

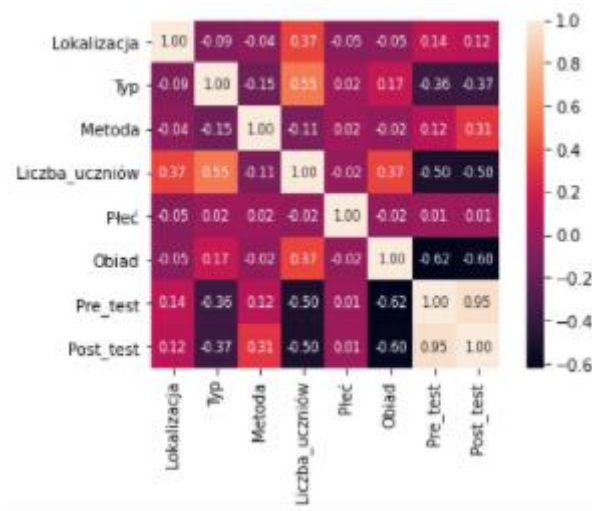
## Raport 4A

Sebastian Krzosek 300136

Celem naszego zadania było zbudowanie modelu szacowania wyniku testu końcowego. Do realizacji tego zadania wykorzystałem algorytmy MLP i Cart.

### Prezentacja rozwiązania:

Na samym początku, zgodnie z poleceniem sprawdziłem poprawność danych a następnie zmienne: „Lokalizacja” , „Typ” , „metoda” , „Plec” , „Obiad” zamieniłem na wartości 0/1. Na podstawie powstałych danych wykonałem macierz korelacji, która wskazała mi, że zmienna „Pre\_test” jest silnie skorelowana z naszą zmienną celu „Post\_test”, natomiast uznałem to za niewystarczające i jako predyktory wykorzystałem wszystkie zmienne.



Po tym kroku, mając już wybrany predyktory i zmienną celu dokonałem podziału danych na zbiór uczący i testowy (70/30) podając swój numer indeksu jako ziarno generatora.

### MLP

Na samym początku algorytmu MLP przeprowadziłem standaryzację danych metoda min-max, aby rola ważności predyktorów nie została zniwelowana. Następnie metodą MLPRegressor zbudowałem sieć, nauczyłem ją na próbnie uczącej i na jej podstawie spróbowałem przewidzieć wyniki. Otrzymałem następujące rezultaty:

