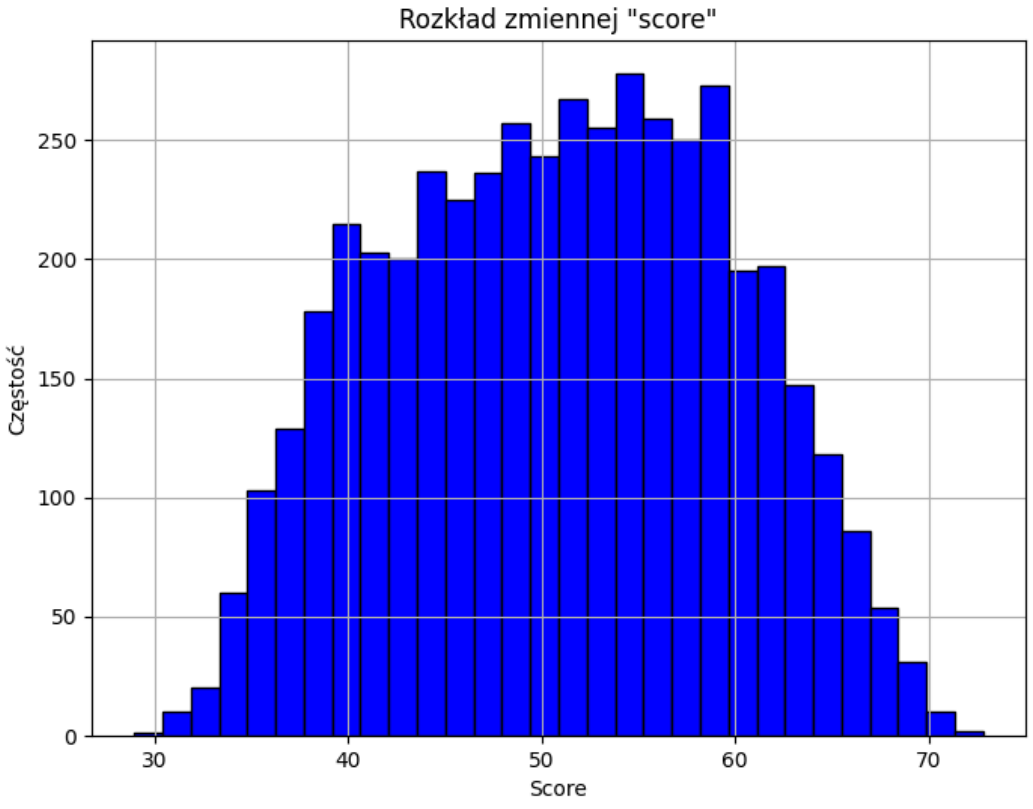
Dokumentacja do Lab03

Cel

Celem projektu jest zbudowanie modelu predykcyjnego, który przewiduje zmienną `score` na podstawie dostarczonego zbioru danych. Proces obejmuje eksplorację danych, inżynierię cech, wybór odpowiedniego modelu i ocenę jego wyników.

Dane

Zbiór danych zawiera 15 kolumn. Zmienna docelowa to `score`, a inne cechy opisują różne aspekty, takie jak `gender`, `ethnicity`, `wage`, `distance`, `tuition` i `education`. Zmienna docelowa `score` ma rozkład stosunkowo równomierny w przedziale 30–70, co można zobaczyć na wykresie poniżej.



