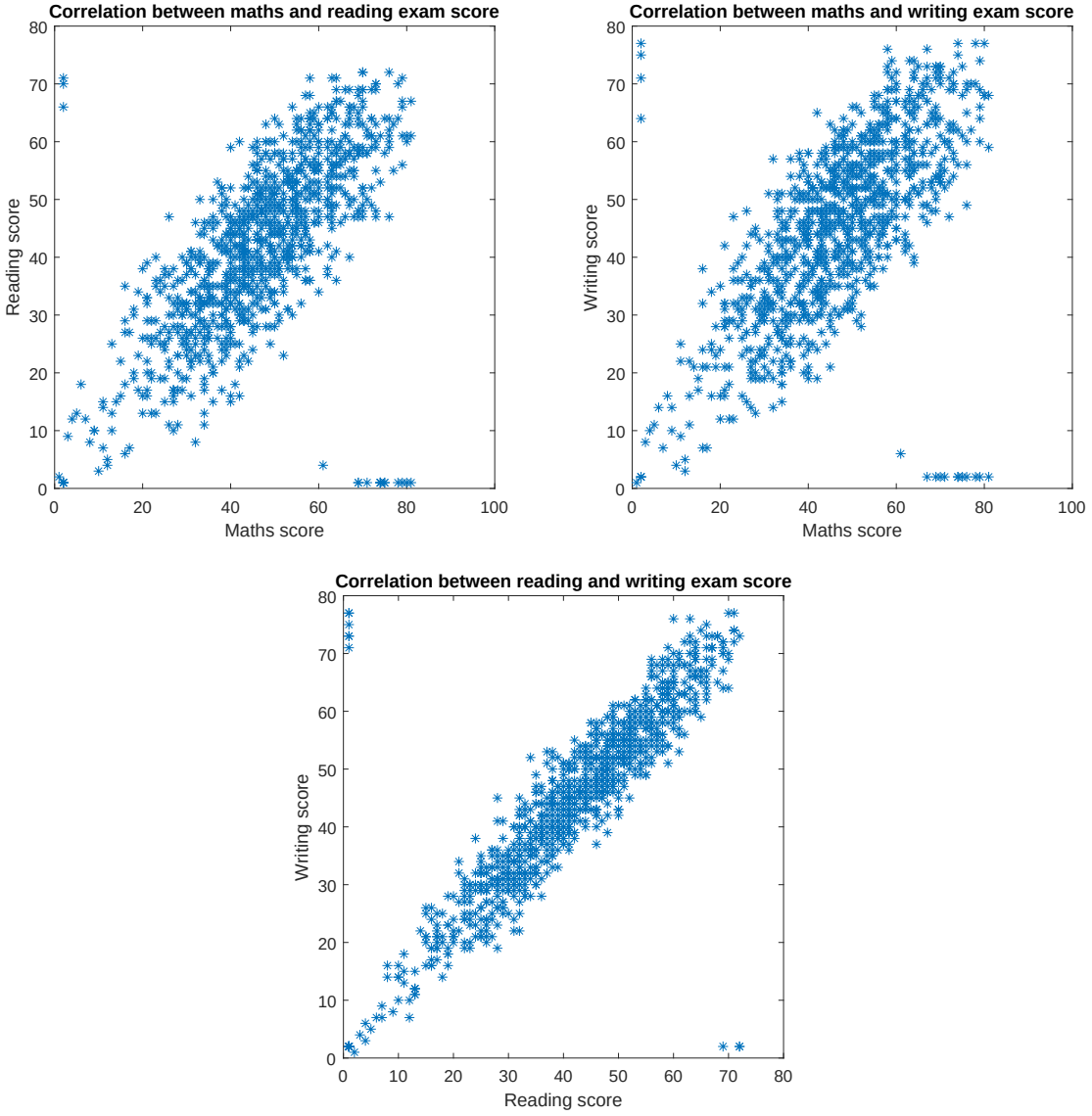
2.2.1 Korelacja pomiędzy wynikami egzaminów

Na początku został wyznaczony współczynnik korelacji pomiędzy wynikami egzaminów. Jako "Egzamin 1" został przyjęty egzamin z części matematycznej, "Egzamin 2" - część z czytania, a "Egzamin 3" egzamin z części pisemnej. Otrzymane wyniki przedstawiono w tabeli poniżej.

Egzamin 1 i 2	Egzamin 1 i 3	Egzamin 2 i 3
0.6453	0.6435	0.8578

Dla wszystkich kombinacji otrzymaliśmy wysokie wartości (większe od 0.6), możemy więc uznać, że dane te są od siebie liniowo zależne. Największa korelacja występuje pomiędzy egzaminem 2 oraz 3, czyli egzaminem z czytania oraz z pisania - współczynnik jest równy 0.86.

Na rysunkach poniżej zależności pomiędzy zbiorami zostały przedstawione w sposób graficzny.



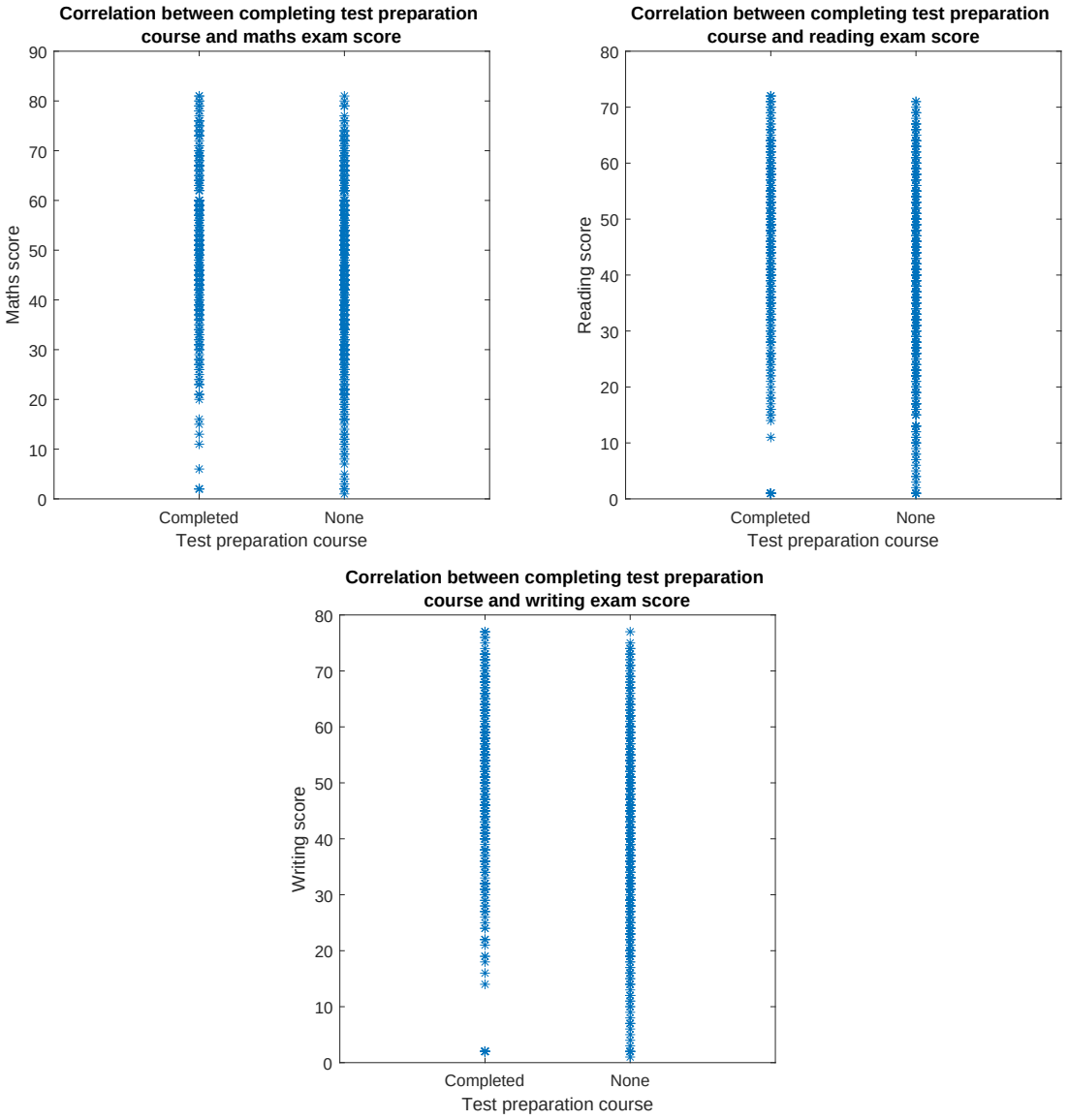
Rysunek 2: Zależności pomiędzy wynikami egzaminów przedstawione w sposób graficzny.