**RMSE na próbnie uczącej: 3.1859**

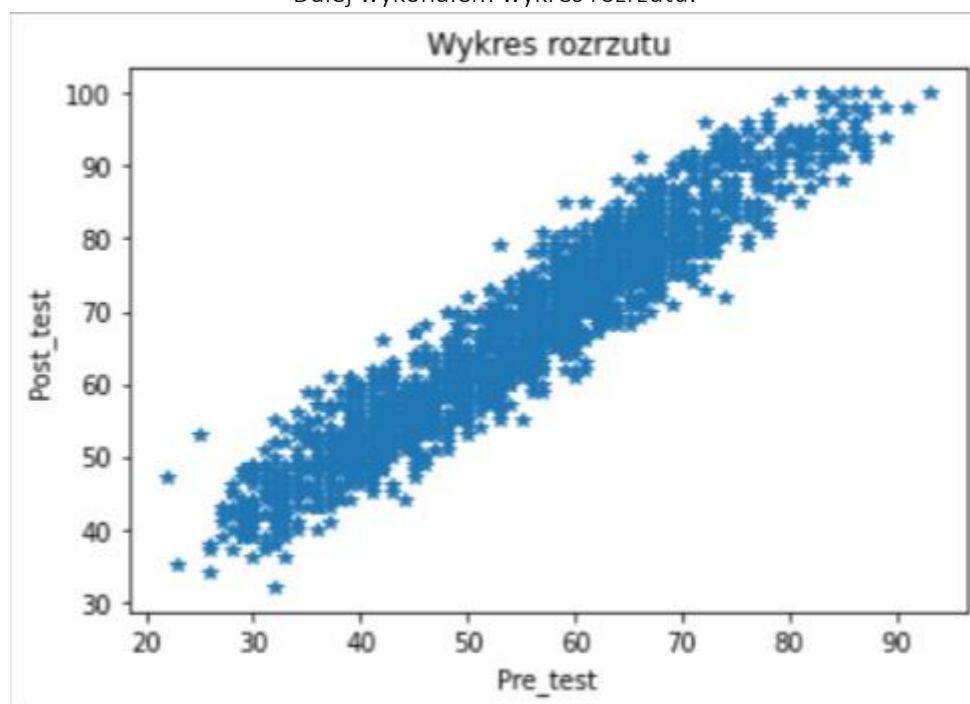
**MAE na próbnie uczącej: 2.5438**

**RMSE na próbnie testowej: 3.7664**

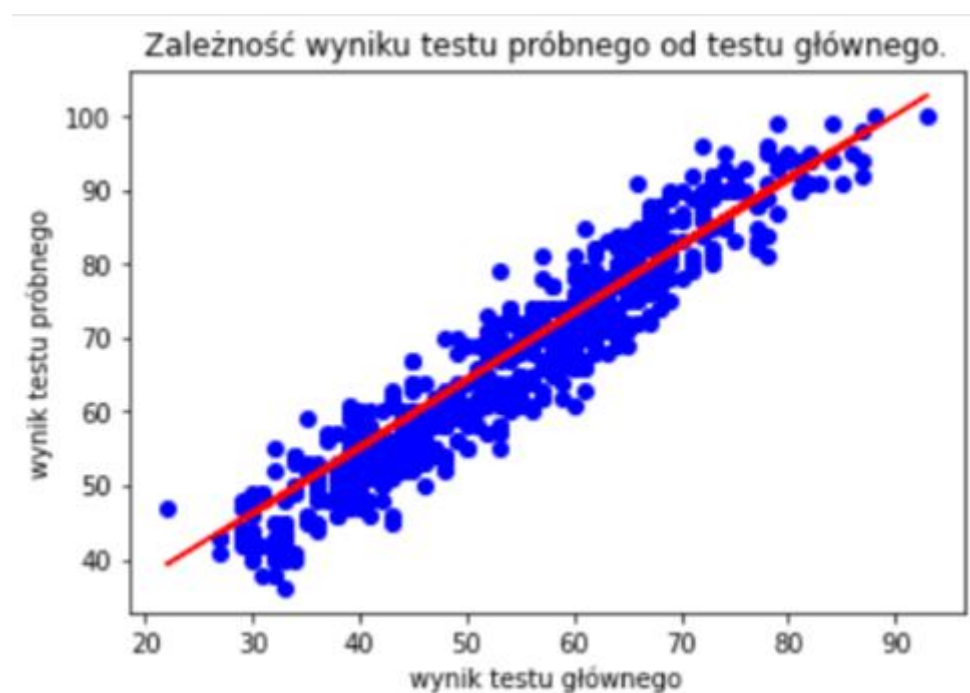
**MAE na próbnie uczącej: 3.0693**

co jest zadowalającym rezultatem, pokazującym że sieć nieuległa przeuczeniu (wartości zarówno na próbnie uczącej jak i testowej są do siebie zbliżone)

