Raport 4A

Dawid Sikorski 291951

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

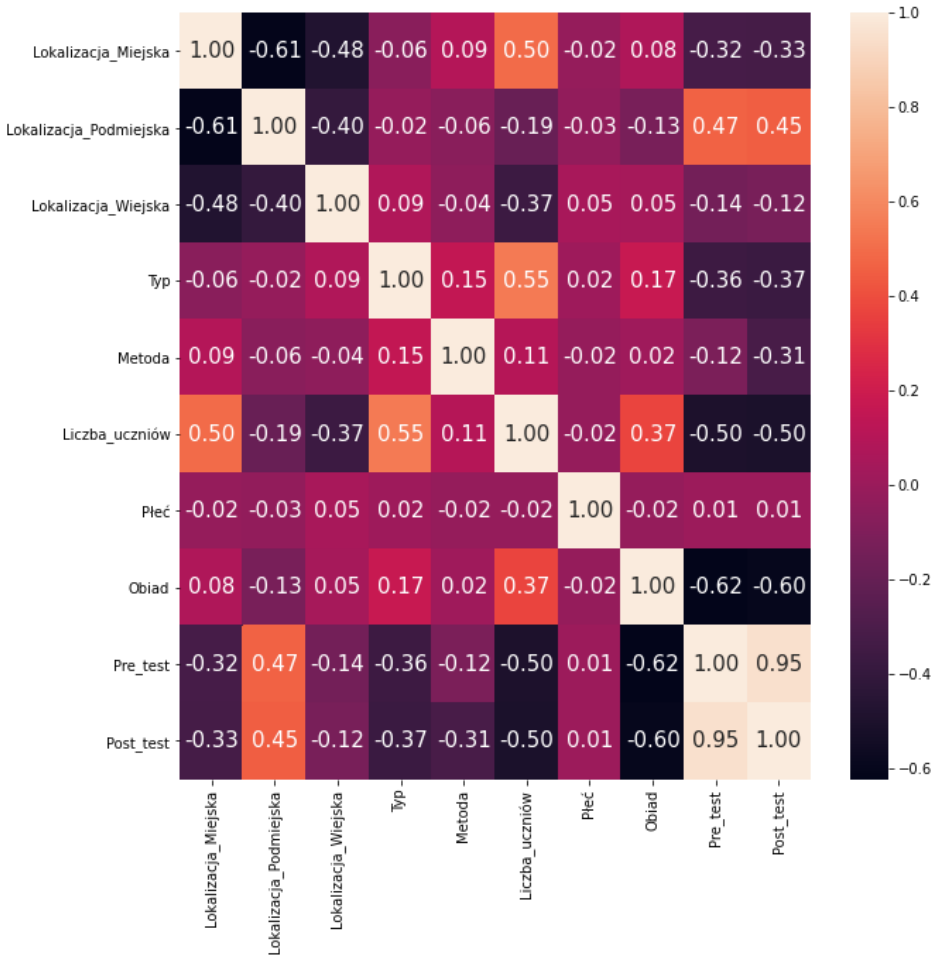
Celem zadania, było przewidzenie ilości punktów z testu dla amerykańskich uczniów z szkół podstawowych. Do realizacji zadania wykorzystano dane zebrane w pliku ‘dane_testy.cvs’, w którym znajduje się 2133 obserwacji z trzema zmiennymi identyfikującymi, pięcioma jakościowymi, oraz trzema ilościowymi w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

```
Szkoła ['ANKYI' 'CCAAW' 'CIMBB' 'CUQAM' 'DNQDD' 'FBUMG' 'GJJHK' 'GOKXL' 'GOOBU'
 'IDGFP' 'KFZMY' 'KZKKE' 'LAYPA' 'OJOBU' 'QOQTS' 'UAGPU' 'UKPGS' 'UUUQX'
 'VHDHF' 'VKWQH' 'VVTVA' 'ZMNYA' 'ZOWMK'] 23
Klasa ['6OL' 'ZNS' '2B1' 'EPS' 'IQN' 'PGK' 'UHU' 'UWK' 'A33' 'EID' 'HUJ' 'PC6'
 '1Q1' 'BFY' 'OMI' 'X6Z' '2AP' 'PW5' 'ROP' 'ST7' 'XXJ' '197' '5LQ' 'JGD'
 'HCB' 'NOR' 'X78' 'YUC' 'ZDT' 'ENO' 'TSA' 'VA6' '18K' 'CXC' 'HKF' 'PBA'
 'U6J' 'W8A' '05H' '98D' 'G2L' 'P2A' 'XZM' '1VD' '21Q' '2BR' '3D0' '5JK'
 'O6A' 'QTU' 'AJ1' 'J8J' 'RA5' '5SZ' '6U9' 'FS3' 'XJ8' '0N7' '3XJ' 'RK7'
 'SUR' 'X2O' 'XZ4' '1SZ' '62L' 'NWZ' 'S98' '08N' '9AW' 'IPU' 'KXB' 'PGH'
 'XXE' '6C1' 'AE1' 'H7S' 'P8I' 'SSP' 'CD8' 'J6X' 'KR1' '341' 'D33' 'DFQ'
 'GYM' 'IEM' '7BL' 'A93' 'TB5' 'YTB' '1UU' '4NN' 'V77' 'CII' 'Q0E' 'QA2'
 'ZBH'] 97
Uczeń ['2FHT3' '3JIVH' '3XOWE' ... 'YDR1Z' 'YUEIH' 'ZVCQ8'] 2133
```

```
Lokalizacja ['miejska' 'podmiejska' 'wiejska']
Typ ['niepubliczna' 'publiczna']
Metoda ['standardowa' 'eksperymentalna']
Liczba_uczniów [20 21 18 15 16 19 17 28 27 24 14 22 23 31 25 26 29 30]
Płeć ['dziewczynka' 'chłopiec']
Obiad ['nie kwalifikuje się' 'kwalifikuje się']
Pre_test [62 66 64 61 63 60 67 57 56 58 54 59 65 55 68 73 70 74 76 69 75 78 72 71
 49 53 48 52 50 46 44 51 47 43 37 40 39 41 38 45 36 42 31 35 33 27 30 34
 32 29 28 23 26 77 79 82 80 85 83 84 86 89 93 88 81 87 91 22 25]
```