2.2.2 Korelacja pomiędzy czynnikami środowiskowymi a wynikami egzaminów

W dalszej części analizy zbioru danych została zbadana zależność pomiędzy czynnikami, będącymi wejściem sieci neuronowej, a wynikami kolejnych egzaminów. Otrzymane współczynniki korelacji przedstawiono w tabeli poniżej.

Czynnik środowiskowy	Egzamin 1	Egzamin 2	Egzamin 3
Płeć	0.1558	-0.1886	-0.2396
Rasa	0.1771	0.0770	0.0907
Wykształcenie rodzica	-0.0584	-0.0306	-0.0571
Przystąpienie do kursu	0.3269	0.1906	0.2182
Dieta	-0.1564	-0.1838	-0.2647

Otrzymane wartości są dość niskie, nie istnieje wyraźna korelacja pomiędzy którąś z tych cech, a wynikami. Najmniejszą zależność obserwujemy pomiędzy wynikami, wykształceniem rodziców - są one najbliższe zeru. Największe znaczenie ma przystąpienie do kursu przygotowawczego, w dalszej kolejności dieta oraz płeć.



Rysunek 3: Zależności pomiędzy wynikami egzaminów a ukończeniem kursu przygotowawczego przedstawiona w sposób graficzny. Na rysunkach można zauważyć, że osoby które ukończyły kurs, uzyskiwały wynik większy niż 15, ale osoby, które nie ukończyły kursu uzyskiwały równie wysokie wyniki, co osoby po kursie. Wyjątek stanowi egzamin z części matematycznej - tu osób, które uzyskały wysoki wynik bez ukończenia kursu jest mniej.

3 Poszukiwanie najlepszej konfiguracji sieci neuronowej

Przeprowadzono badanie błędu średniokwadratowego dla uczenia oraz testu różnych konfiguracji funkcji aktywacji, ilości warstw ukrytych oraz ilości neuronów w poszczególnych warstwach. Przetestowano konfigurację sieci o jednej oraz dwóch warstwach, z wykorzystaniem różnych kombinacji funkcji:

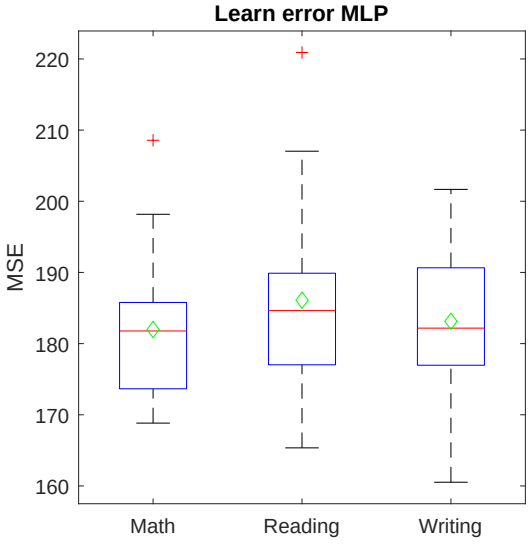
1. `logsig`
2. `tansig`
3. `purelin`
4. `radbas`

Dla każdej konfiguracji zbadano również wpływ ilości neuronów w poszczególnych warstwach. Ilość neuronów zmieniano między 10 a 100 ze skokiem 10. Wykresy błędów dla wybranych konfiguracji znajdują się na następnych stronach.

3.1 Wykresy pudełkowe

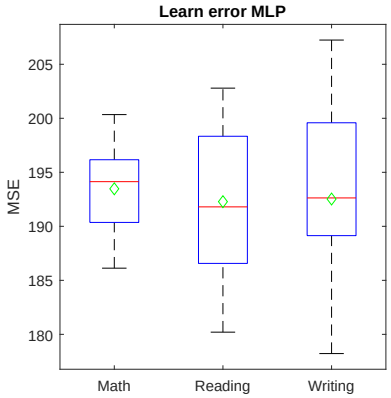
Otrzymane błędy uczenia oraz testu dla różnych konfiguracji przedstawiono przy pomocy tak zwanych wykresów pudełkowych z wąsami. Przykład takiego wykresu pokazano na rysunku 4.

Na wykresie zielone diamenty symbolizują wartości średnie z N prób, czerwona linia wewnątrz oznacza medianę. Górna granica pudełka to trzeci kwartyl (Q_3), natomiast dolna to kwaryl pierwszy (Q_1). Dolny i górny "wąs" reprezentują kolejno wartości $Q_1 - 1.5IQR$ oraz $Q_3 + 1.5IQR$, gdzie IQR to rozstęp ćwiartkowy, obliczany jako $IQR = Q_3 - Q_1$. Czerwonymi plusami zaznaczono wartości odstające.

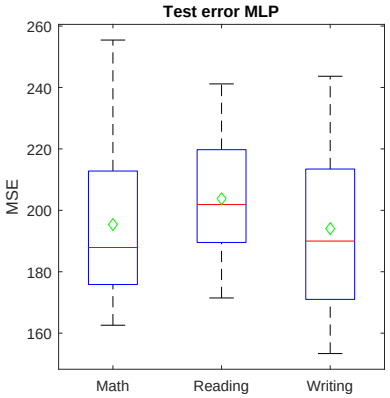


Rysunek 4: Przykład wykresu pudełkowego pokazującego błędy dla różnych egzaminów.