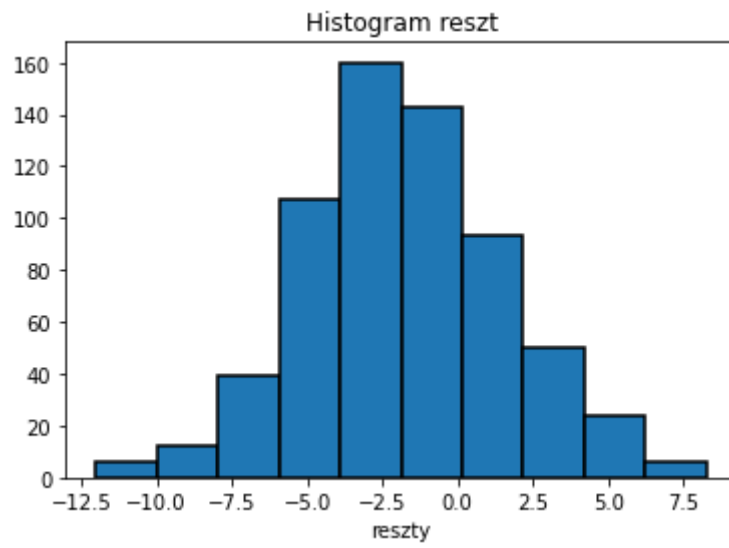
Dalej wykonałem wykres rozrzutu:



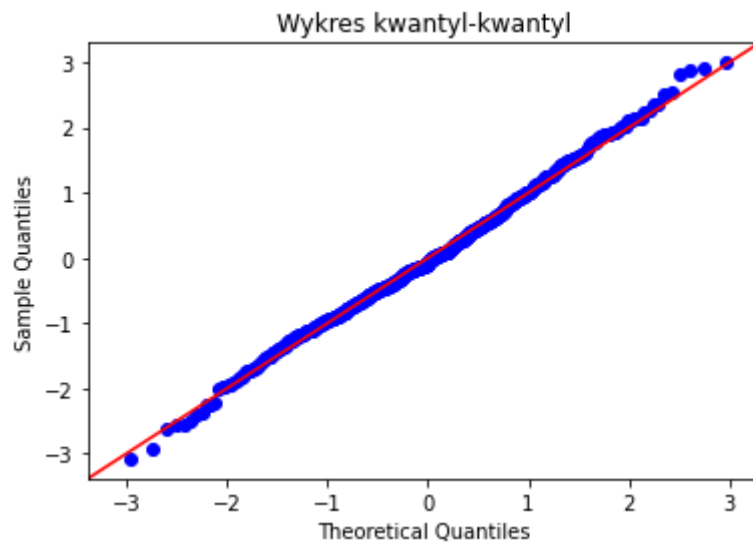
Oraz wykres zależności wyniku testu próbnego od testu głównego:



Histogram dla reszt:

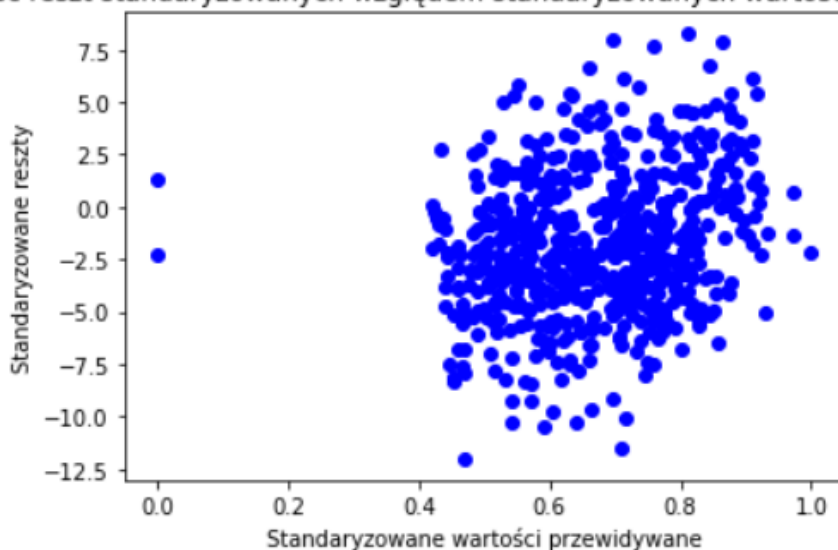


Jak i wykres kwantyl-kwantyl, na którym widzimy że linia układa się bardzo dobrze względem punktów.