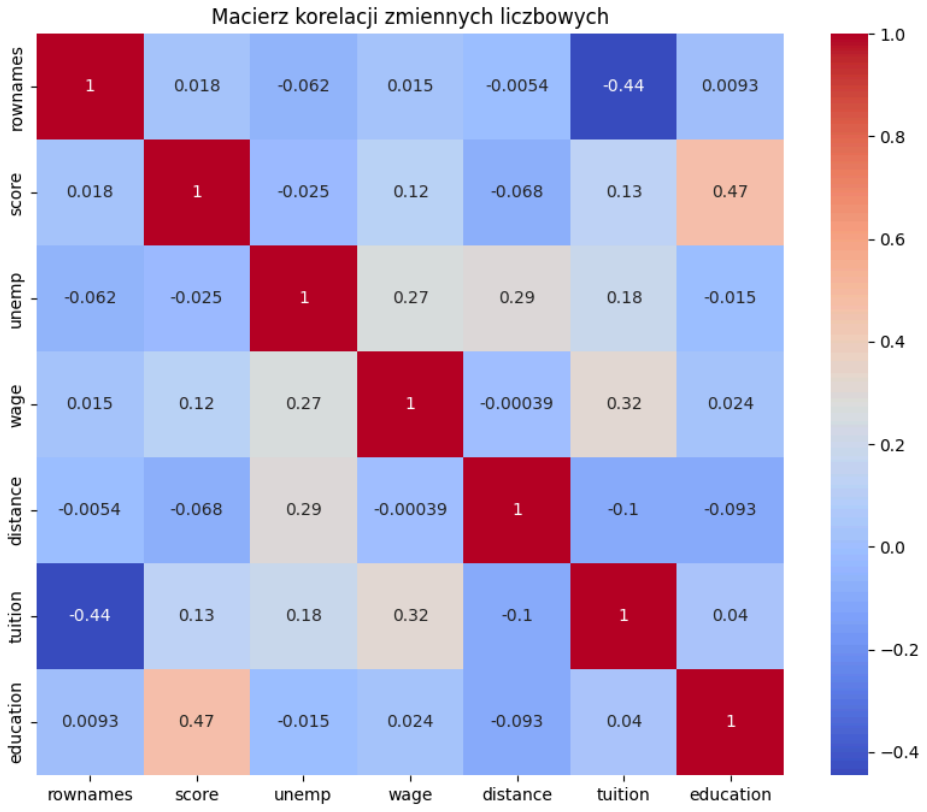
Wykres 1: Rozkład zmiennej `score`

Opis: Histogram powyżej przedstawia rozkład zmiennej `score`. Większość wyników oscyluje w przedziale od 40 do 60, co sugeruje, że wartości zmiennej `score` są rozproszone wokół tego zakresu.

Eksploracja i analiza danych

Macierz korelacji

Aby lepiej zrozumieć zależności między zmiennymi liczbowymi w zbiorze danych, obliczono macierz korelacji, która ilustruje siłę zależności między różnymi zmiennymi.