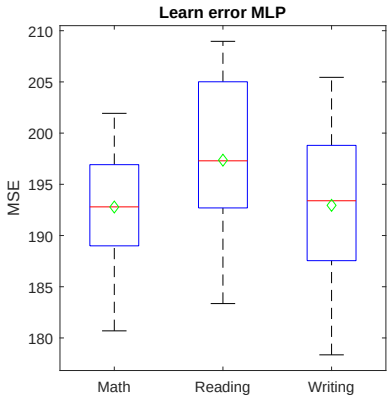
3.2 Otrzymane błędy uczenia i testu dla różnych konfiguracji



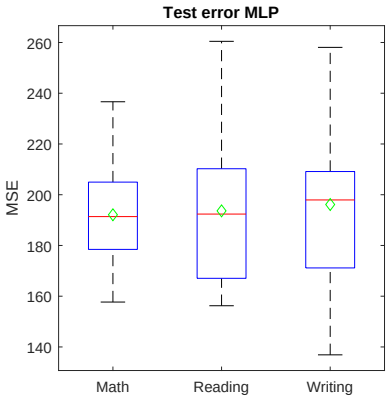
(a) MSE uczenia dla 20 neuronów.



(b) MSE testu dla 20 neuronów.



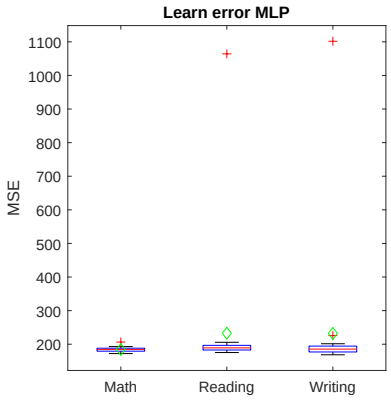
(c) MSE uczenia dla 50 neuronów.



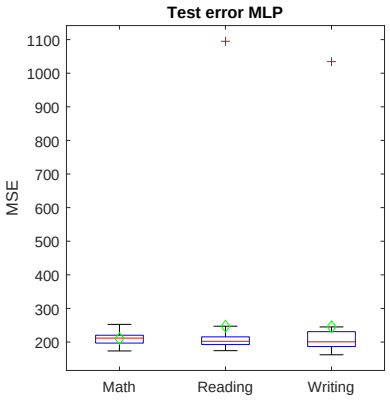
(d) MSE testu dla 50 neuronów.

Rysunek 5: Błąd uczenia i testowania sieci dla X prób dla kolejnych egzaminów, **funkcje: purelin, purelin**.

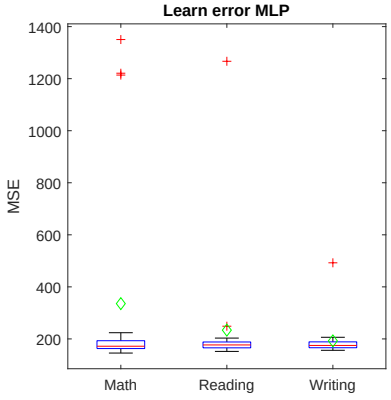
Pierwszym omawianym zestawieniem jest prosty przykład dla kombinacji funkcji purelin oraz purelin. Otrzymane wartości błędów były jednymi z najmniejszych wśród wszystkich konfiguracji, z akceptowalnym odchyleniem. W przypadku tak prostej sieci ilość neuronów nie wpłynęła praktycznie wcale na otrzymane wyniki.



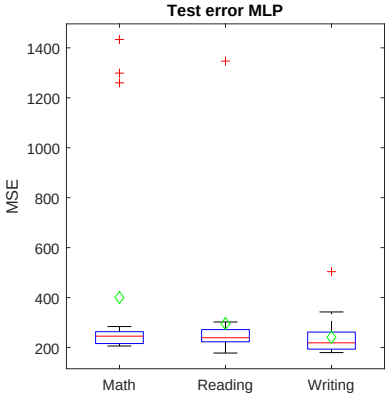
(a) MSE uczenia dla 20 neuronów.



(b) MSE testu dla 20 neuronów.



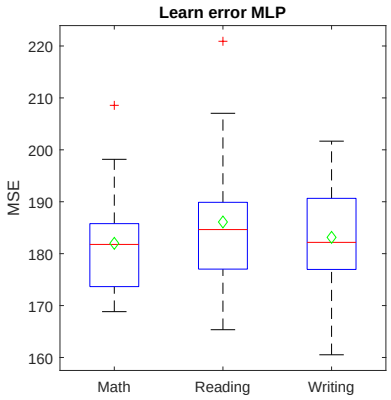
(c) MSE uczenia dla 50 neuronów.



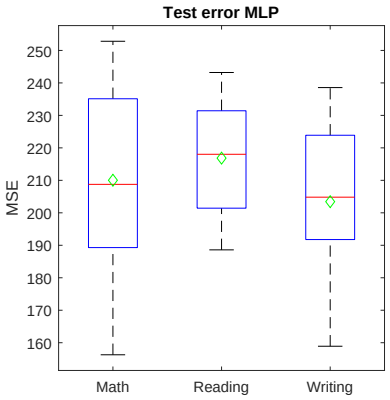
(d) MSE testu dla 50 neuronów.

Rysunek 6: Błąd uczenia i testowania sieci dla X prób dla kolejnych egzaminów, **funkcje: tansig, tansig**.

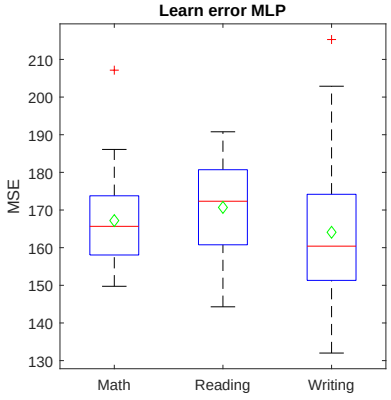
Przykładem jednej z konfiguracji, która daje mierne rezultaty jest konfiguracja z wykorzystaniem funkcji tansig oraz tansig na wyjściu. Wartości błędów były jednymi z większych, jednak istotny jest fakt, że wyniki dla trzeciego egzaminu były zbieżne, w przeciwieństwie do większości pozostałych konfiguracji. Ilość neuronów nie miała dużego wpływu na otrzymywane błędy.



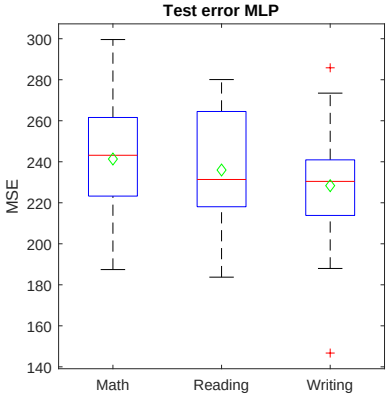
(a) MSE uczenia dla 20 neuronów.



(b) MSE testu dla 20 neuronów.



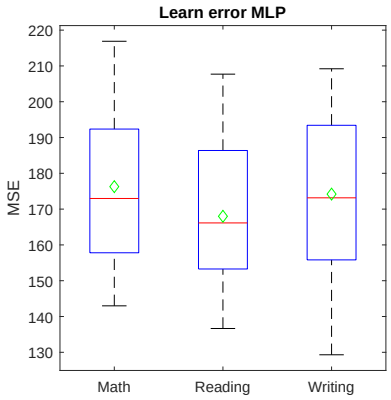
(c) MSE uczenia dla 50 neuronów.



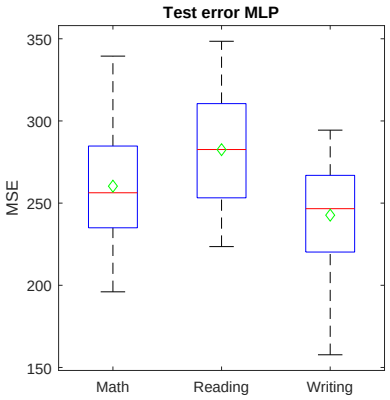
(d) MSE testu dla 50 neuronów.