Kolejno wykonałem wykres zależności reszt standaryzowanych względem standaryzowanych wartości przewidywanych. Widzimy tutaj, że punkty nie układają się w żadne dziwne kształty, nie formują „lejków” tylko składają się w jedną chmurkę poza pojedynczymi obserwacjami odstającymi.

Zależność reszt standaryzowanych względem standaryzowanych wartości przewidywanych



Na sam koniec wykonałem testy Shapiro-Wilka oraz Dubrina-Watsona, które okazały następujące wyniki:

#### Test Shapiro-Wilka:

p-value : 0.1818 – wskazała nam, że nie ma powodów do odrzucania  $H_0$  mówiącej o tym, że rozkład reszt jest normalny, natomiast w teście **Dubrina-Watsona** otrzymaliśmy wartość odległą od 2 (1.5898) co stwierdza że reszty nie mają rozkładu normalnego. Jest to pierwszy problem, na który się natknąłem. Przez to nasz model przewidywania jest mniej wiarygodny.

#### Cart

W tym wypadku korzystając z wcześniej przygotowanego podziału na zbiór uczący oraz testowy, przy pomocy metody `DecisionTreeRegressor` stworzyłem drzewo CART (drzewo zostało załączone w osobnym pliku graficznym pod nazwa `cart_tree`), na którym to następnie przeprowadziłem predykcję. Otrzymałem następujące wyniki:

