Dokumentacja Lab3

Oskar Paciorkowski, s25488

Spis treści

1. Wstęp

Celem projektu jest opracowanie modelu przewidującego wyniki testów w oparciu o czynniki społeczne i ekonomiczne, z wykorzystaniem zbioru danych „*CollegeDistance.csv*”. Zbiór ten zawiera takie zmienne jak: płeć, pochodzenie etniczne, dochody rodziny czy poziom wykształcenia rodziców, które mogą istotnie wpływać na rezultaty osiągane przez uczniów. Analiza ta, ma na celu zbadanie wpływu tych czynników na osiągnięcia edukacyjne oraz stworzenie modelu, który pozwoli precyzyjnie prognozować wyniki testów, dostarczając przy tym cennych informacji o wpływie warunków społeczno-ekonomicznych na postępy uczniów.

2. Zbiór danych

W projekcie wykorzystano zbiór danych *CollegeDistance* z pakietu AER, który został opracowany przez Departament Edukacji. Zawiera on 4739 obserwacji i 15 zmiennych.

Zmienne zawarte w zbiorze obejmują między innymi:

- **gender**: płeć ucznia,
- **ethnicity**: pochodzenie etniczne (Afroamerykanin, Latynos, inne),
- **score**: wynik testu (zmienna przewidywana),
- **fcollege** i **mcollege**: czy ojciec/matka mają ukończone studia,
- **home**: czy rodzina posiada własne mieszkanie,
- **urban**: czy szkoła znajduje się na obszarze miejskim,
- **unemp**: stopa bezrobocia,
- **wage**: średnia płaca godzinowa,
- **distance**: odległość od najbliższego college'u,
- **tuition**: średnie czesne,
- **education**: liczba lat edukacji ucznia,
- **income**: czy dochód rodziny przekracza 25 000 USD rocznie,
- **region**: region zamieszkania ucznia (Zachód lub inne).

3. Przygotowanie danych

Podczas wczytywania danych została usunięta pierwsza kolumna z indeksem rzędu, gdyż nie była ona w żaden sposób skorelowana z danymi i służyła wyłącznie do organizacji danych w pliku CSV.

```
data = pd.read_csv(path)

#Usunięcie pierwszej kolumny
data = data.drop(columns=data.columns[0])
```

Następnie wszystkie zmienne tekstowe zostały przekształcone na wartości liczbowe za pomocą kodowania one-hot, co umożliwia modelowi ich właściwe wykorzystanie podczas uczenia.

```
#Zamiana wartości tekstowych na liczbowe
categorical_columns = data.select_dtypes(include=['object']).columns.tolist()
data_encoded = pd.get_dummies(data, columns=categorical_columns, drop_first=True)
```

Kolejnym krokiem była normalizacja danych, mająca na celu ujednolicenie skali zmiennych oraz eliminację wpływu różnic w jednostkach poszczególnych zmiennych, co mogłoby zaburzać działanie modelu.

```
#Usunięcie score z X, gdyż to go będziemy chcieli przewidywać
X = data_encoded.drop(columns=["score"])

y = data_encoded['score']

#Normalizacja danych
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Na koniec dane zostały podzielone na **zbiór treningowy (80%)** i **testowy (20%)** w celu oceny skuteczności modeli w przewidywaniu wyników testów uczniów na podstawie zmiennych społeczno-ekonomicznych.

```
#Podział 80:20 na dane treningowe i testowe
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

Warto zauważyć, że w zbiorze danych nie wystąpiły żadne braki. Dzięki temu mogliśmy pominąć krok z usuwaniem, bądź uzupełnianiem danych.

4. Wybór i trenowanie danych

Trzy modele regresyjne zostały wybrane i przetestowane w celu przewidywania wyników testów na podstawie cech społeczno-ekonomicznych:

- ***Regresja liniowa***: model podstawowy bez dodatkowych hiperparametrów, pozwalający na uchwycenie ogólnych zależności liniowych.
- ***Las losowy****:* model lasu losowego z optymalizacją hiperparametrów (*n_estimators* i *max_depth*) za pomocą siatki *GridSearchCV,* co umożliwiło wybór najlepszego zestawu parametrów pod kątem współczynnika determinacji R^2.
- ***XGBoost***: model gradientowego wzmocnienia XGBoost, również dostrojony z użyciem GridSearchCV na podstawie *n_estimators*, *learning_rate* oraz *max_depth*.