Rysunek 7: Błąd uczenia i testowania sieci dla X prób dla kolejnych egzaminów, **funkcje: tansig, purelin**.

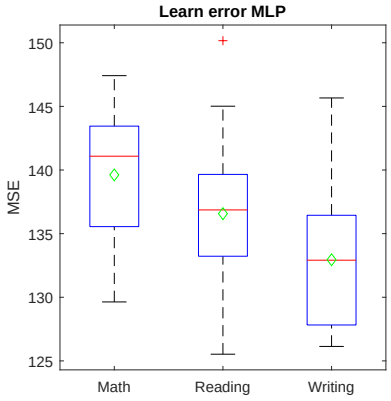
Kombinacja tansig oraz purelin daje bardzo niskie wartości błędu w porównaniu z pozostałymi metodami, jednak z dość dużym rozstrzałem między otrzymywanymi wartościami. Błedy uczenia nieznacznie spadały wraz ze wzrostem ilości neuronów w warstwie, jednak tendencja dla testu była odwrotna. Najlepsze wyniki udało się uzyskać dla 20 oraz 30 neuronów.



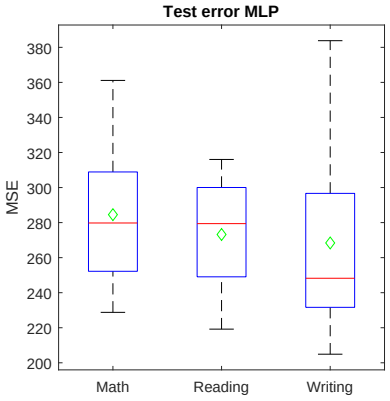
(a) MSE uczenia dla 40 neuronów w pierwszej warstwie, 20 w drugiej.



(b) MSE testu dla 40 neuronów w pierwszej warstwie, 20 w drugiej.



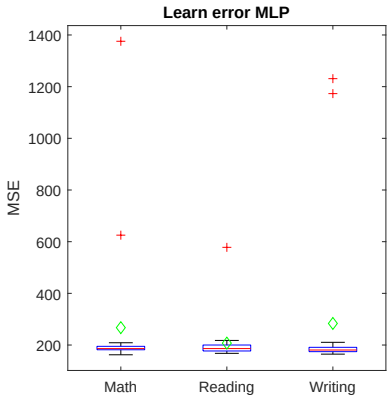
(c) MSE uczenia dla 100 neuronów w pierwszej warstwie, 50 w drugiej.



(d) MSE testu dla 100 neuronów w pierwszej warstwie, 50 w drugiej.

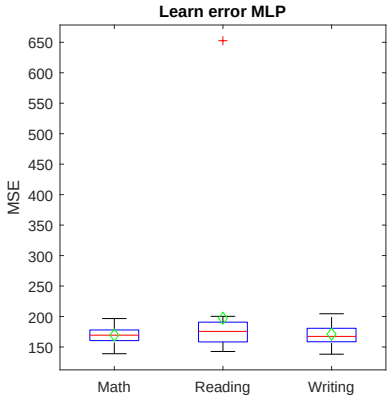
Rysunek 8: Błąd uczenia i testowania sieci dla X prób dla kolejnych egzaminów, **funkcje: radbas, tansig, purelin**.

Postanowiono również przetestować sieci o dwóch warstwach w bardziej egzotycznych konfiguracjach, dla przykładu konfiguracja radbas, tansig, purelin dała zadziwiająco niskie wartości błędów dla próby uczenia, który gwałtownie zmieniał się wraz z dokładaniem ilości neuronów. Pod względem błędu uczenia była to najlepsza konfiguracja.

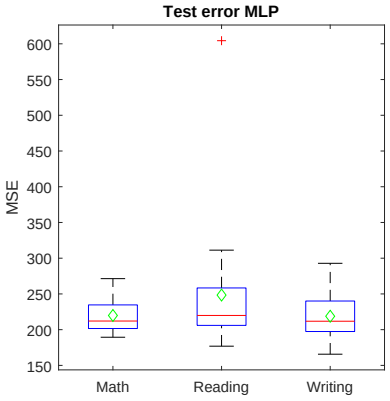


(a) MSE uczenia dla 50 neuronów w pierwszej warstwie, 50 w drugiej.

(b) MSE testu dla 50 neuronów w pierwszej warstwie, 50 w drugiej.



(c) MSE uczenia dla 50 neuronów w pierwszej warstwie, 50 w drugiej.



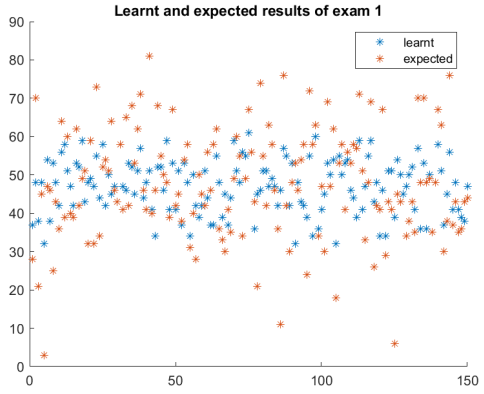
(d) MSE testu dla 50 neuronów w pierwszej warstwie, 50 w drugiej.

Rysunek 9: Błąd uczenia i testowania sieci dla X prób dla kolejnych egzaminów, **funkcje: logisg, tansig, tansig**.

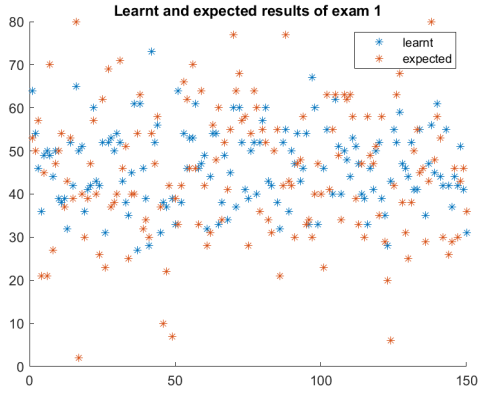
Kombinacja logsig, tansig oraz tansig dała ciekawe rezultaty, wartości błędu dla wszystkich prób nie licząc trzeciego egzaminu były do siebie zbliżone. Wyniki dla trzeciego egzaminu jednak zbyt odstają od reszty, by móc brać tą konfiguracje w kolejnych badaniach.

W zestawieniu nie objęto wykresów dla pozostałych kombinacji (na przykład tansig, tansig), gdyż niezależnie od ilości neuronów wyniki były bardzo złe.

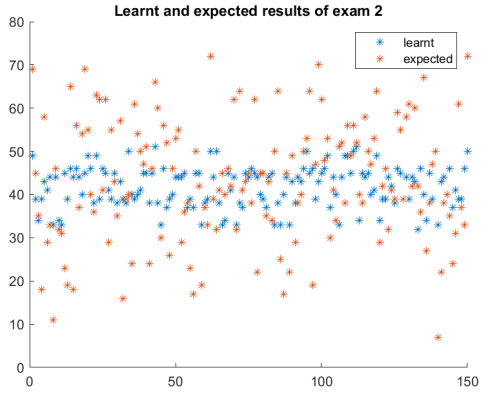
Poniżej przedstawione zostały przykładowe wyniki dla najlepszej (purelin, purelin, 20 neuronów) i najgorszej sieci (tansig, tansig, 50 neuronów w warstwie ukrytej), dla zbioru testującego.