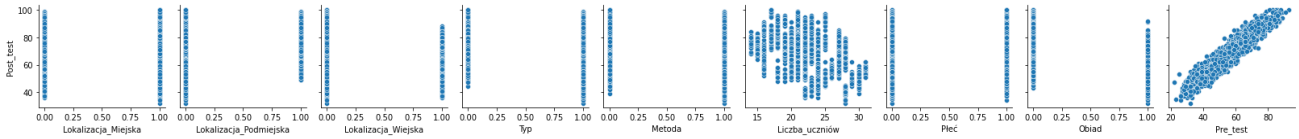
Analiza danych z pliku

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’.

```
Lokalizacja_Miejska  ważność: 0.0
Lokalizacja_Podmiejska  ważność: 0.0
Lokalizacja_Wiejska  ważność: 0.0
Typ  ważność: 0.0
Metoda  ważność: 0.030112398565367782
Liczba_uczniów  ważność: 0.0016779715634351248
Płeć  ważność: 0.0
Obiad  ważność: 0.0
Pre_test  ważność: 0.9682096298711971
```

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’.

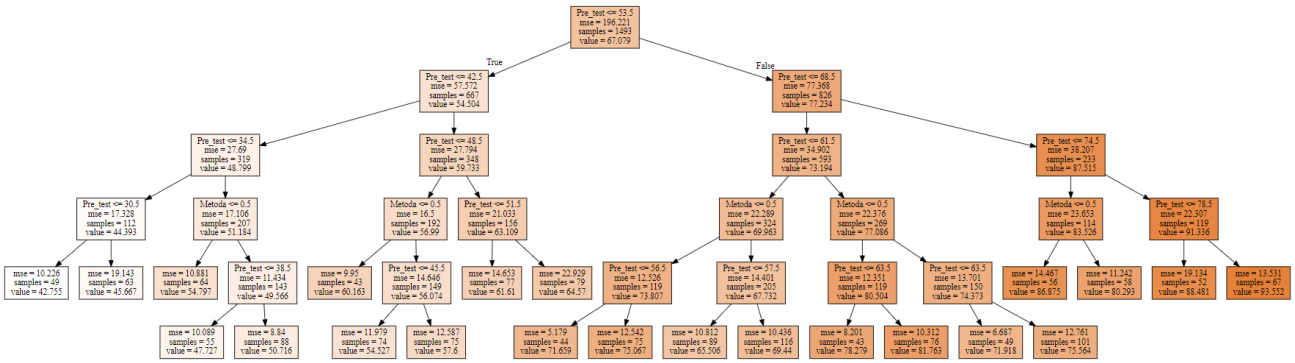
Drzewo CART

Parametry:

