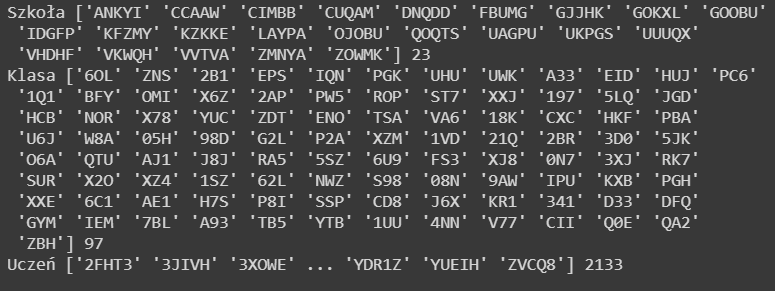
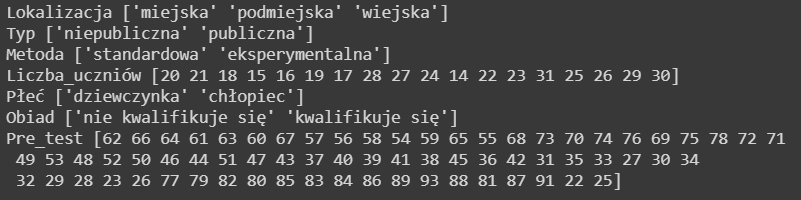
Raport 4A

Dawid Sikorski 291951

<https://colab.research.google.com/drive/10mODBszrAmavSkqzaZ-odqEptejif9-Q?usp=sharing>

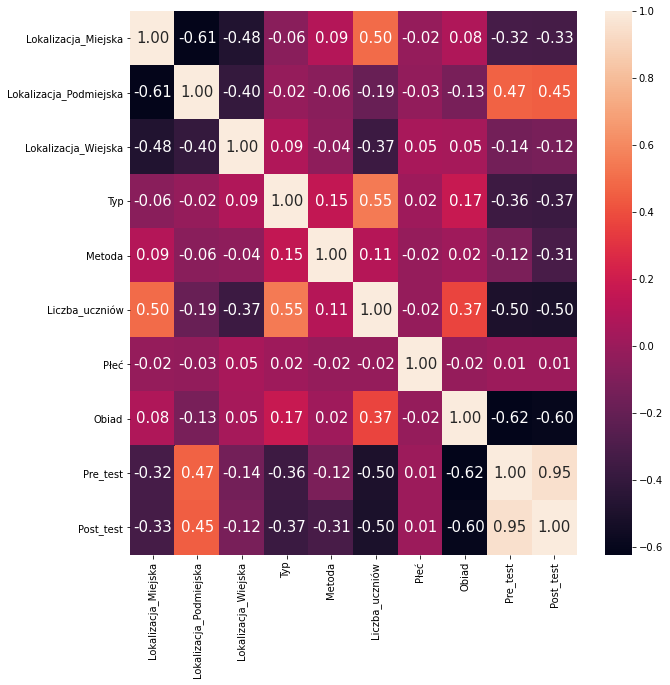
Celem zadania, było przewidzenie ilości punktów z testu dla amerykańskich uczniów z szkół podstawowych. Do realizacji zadania wykorzystano dane zebrane w pliku ‘dane\_testy.cvs’, w którym znajduje się 2133 obserwacji z trzema zmiennymi identyfikującymi, pięcioma jakościowymi, oraz trzema ilościowymi w czym wliczona została zmienna celu ‘post\_test’.





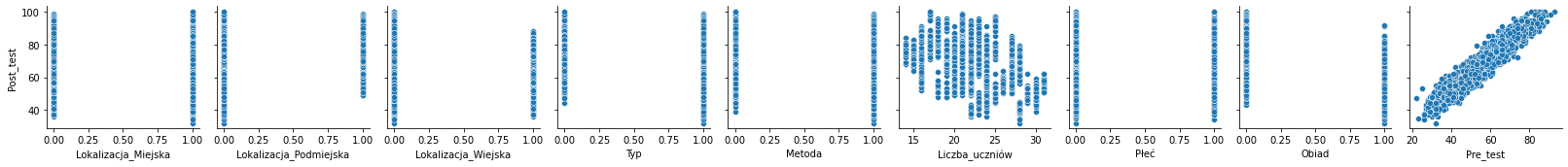
*Analiza danych z pliku*

Plik nie zawierał braków danych. W zbiorze umieszczono dane z 23 szkół. Każdy rekord zawiera informacje o innym uczniu, tak więc nie ma możliwości powtórzeń. Kolumna ‘Typ’ zawierała informację o rodzaju szkoły, dla szkoły publicznej została ustawiona wartość 1, a dla prywatnej 0. Płeć została zamieniona na 1 dla chłopców oraz 0 dla dziewczynek. Dla kolumny ‘Obiad’ określiłem 1 dla kwalifikujących się oraz 0 dla nie. Lokalizacja zawierała trzy wartości, ‘miejska’, ‘podmiejska’, ‘wiejska’, tak więc postanowiłem rozdzielić tą kolumnę na trzy, tak aby pojedyncza kolumna trzymała informację czy dany uczeń uczęszcza do szkoły w danej lokalizacji. Liczba uczniów mieściła się pomiędzy 14 a 31. Wartości ‘Pre\_test’ mieściły się w przedziale od 22 do 93.