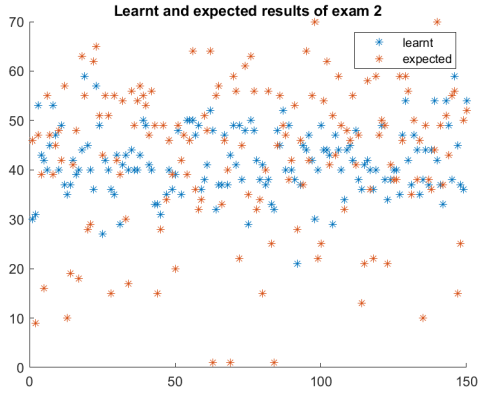
(a) Oczekiwane i znalezione wyniki dla najlepszej sieci dla egzaminu 1 (matematyka).



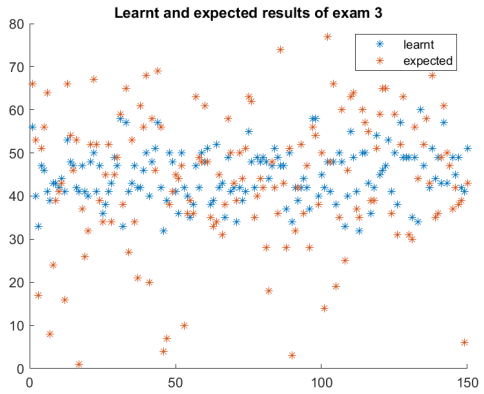
(b) Oczekiwane i znalezione wyniki dla najgorszej sieci dla egzaminu 1 (matematyka).



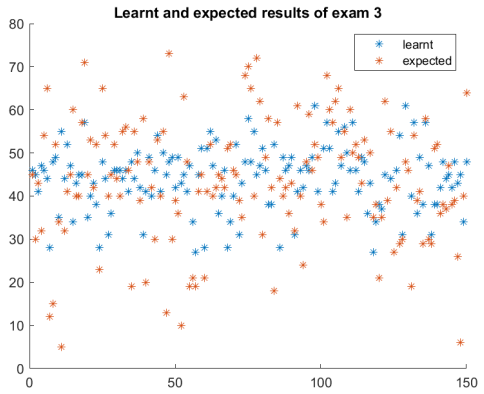
(c) Oczekiwane i znalezione wyniki dla najlepszej sieci dla egzaminu 2 (czytanie).



(d) Oczekiwane i znalezione wyniki dla najgorszej sieci dla egzaminu 2 (czytanie).



(e) Oczekiwane i znalezione wyniki dla najlepszej sieci dla egzaminu 3 (pisanie).



(f) Oczekiwane i znalezione wyniki dla najgorszej sieci dla egzaminu 3 (pisanie).

Powyższe wyniki pochodzą z pojedynczego uruchomienia programu, przy każdym uczeniu wynik działania będzie trochę inny.

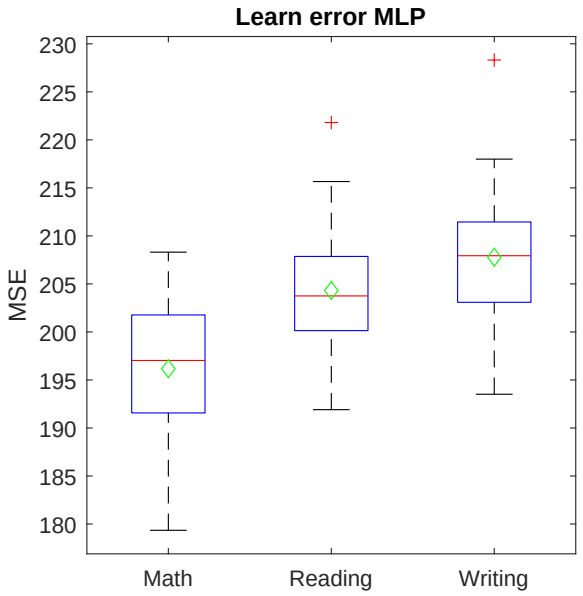
Na rysunkach możemy zauważyć, że dla najlepszej sieci wyniki pochodzące z sieci są bardziej skupione w okolicy średniej liczby punktów. W przypadku najgorszej sieci otrzymane dane są bardziej "rozproszone" - nieliniowy kształt odpowiedzi niekorzystnie wypływa na wyniki sieci.

4 Badanie wpływu ilości kolumn na jakość sieci

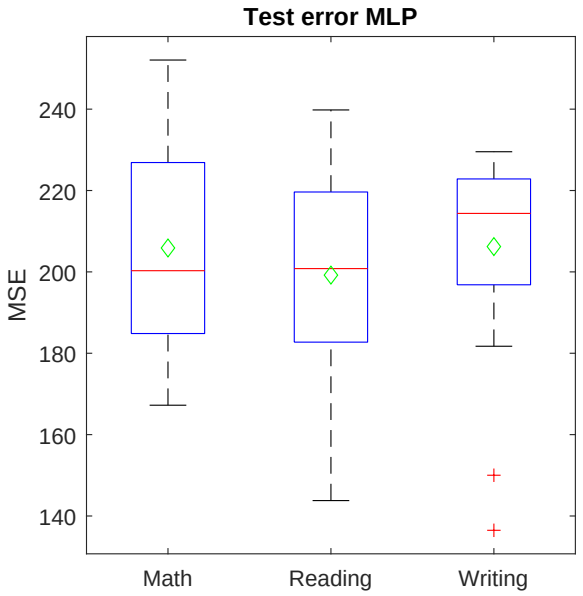
Dla wybranych najlepszych parametrów sieci zostało przeprowadzone uczenie ze zmniejszoną ilością kolumn. Celem tego zabiegu było zbadanie, jaki wpływ mają te czynniki na jakość działania sieci.

4.1 Uzyskane błędy w przypadku usunięcia pierwszej kolumny

Ze zbioru uczącego została usunięta kolumna informująca o **płci** studenta. Wyniki działania takiej sieci przedstawiono poniżej.



(a) MSE uczenia dla 20 neuronów w pierwszej warstwie.

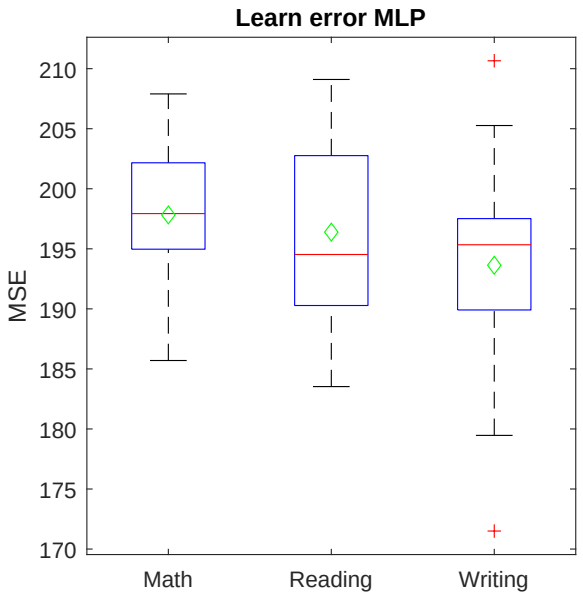


(b) MSE testu dla 20 neuronów w pierwszej warstwie.