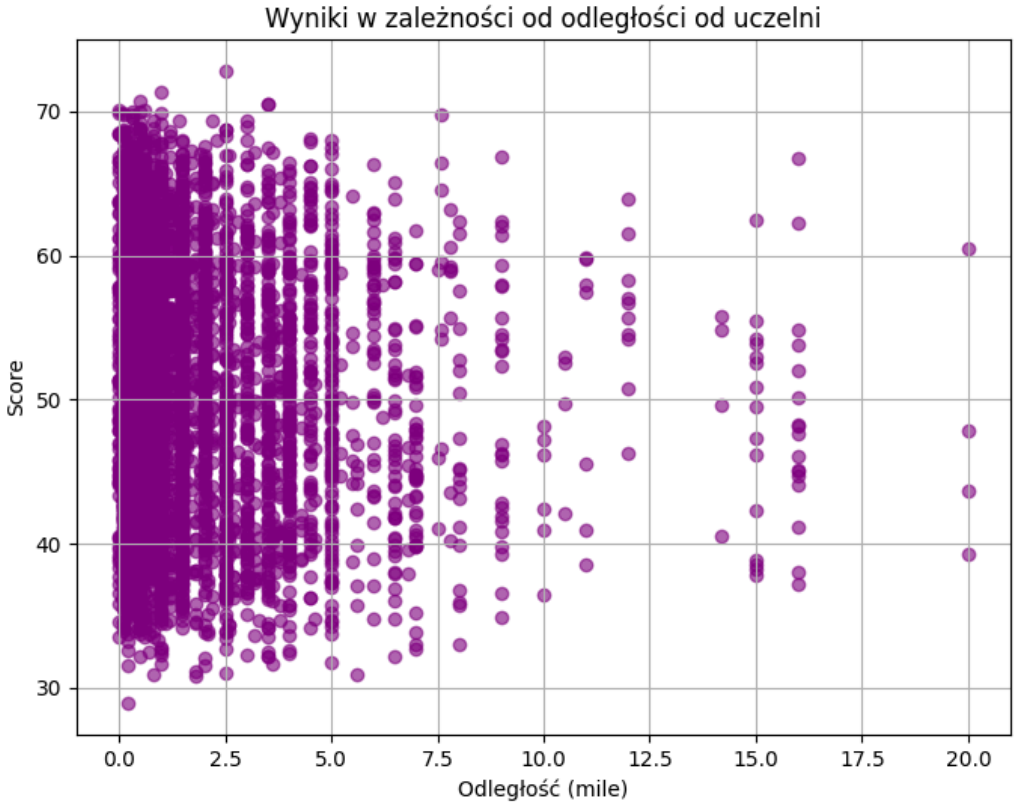


Wykres 2: Macierz korelacji zmiennych liczbowych

Opis: Macierz korelacji wskazuje, które zmienne mają najsilniejszy wpływ na zmienną docelową `score`. Najwyższą korelację z `score` zauważamy w zmiennej `education` (0.47), co sugeruje, że poziom wykształcenia jest istotnym czynnikiem wpływającym na wynik.

Zależność wyników od odległości od uczelni

Jednym z kluczowych aspektów analizy jest sprawdzenie, jak odległość od uczelni wpływa na wyniki uczniów (zmienna `score`).



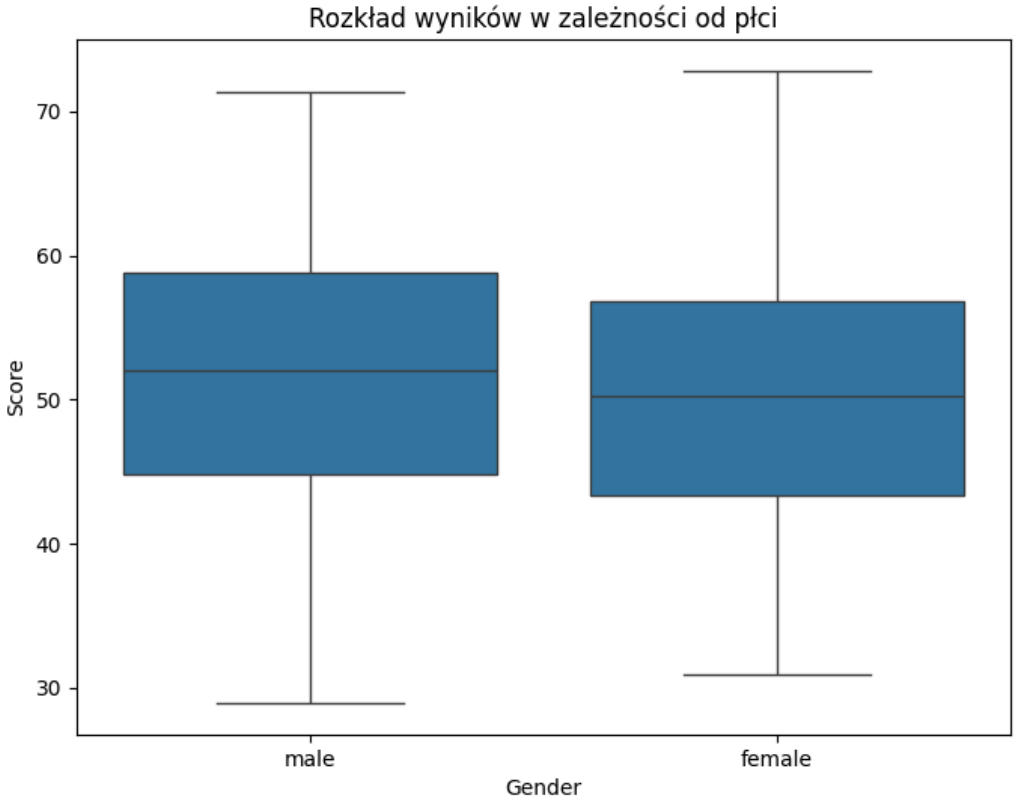
Wykres 3: Wyniki w zależności od odległości od uczelni

Opis: Wykres rozrzutu pokazuje, że wraz ze wzrostem odległości od uczelni, wyniki `score` nie wykazują jednoznacznego trendu. Większość obserwacji skupia się przy małych wartościach odległości (poniżej 5 mil), co może wskazywać na większą liczbę osób mieszkających blisko uczelni.

Analiza w zależności od cech demograficznych

Rozkład wyników w zależności od płci

Poniższy wykres pudełkowy przedstawia, jak wyniki `score` różnią się w zależności od płci uczniów.