- Criterion = mse
- Max-depth = 5
- Min_samples_split = 10
- Min_samples_leaf = 40
- Random_state = 291951

```
Metoda  ważność: 0.03013734285083393
Pre_test  ważność: 0.9698626571491661
```

Ważność predyktorów dla CART



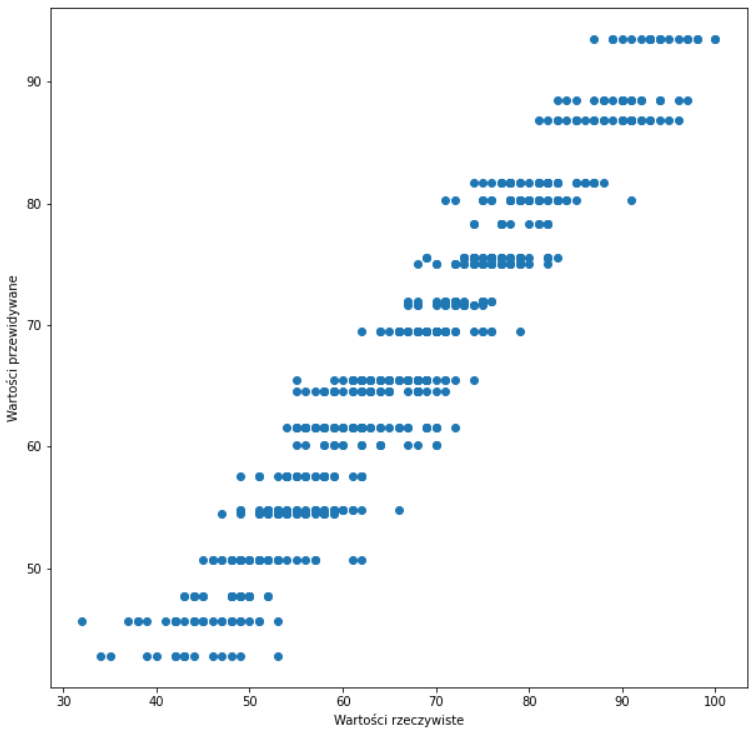
Drzewo CART

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’.

```
MSE dla uczacej  12.31893753879136
MAE dla uczacej   2.82227067078129
RMSE dla uczacej  3.509834403328932

MSE dla testowej 14.710357083383645
MAE dla testowej  3.0421587078605383
RMSE dla testowej 3.835408333330839
```

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:

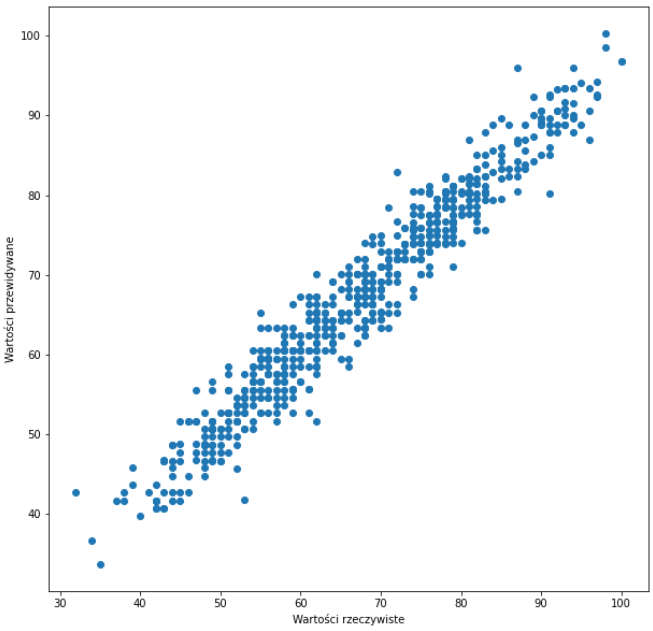
- Hidden_layer_size = (10,)
- Activation = ‘tanh’
- Solver = ‘lbfgs’
- Alpha = 0.0001
- Max_iter = 10000
- Random_state = 291951

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.

```
MSE dla uczacej  10.43688987233718
MAE dla uczacej   2.598524673078002
RMSE dla uczacej  3.230617568257992

MSE dla testowej 10.980476439657515
MAE dla testowej  2.614234546839702
RMSE dla testowej 3.3136801957427204
```

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.