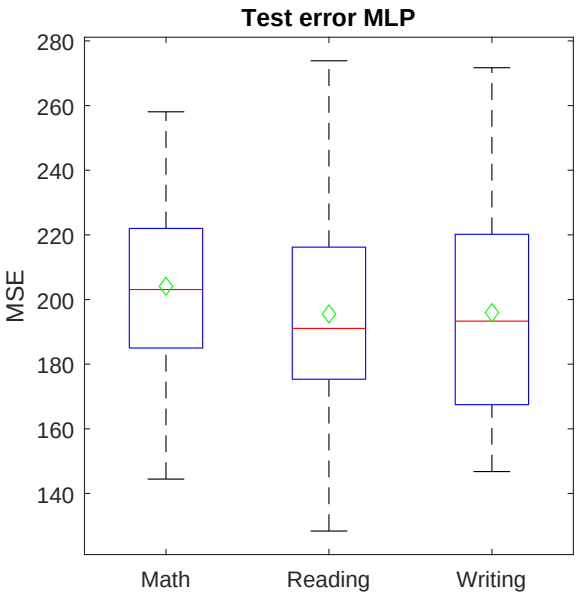
Rysunek 11: Błąd uczenia i testowania sieci w przypadku usunięcia pierwszej kolumny z danymi, **funkcje: purelin, purelin**.

4.2 Uzyskane błędy w przypadku usunięcia drugiej kolumny

Poniżej przedstawione zostały błędy uczenia oraz testowania sieci przy usunięciu kolumny informującej o **rasie** studenta.



(a) MSE uczenia dla 20 neuronów w pierwszej warstwie.

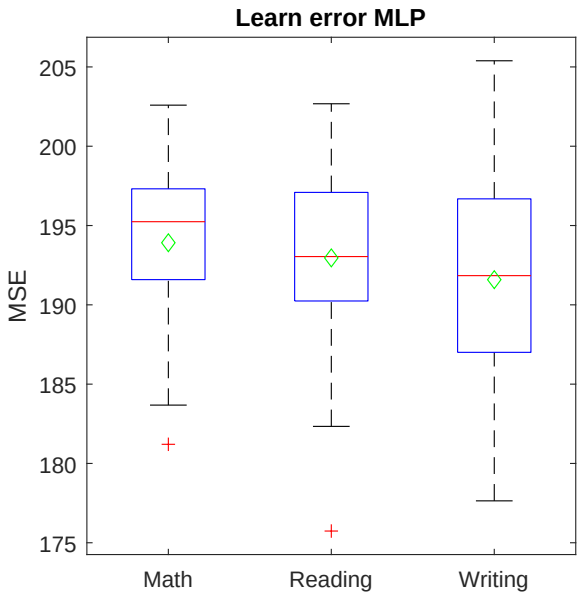


(b) MSE testu dla 20 neuronów w pierwszej warstwie.

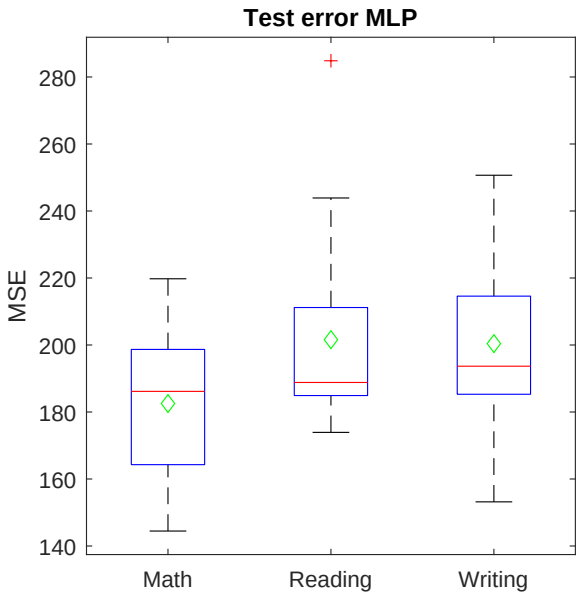
Rysunek 12: Błąd uczenia i testowania sieci w przypadku usunięcia drugiej kolumny z danymi, **funkcje: purelin, purelin**.

4.3 Uzyskane błędy w przypadku usunięcia trzeciej kolumny

Ze zbioru uczącego usunięta została kolumna informująca o wykształceniu rodzica.



(a) MSE uczenia dla 20 neuronów w pierwszej warstwie.

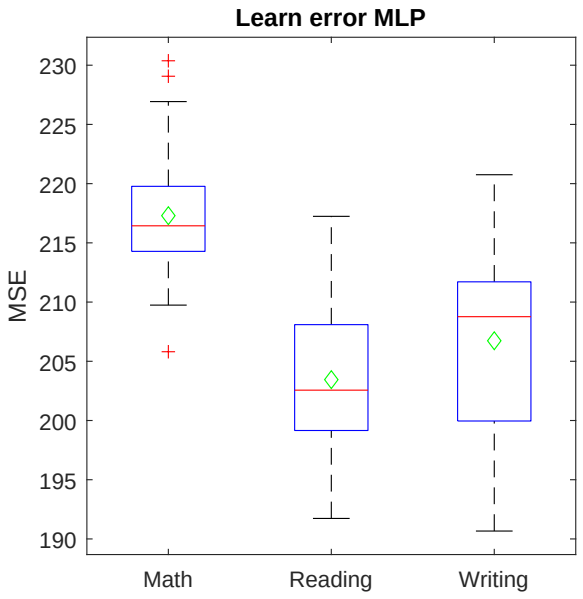


(b) MSE testu dla 20 neuronów w pierwszej warstwiej.

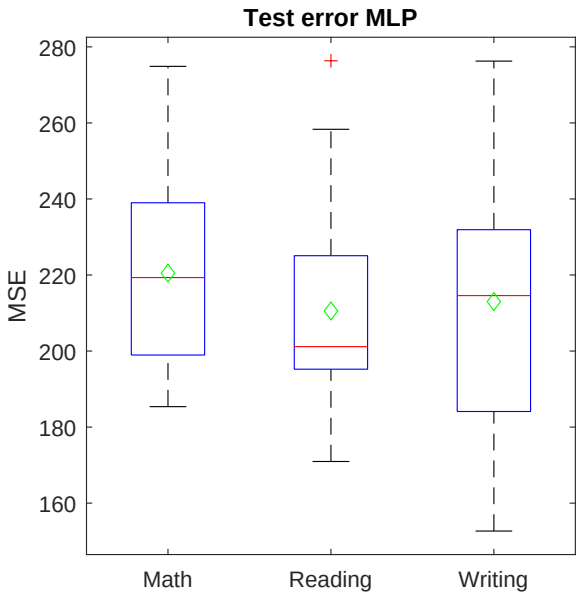
Rysunek 13: Błąd uczenia i testowania sieci w przypadku usunięcia trzeciej kolumny z danymi, **funkcje: purelin, purelin**.

4.4 Uzyskane błędy w przypadku usunięcia czwartej kolumny

Poniżej przedstawiono błędy sieci w przypadku usunięcia czwartej kolumny, informującej o diecie studenta.