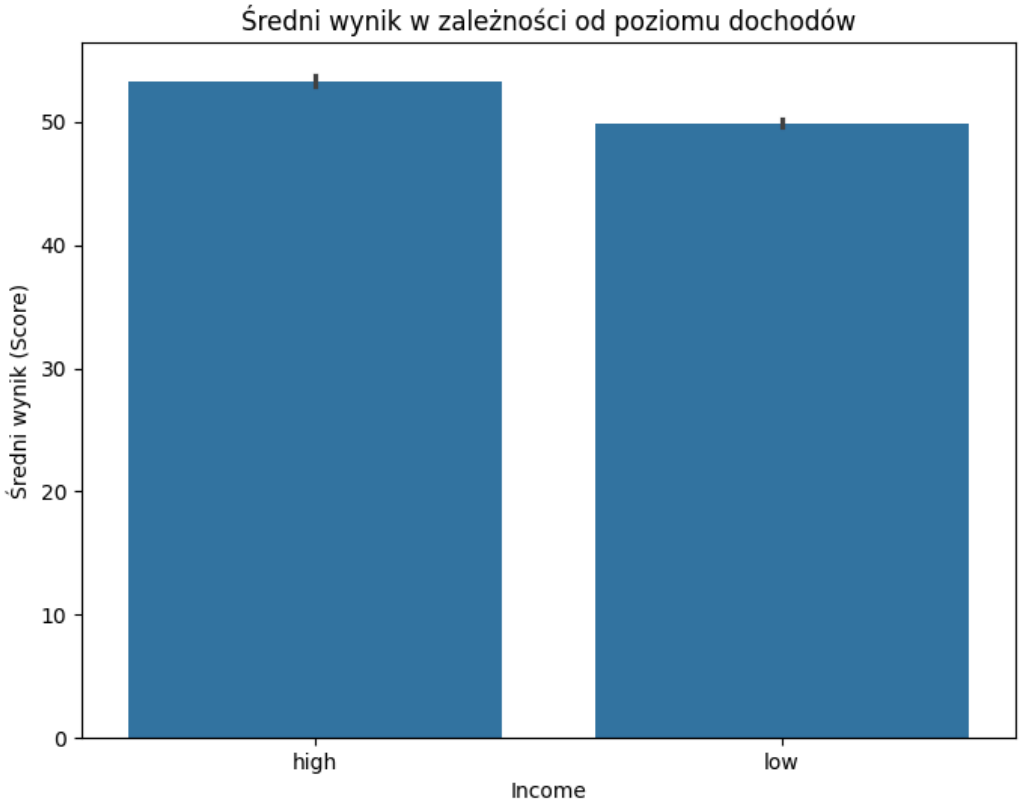


Wykres 4: Rozkład wyników w zależności od płci

Opis: Wyniki `score` są podobnie rozłożone zarówno dla mężczyzn, jak i kobiet. Średnie wyniki dla obu płci są zbliżone, choć mężczyźni wydają się mieć nieco szerszy rozrzut wyników, podczas gdy u kobiet widoczny jest mniejszy rozkład wyników.

Średnie wyniki w zależności od poziomu dochodów

Na poniższym wykresie przedstawiono średnie wyniki `score` dla dwóch grup dochodowych: niskich i wysokich dochodów.



Wykres 5: Średnie wyniki w zależności od poziomu dochodów

Opis: Wyniki `score` są nieco wyższe dla osób z wyższymi dochodami, co może sugerować pewną zależność pomiędzy poziomem dochodów a wynikami uczniów. Różnica jednak nie jest znacząca, co może wskazywać, że inne czynniki (np. edukacja) mogą mieć większy wpływ na wyniki.

Inżynieria cech

Na potrzeby modelu:

1. Zastosowano **standaryzację** dla cech liczbowych, takich jak `unemp`, `wage`, `distance`, `tuition`, i `education`, aby ujednolicić ich zakresy wartości.
2. Zmienione wartości kategoryczne, takie jak `gender` i `ethnicity`, zakodowano przy użyciu **One-Hot Encoding**.

Po przygotowaniu danych podzielono je na zbiór treningowy i testowy (80% treningowy, 20% testowy).