Po trenowaniu każdego z modeli i ocenie wyników na zbiorze testowym, dla każdego modelu obliczono kluczowe miary jakości, takie jak MSE (Mean Squared Error), MAE (Mean Absolute Error), R^2 oraz procentowy wskaźnik sukcesu modelu. Najlepszy model został wybrany na podstawie najwyższego współczynnika R^2, co pozwoliło na wyłonienie modelu o najwyższej przewidywalności.

W celu zbudowania modelu przewidującego wyniki testów wybrano trzy modele, które różnią się złożonością oraz sposobem radzenia sobie z różnymi typami danych:

1. **Regresja liniowa** – model ten jest stosunkowo prosty, a jego interpretacja jest przejrzysta. Regresja liniowa zakłada liniową relację między zmiennymi niezależnymi a wynikami testów, co pozwala uchwycić podstawowe wzorce w danych.
2. **Las losowy** – model lasu losowego składa się z licznych drzew decyzyjnych, które wspólnie poprawiają przewidywalność modelu poprzez wprowadzenie losowego wyboru próbek i predyktorów przy każdym podziale. Dla lasu losowego zastosowano ***optymalizację hiperparametrów*** za pomocą *GridSearchCV*, aby znaleźć najlepsze wartości dla *n_estimators* (liczby drzew) oraz *max_depth* (maksymalnej głębokości drzewa). Siatka hiperparametrów została przetestowana za pomocą trzykrotnej ***walidacji krzyżowej*** na zbiorze treningowym, maksymalizując współczynnik R^2. Ostateczny model lasu losowego jest bardziej odporny na overfitting niż pojedyncze drzewo decyzyjne i pozwala uchwycić złożone, nieliniowe zależności między zmiennymi.
3. **XGBoost (Extreme Gradient Boosting)** – XGBoost to model, który tworzy kolejne drzewa na podstawie

błędów wcześniejszych predykcji, co umożliwia skuteczne wychwycenie złożonych zależności w danych. Podobnie jak w przypadku lasu losowego, dla XGBoost przeprowadzono optymalizację hiperparametrów (*n_estimators*, *learning_rate* oraz *max_depth*) z użyciem *GridSearchCV*. Dzięki tej metodzie model buduje mocne drzewo na podstawie błędów poprzednich drzew, co pozwala lepiej dopasować model do trudnych i nieliniowych zależności między cechami a wynikami testów.

4.1. Ewaluacja modeli

Każdy z wybranych modeli został przetestowany pod kątem kilku metryk jakości:

- **Mean Squared Error (MSE)**: mierzy średni kwadratowy błąd predykcji, a jego niska wartość oznacza dobre dopasowanie modelu.
- **Mean Absolute Error (MAE)**: mierzy średni absolutny błąd, czyli średnią różnicę między przewidywaną a rzeczywistą wartością.
- **Współczynnik determinacji R^2**: ocenia, jaki procent zmienności wyników testów jest wyjaśniony przez model. Wyższy R^2 świadczy o lepszym dopasowaniu modelu do danych.

Czasy wykonania zostały również zapisane, aby porównać wydajność obliczeniową modeli, co jest szczególnie istotne

w przypadku dużych zbiorów danych i bardziej skomplikowanych algorytmów, takich jak XGBoost.

4.2. Wybór najlepszego modelu

Na podstawie wyników uzyskanych na zbiorze testowym wybrano model o najwyższym współczynniku R^2. Wartość tego współczynnika pokazuje, który z modeli najdokładniej przewiduje wyniki testów, wykorzystując dostępne cechy. Najlepszy model został następnie przeanalizowany pod kątem znaczenia cech, co pozwoliło określić, które zmienne (takie jak: wykształcenie rodziców, dochód rodziny czy region) miały największy wpływ na przewidywany wynik testu.

W rezultacie proces wyboru i trenowania modelu pozwolił na identyfikację najbardziej skutecznego algorytmu, który nie tylko przewiduje wyniki z dużą dokładnością, ale również umożliwia zrozumienie wpływu poszczególnych zmiennych na wyniki uczniów, co jest cenną informacją w analizie społeczno-ekonomicznej.