*Macierz korelacji*

Z macierzy korelacji możemy odczytać, iż największy wpływ na zmienną celu ma wpływ ‘Pre\_test’. Praktyczny brak korelacji można zaobserwować dla zmiennej Płeć. Kolejnym ważnym spostrzeżeniem jest iż, wszystkie zmienne z wyjątkiem ‘Metody’ są praktycznie w tym samym stopniu skorelowane z zmienną ‘Pre\_test’ oraz ‘Post\_test’, tak więc mogą one mieć największy wpływ podczas procesu regresji.



*Macierz wykresów rozrzutów*

Z macierzy rozrzutu obserwujemy iż wyłącznie ‘Pre\_test’ układa się w miarę liniowo, tak więc prosta regresja liniowa nie jest zalecana do zastosowania.

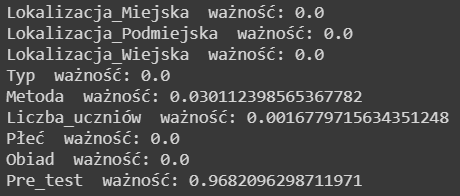
W celu przewidzenia wyników postanowiłem wybrać **sieć neuronową** oraz **drzewo CART**.

**Drzewo CART z wszystkimi predyktorami**

Parametry:

* Criterion = mse
* Max-depth = 5
* Min\_samples\_split = 10
* Min\_samples\_leaf = 40
* Random\_state = 291951

Początkowym podejściem było wykorzystanie wszystkich zmiennych. Otrzymałem drzewo, które brało w 3% ‘metoda’ oraz w niecałych 97% wartość ‘pre\_test’.



*Ważność predyktorów dla CART dla wszystkich predyktorów*

Na podstawie macierzy korelacji oraz wstępnego modelu drzewa CART postanowiłem w całym procesie użyć wyłącznie wartości z kolumn ‘metoda’ oraz ‘pre\_test’.

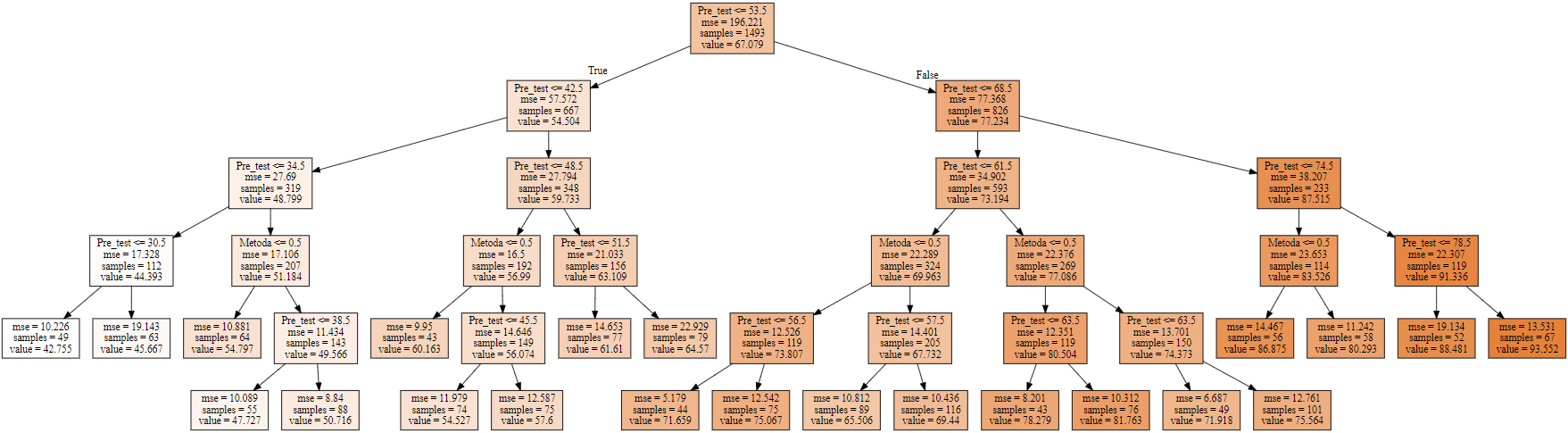
**Drzewo CART**

Parametry:

* Criterion = mse
* Max-depth = 5
* Min\_samples\_split = 10
* Min\_samples\_leaf = 40
* Random\_state = 291951