Wybór modelu

Zdecydowaliśmy się na model **Gradient Boosting Regressor**, ponieważ:

- **Boosting** to technika, która łączy kilka słabszych modeli w jeden silny model. W przypadku regresji daje dobre wyniki dla skomplikowanych zależności w danych.
- **Gradient Boosting** automatycznie dobiera odpowiednie parametry modelu i potrafi lepiej zrównoważyć ryzyko przetrenowania (overfittingu) w porównaniu do prostszych modeli.

Porównanie modeli

Model	MAE	MSE	R^2
Regresja Liniowa	5.75	49.11	0.35
Random Forest	**5.68**	**49.19**	**0.35**
Gradient Boosting	**5.58**	**47.9**	**0.36**

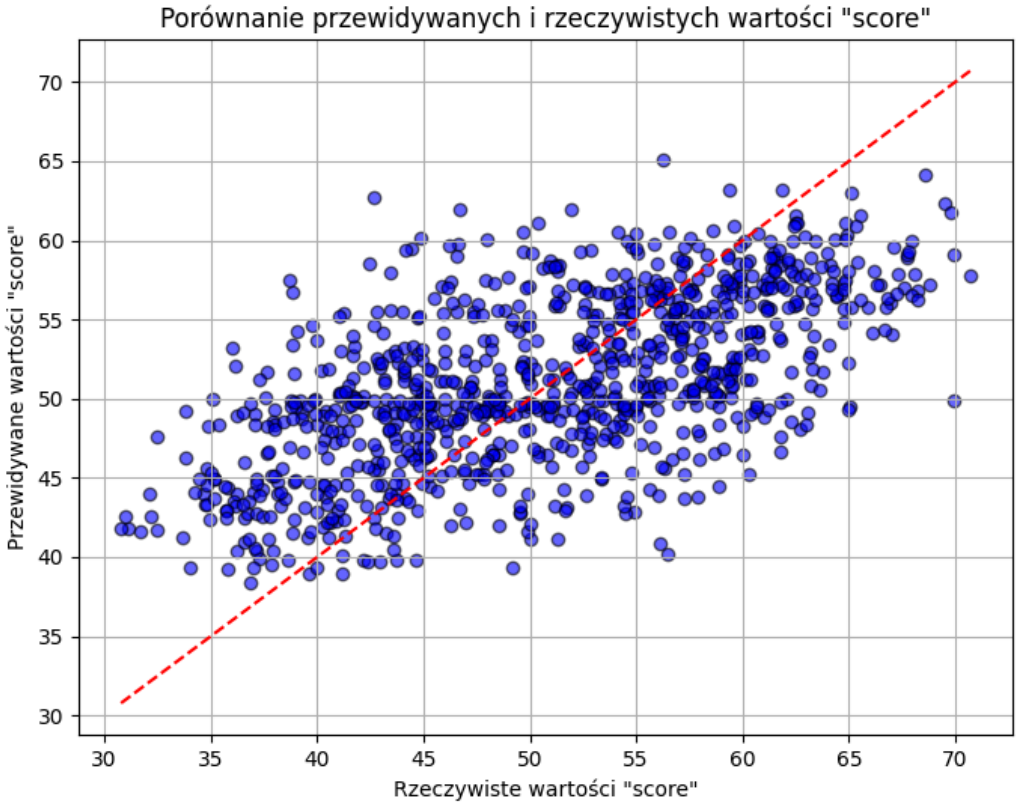
Gradient Boosting osiąga najlepsze wyniki z trzech testowanych modeli pod względem MAE, MSE oraz współczynnika determinacji R^2.

Ocena modelu

Na podstawie zbioru testowego, model **Gradient Boosting** osiągnął następujące wyniki:

- **Mean Absolute Error (MAE):** 5.6857425283098975
- **Mean Squared Error (MSE):** 47.99146435453912
- **R^2 (Współczynnik determinacji):** 0.3671374809985316

Wykres porównania przewidywanych i rzeczywistych wartości `score`



Na wykresie widać, że przewidywane wartości `score` dobrze odpowiadają rzeczywistym danym, choć w niektórych miejscach pojawiają się pewne odchylenia.

Podsumowanie

Model Gradient Boosting został wybrany ze względu na jego zdolność do skutecznego modelowania złożonych zależności w danych oraz lepsze wyniki w porównaniu do innych modeli, takich jak regresja liniowa czy lasy losowe.

Model osiągnął:

- **MAE:** 5.6857425283098975
- **MSE:** 47.99146435453912
- **R^2:** 0.3671374809985316