4.3. Dopasowanie hiperparametrów

Podczas trenowania modeli priorytetem było osiągnięcie jak najkrótszego czasu wykonania przy zachowaniu wysokiej jakości wyników. W początkowej konfiguracji algorytmy XGBoost i Random Forest potrzebowały kilku minut na wykonanie, a ich wyniki wciąż tylko nieznacznie przewyższały te uzyskane przy użyciu regresji liniowej. W

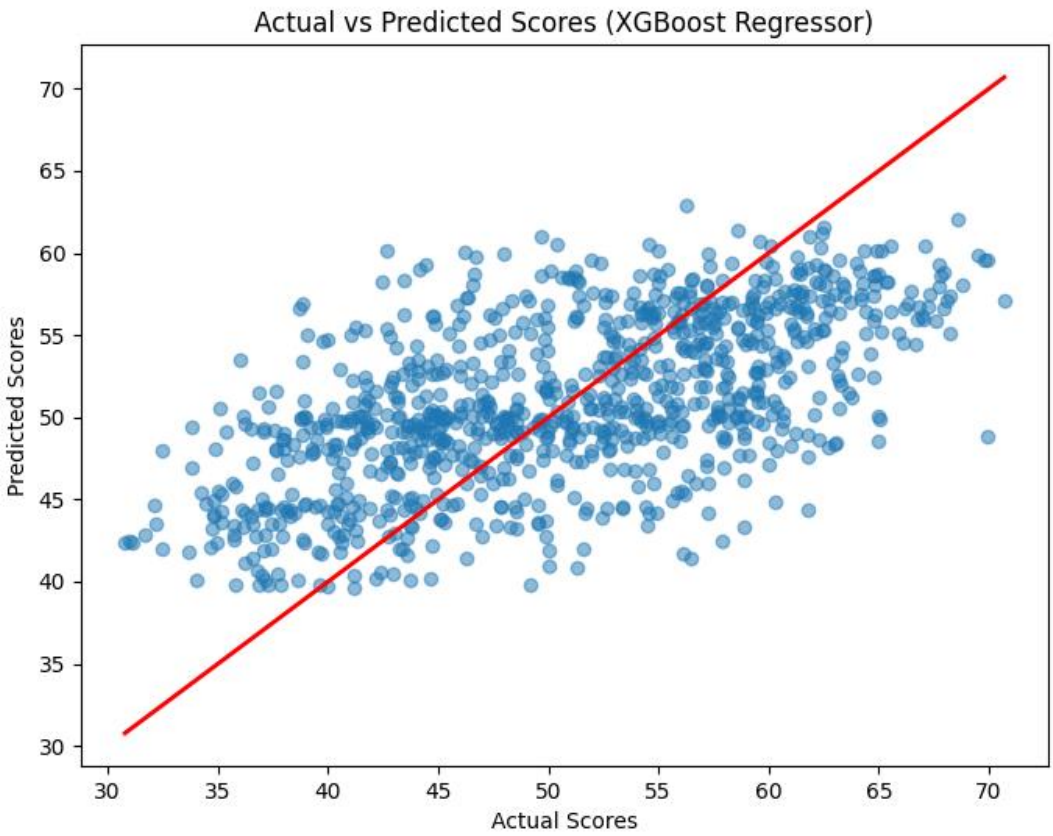
związku z tym zdecydowałem się na dostrojenie parametrów, aby zoptymalizować czas trwania procesów i zminimalizować wykorzystanie limitowanego dostępu do Github Actions.

5. Wnioski

Analizowane modele regresyjne – regresja liniowa, las losowy i XGBoost – różniły się pod względem skuteczności przewidywania wyników testów uczniów. Najważniejsze metryki każdego z modeli przedstawiały się następująco:

- **Regresja liniowa**: Współczynnik determinacji R^2 wyniósł 0,35, co oznacza, że model wyjaśniał jedynie 35% zmienności wyników testów. Jest to model najszybszy w wykonaniu (0,02 sekundy), lecz także najprostszy i najmniej dokładny w tej analizie.
- **Las losowy**: Model lasu losowego również osiągnął współczynnik R^2 w okolicach 0,35, co wskazuje na podobną dokładność w porównaniu z regresją liniową. Las losowy potrzebował jednak znacznie więcej czasu na wykonanie (8,79 sekundy), co sugeruje, że nie jest idealnym wyborem przy tak małej poprawie dokładności.
- **XGBoost**: Model okazał się najlepszy spośród trzech testowanych modeli, osiągając $R^2 = 0,36$. Czas wykonania wyniósł 3,9 sekundy – dłużej niż regresja liniowa, ale znacznie szybciej niż las losowy. Wobec tego modelu uzyskano najlepszą kombinację skuteczności i czasu przetwarzania, co czyni go **najlepszym wyborem**.