(a) MSE uczenia dla 20 neuronów w pierwszej warstwie.

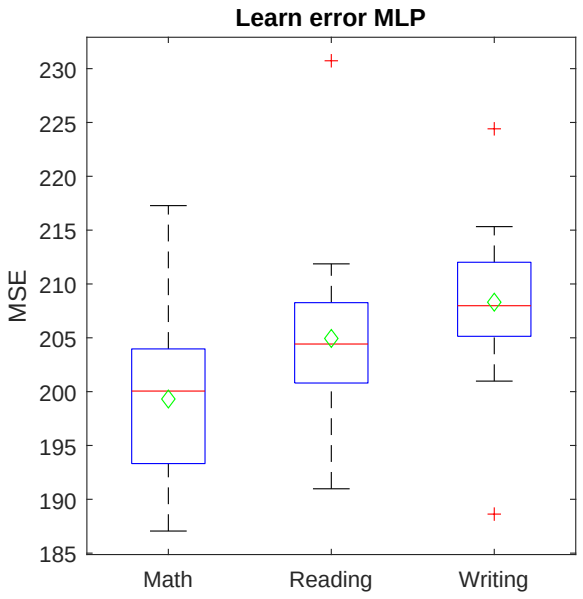


(b) MSE testu dla 20 neuronów w pierwszej warstwiej.

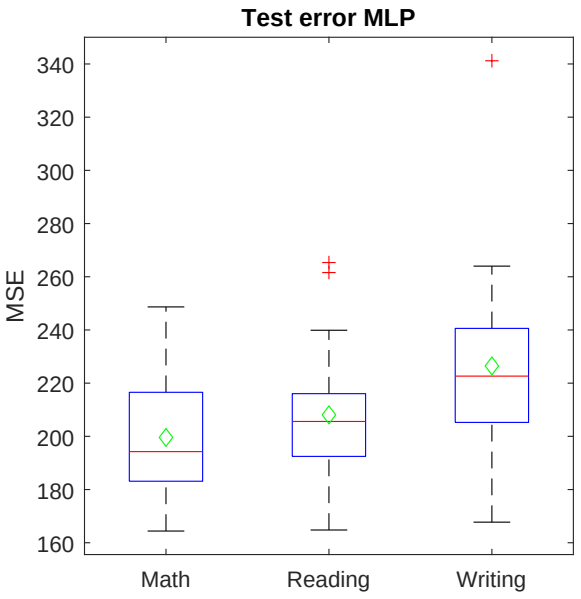
Rysunek 14: Błąd uczenia i testowania sieci w przypadku usunięcia czwartej kolumny z danymi, **funkcje: purelin, purelin**.

4.5 Uzyskane błędy w przypadku usunięcia piątej kolumny

Jako ostatnia ze zbioru uczącego została usunięta kolumna informująca o ukończeniu kursu przygotowawczego. Otrzymane wyniki przedstawiono poniżej.



(a) MSE uczenia dla 20 neuronów w pierwszej warstwie.



(b) MSE testu dla 20 neuronów w pierwszej warstwiej.

Rysunek 15: Błąd uczenia i testowania sieci w przypadku usunięcia piątej kolumny z danymi, **funkcje: purelin, purelin**.

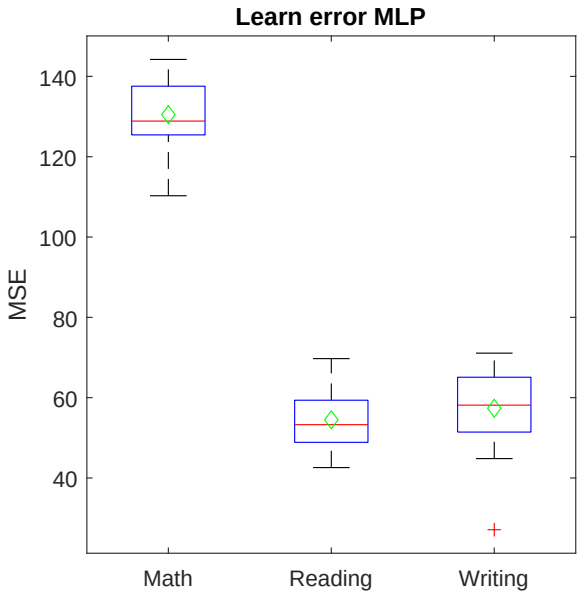
4.6 Wnioski

Badano wersję z funkcjami purelin, purelin oraz 20 neuronami w pierwszej warstwie.

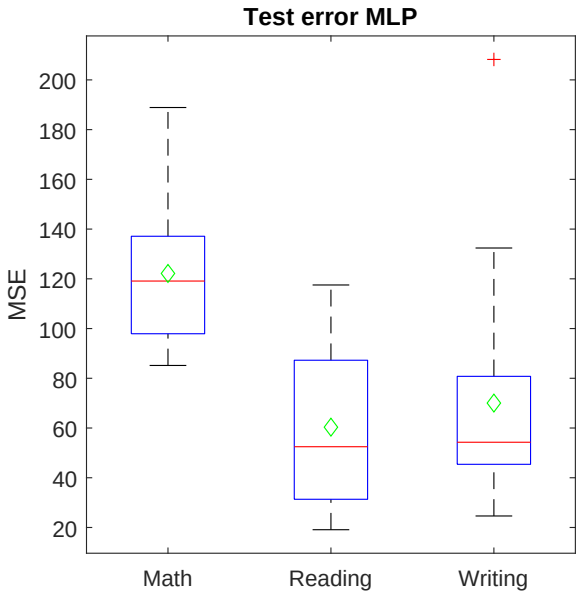
5 Przewidywanie wyniku egzaminu bazując na rezultatach pozostałych egzaminów

Jak zostało zauważone podczas badania korelacji, wyniki egzaminów są od siebie zależne (wysoki współczynnik korelacji). Korzystając z wyników z dwóch egzaminów podjęto próbę stworzenia sieci obliczającą wynik z trzeciego egzaminu. Sieć została skonfigurowana według najlepszych znalezionych parametrów.

Otrzymane błędy testowania zostały przedstawione poniżej.



(a) MSE uczenia dla 20 prób, dla kolejnych egzaminów.

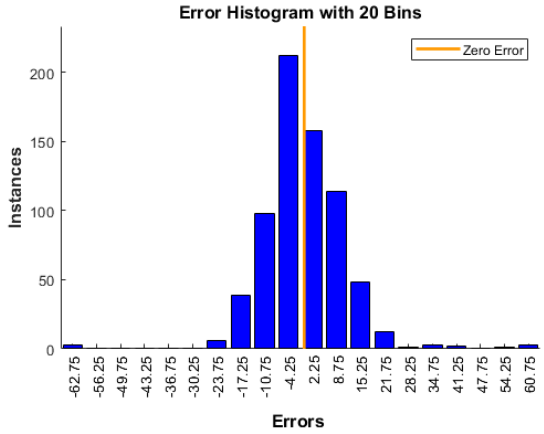


(b) MSE testowania dla 20 prób, dla kolejnych egzaminów.