**RMSE** na próbie uczącej: **3.2135**

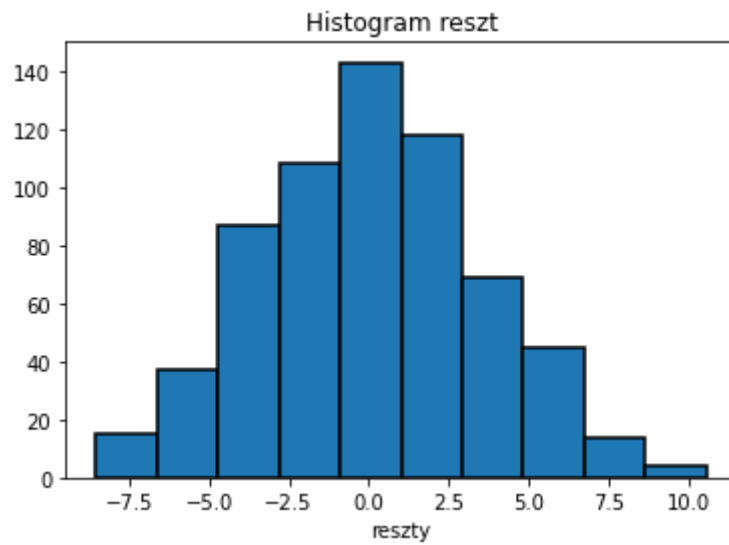
**MAE** na próbie uczącej: **2.5431**

**RMSE** na próbie testowej: **3.4535**

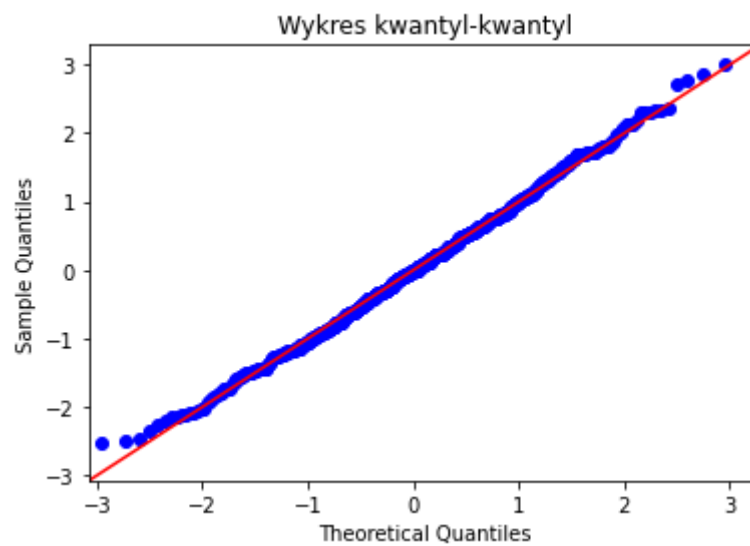
**MAE** na próbie uczącej: **2.7723**

co również jest zadowalającym rezultatem, pokazującym że nie doszło tu do przeuczenia (wartości zarówno na próbie uczącej jak i testowej są do siebie zbliżone)

Idąc dalej wykonałem histogram dla reszt:

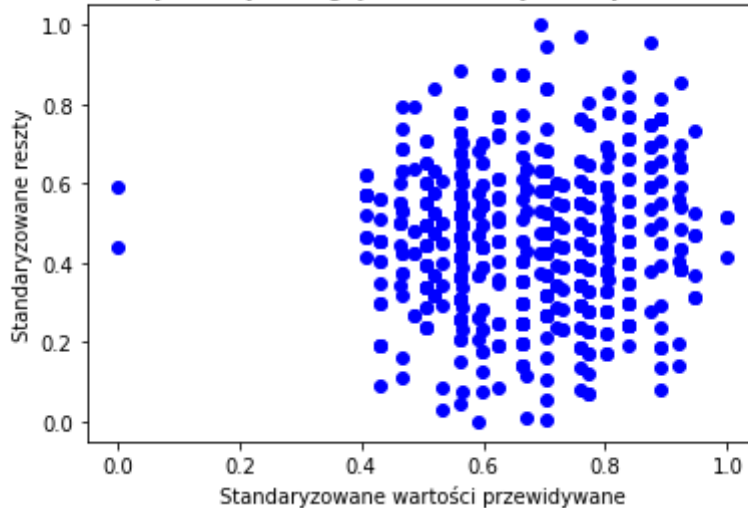


Oraz wykres kwantyl-kwantyl, który bardzo dobrze dopasowuje się do punktów.



Następnie wykonałem wykres zależności reszt standaryzowanych względem standaryzowanych wartości przewidywanych. Jak przy poprzednim algorytmie punkty nie układają się w żadne dziwne kształty, nie formują „lejków” tylko składają się w jedną chmurkę. Jedynymi potencjalnymi problemami ponownie są dwie obserwacje odstające.

Zależność reszt standaryzowanych względem standaryzowanych wartości przewidywanych.



ostatnim krokiem jaki wykonałem było przeprowadzenie testów Shapiro-Wilka oraz Dubrina-Watsona, które zwróciły następujące wyniki:

#### **Test Shapiro-Wilka:**

p-value : 0.2702 – na jej podstawie nie mamy powodów do odrzucenia  $H_0$ , czyli stwierdzenia że występuje rozkład normalny.

**Test Dubrina-Watsona** zwrócił wartość również bardzo bliską 2 (1.9863) co również świadczy o tym, że mamy do czynienia z rozkładem normalnym.

#### **Podsumowanie:**

W obu przypadkach modele działają dobrze, zwracają bardzo podobne RMSE rzędu (około 3-4) oraz również podobne MAE (około 2-3). Problemem MLP jest to, że nie przechodzi testu Dubrina-Watsona, co obniża jego wiarygodność. Dlatego też uważam, że CART w tym zadaniu sprawdził się lepiej, jednakże praca z MLP wydała się łatwiejsza i przyjemniejsza, być może ze względu na wcześniejszą znajomość i subiektywnie większą prostotę przy budowie oraz czytaniu modelu.