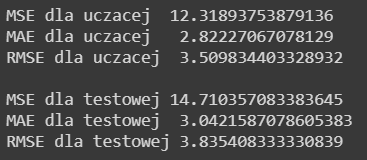
****

*Ważność predyktorów dla CART*

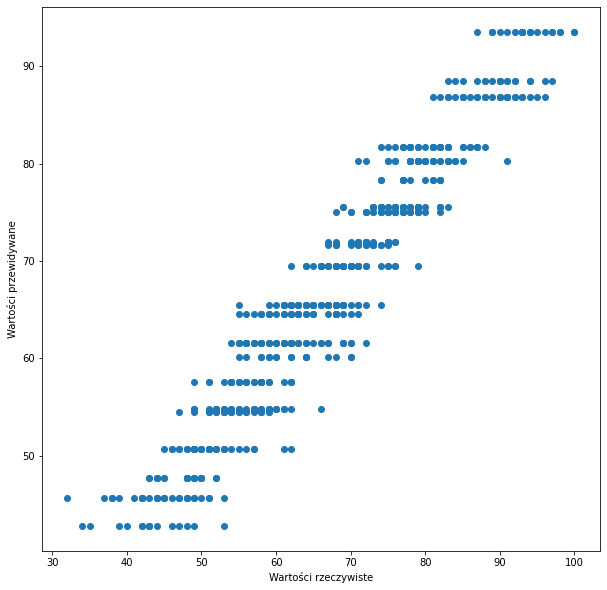
****

*Drzewo CART*

Największy wpływ miał predyktor ‘pre\_test’ wokół którego odbywało się najwięcej pytań. Z lewej strony znajdowały się osoby posiadające mniej niż 54 punkty. Na drugim oraz trzecim poziomie nadal występowało pytanie o tą samą zmienną. W czwartym poziomie większość pytań bazowała o ‘metodę’. W ostatnim poziomie ponownie stawiano pytanie wyłącznie o wartość ‘pre\_test’.

**

*Jakość modelu - CART*



*Zestawienie wartości rzeczywistych z przewidywanymi dla CART*

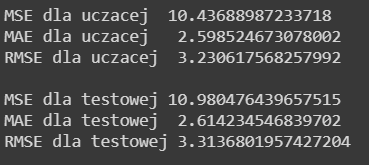
Na podstawie analizy wyników otrzymanych podczas liczenia błędów ‘MAE’ oraz ‘MSE’ dla próby treningowe i testowej możemy stwierdzić iż model nie uległ przeuczeniu, a na zestawieniu otrzymano „chmurkę”, tak więc możemy stwierdzić iż model przewiduje poprawnie.

**Sieć neuronowa**

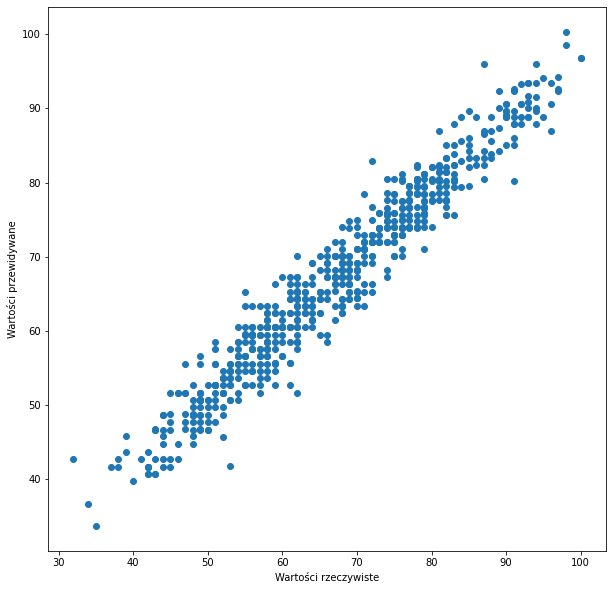
Parametry:

* Hidden\_layer\_size = (10,)
* Activation = ‘tanh’
* Solver = ‘lbfgs’
* Alpha = 0.0001
* Max\_iter = 10000
* Random\_state = 291951

Tak samo jak w drzewie CART, wybrałem predyktory ‘**metoda**’ oraz ‘**pre\_test**’ do przewidzenia wyniku końcowego. Początkowo dane ilościowe zostały znormalizowane do przedziału od 0 do 1, zgodnie z założeniami sieci neuronowej. Metodą prób i błędów, określanie większej ilości warstw ukrytych lub też zwiększanie ilości neuronów w pierwszej warstwie nie wpływa znacząco na wynik.



*Jakość modelu – sieć neuronowa*



*Zestawienie wartości rzeczywistych z przewidywanymi dla sieci neuronowej*

Na podstawie analizy wyników otrzymanych podczas liczenia błędów ‘MAE’ oraz ‘MSE’ dla próby treningowe i testowej możemy stwierdzić iż model nie uległ przeuczeniu, a na zestawieniu otrzymano „chmurkę”, tak więc możemy stwierdzić iż model przewiduje poprawnie.

Podsumowując, obydwa modele przewidują poprawnie. Dokładniejszy wynik otrzymamy z sieci neuronowej. Spowodowane jest to wyliczaniem wartości na podstawie danych wejściowych, w przeciwieństwie do drzewa CART gdzie następuje grupowanie próbek ze względu na spełnianie odpowiednich kryteriów.