Model przewiduje zmienną score, na podstawie danych "CollegeDistance (1).csv".
W modelu został użyty algorytm XGBoost.

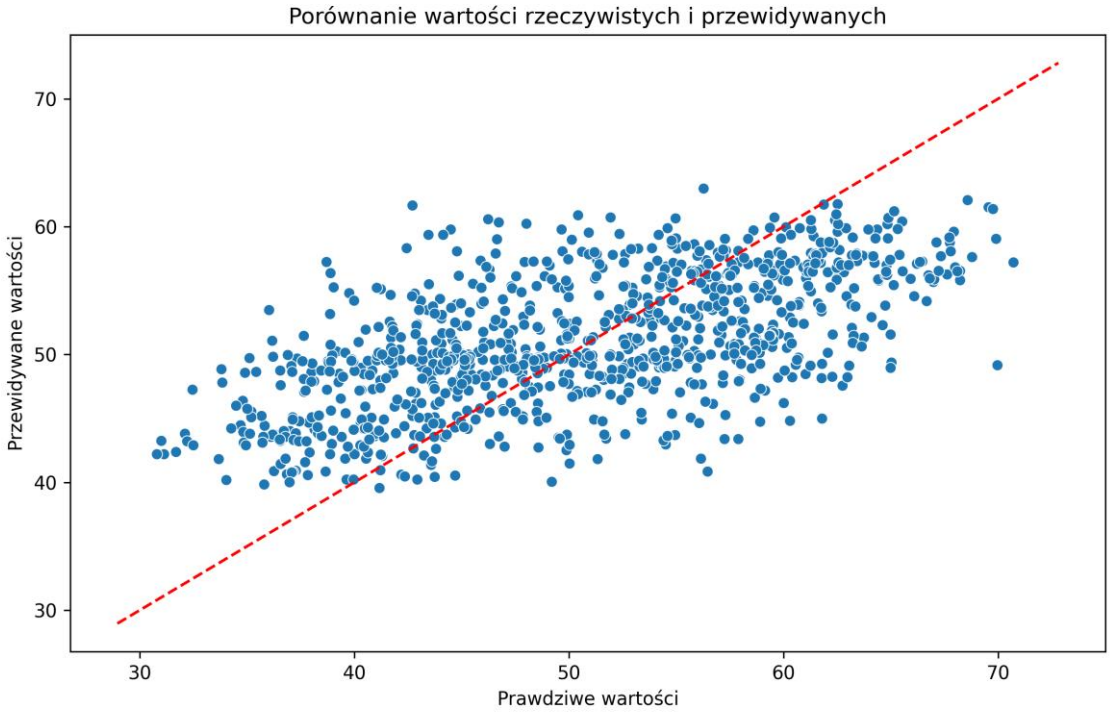
ETAPY

1. Eksploracja i wstępna analiza danych
Dane zostały dostarczone z pliku "CollegeDistance (1).csv", zawartego w folderze z zadaniem.
Dane zostały wczytane przy pomocy biblioteki Pandas.
Została przeprowadzona analiza brakujących wartości w kolumnach i ich ilość została wyświetglona na ekranie.
Wiersze, gdzie liczba brakujących informacji przekracza 30%, zostały usunięte.
Jeśli ilość brakujących informacji była mniejsza niż 30% zostały uzupełnione najczęściej występującą wartoscią z danej kolumny.

2. Inżynieria cech i przygotowanie danych
Kolumny typpu object zostały zakodowane na wartości typu numerycznego, dzięki użyciu one-hot encoding.
Nastąpił podział danych na zbiór testowy oraz treningowy.
Dane zostały znormalizowane, dzięki StandardScaler, co przyczyniło się do lepszego dopasowania modelu.

3. Wybór i trenowanie modelu
Został zastosowany model XGBRegressor, który jest algorytmem regresji.
Trenowanie modelu odbyło się na danych treningowych z wykorzystaniem techniki wyszukiwania siatki do
optymalizacji hiperparametrów.

4. Ocena i optymalizacja modelu
Ocena modelu zostałą przeprowadzona na danych testowych, przy użyciu metryk mse, msa, r2.
Została również przeprowadzona walidacja krzyżowa, w celu oceny stabilności modelu.




```
Mean Squared Error: 47.68575825213402
Mean Absolute Error: 5.678076301445941
R² Score: 0.3711688215013354
Cross-validated MSE: 52.76195960297905
```