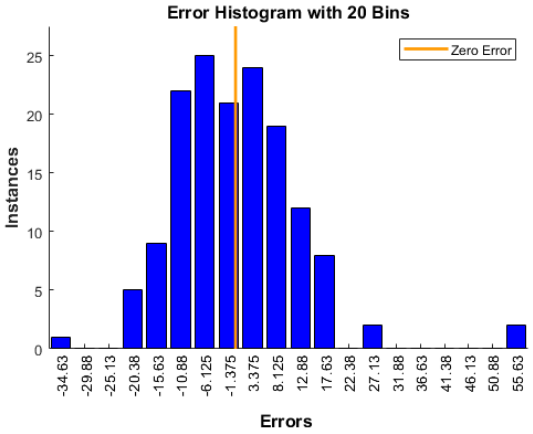
Na podstawie obserwacji wielkości błędów, można zauważyć, że dla egzaminów 2 oraz 3 sieć osiąga dobre wyniki - na poziomie wartości 50. Są one o wiele niższe niż dla sieci uczonej czynnikami środowiskowymi.

Otrzymane wyniki można powiązać z analizowaną wcześniej korelacją danych - wyniki egzaminów 2 i 3 są ze sobą w większym stopniu powiązane, niż egzamin 1 z egzaminem 2 lub 3.

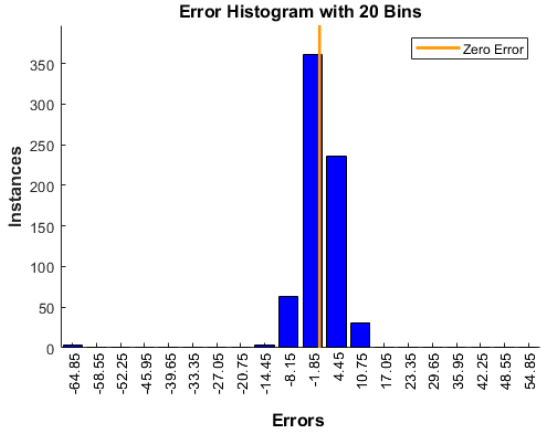
Na rysunkach poniżej został przedstawiony za pomocą histogramu rozkład błędów, postały przy pojedynczym uruchomieniu. Za każdym razem histogramy te różniły się między sobą lecz zaobserwowane tendencje pozostawały takie same.



(a) Rozkład błędów dla zbioru uczącego dla egzaminu 1 (matematyka).

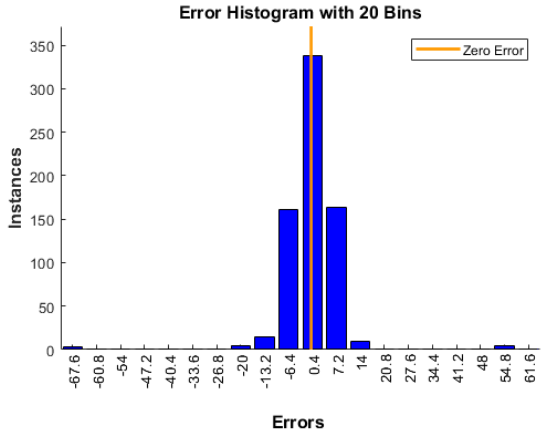


(b) Rozkład błędów dla zbioru testującego dla egzaminu 1 (matematyka).

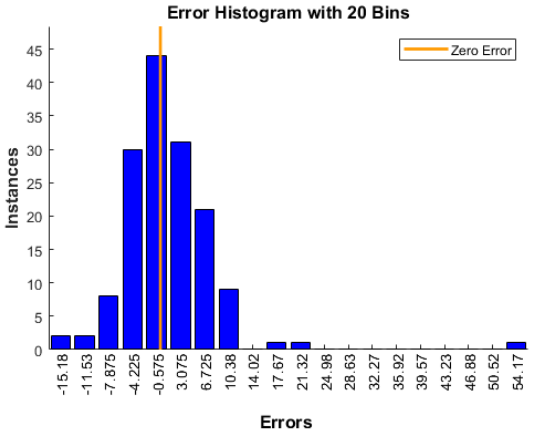


(c) Rozkład błędów dla zbioru uczącego dla egzaminu 2 (czytanie).

(d) Rozkład błędów dla zbioru testującego dla egzaminu 2 (czytanie).



(e) Rozkład błędów dla zbioru uczącego dla egzaminu 3 (pisanie).



(f) Rozkład błędów dla zbioru testującego dla egzaminu 3 (pisanie).

Największa ilość błędów we wszystkich przypadkach to te bliskie zeru. Przy testowaniu zbioru uczącego wartości są bardziej skoncentrowane, szczególnie dla egzaminów 2 oraz 3. Chociaż różnica pomiędzy wartością oczekiwaną a doświadczalną jest czasem duża, sieć w większości przypadków potrafi przewidzieć wynik egzaminu z dokładnością do 10 punktów.

6 Podsumowanie

W przypadku uczenia sieci wynikami egzaminów, jakość jej działania jest o wiele wyższa niż w przypadku uczenia jej czynnikami środowiskowymi. Powiązane jest to z faktem, że korelacja pomiędzy kolumnami z wynikami egzaminów jest o wiele większa niż pomiędzy egzaminami a czynnikami środowiskowymi. Jednak, gdy chcielibyśmy użyć sieci w prawdziwym życiu, przed egzaminami nie dysponujemy taką wiedzą. Można by było posłużyć się wynikami egzaminów próbnych - jednak jako że nie były one zawarte w naszym zbiorze, więc ta hipoteza musiałaby zostać zbadana.

Oprócz wyników egzaminu, duży wpływ ma ukończenie kursu przygotowawczego - szczególnie z matematyki. Usunięcie tej kolumny pomniejszyło jakość działania sieci. Dla pozostałych dwóch zbiorów podobnie duży wpływ ma jeszcze płeć oraz dieta studenta.

W żadnej konfiguracji nie udało nam się uzyskać bardzo dobrych wyników, co można uznać za pozytywną informację dla studentów. Ich wyniki nie są zdeterminowane rasą, płcią czy rezultatami pozostałych egzaminów, a ich realną pracą i wysiłkiem włożonym w przygotowanie (choć samo ukończenie kursu nie ma bardzo dużego znaczenia). Do stworzenia idealnej sieci potrzebna by była większa ilość danych, na przykład na temat samodzielnej pracy studenta, czy wyników egzaminów próbnych.

Gdy przyjmiemy, że nasza sieć powinna znajdować wynik w granicach błędu np 5 punktów (na 100 wszystkich możliwych) jakość wyników sieci wzrasta.

Literatura

[1] Zbiór danych *Students Performance in Exams* - https://www.kaggle.com/spscientist/students-performance-in-exams

[2] Platforma *Kaggle* - https://www.kaggle.com/

[3] Perceptron wielowarstwowy - https://pl.wikipedia.org/wiki/Perceptron_wielowarstwowy

[4] ASSESSMENT OF ARTIFICIAL NEURAL NETWORK FOR BATHYMETRY ESTIMATION USING HIGH RESOLUTION SATELLITE IMAGERY IN SHALLOW LAKES: CASE STUDY EL BURULLUS LAKE. - https://www.researchgate.net/publication/303875065_ASSESSMENT_OF_ARTIFICIAL_NEURAL_NETWORK_FOR_BATHYMETRY_ESTIMATION_USING_HIGH_RESOLUTION_SATELLITE_IMAGERY_IN_SHALLOW_LAKES_CASE_STUDY_EL_BURULLUS_LAKE