5.1. Analiza wykresów

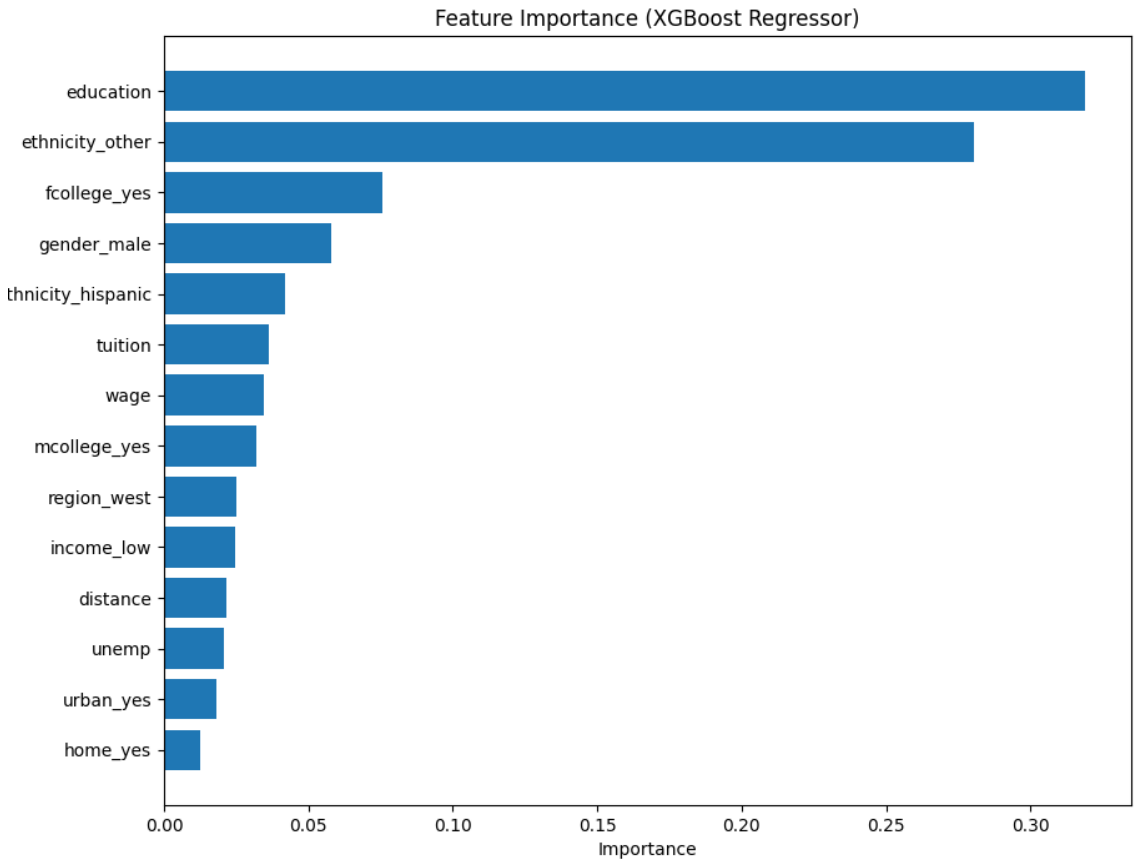


Wykres 1: „Actual vs Predicted Scores (XGBoost Regressor)”

Na wykresie przedstawiono zależność między rzeczywistymi a przewidywanymi wynikami testów przez model XGBoost.

- **Oś X** reprezentuje rzeczywiste wyniki testów, natomiast **oś Y** to wyniki przewidywane przez model.
- Czerwona linia przekątna reprezentuje idealne dopasowanie, gdzie przewidywane wyniki są identyczne z rzeczywistymi.

- Można zauważyć, że większość punktów skupia się w okolicach linni, więc model XGBoost z pewnym sukcesem przewiduje wyniki testów, ale z zauważalnym rozproszeniem wyników. To rozproszenie pokazuje, że model ma ograniczoną dokładność i nie zawsze trafnie przewiduje wartości ekstremalne (bardzo wysokie lub bardzo niskie wyniki).

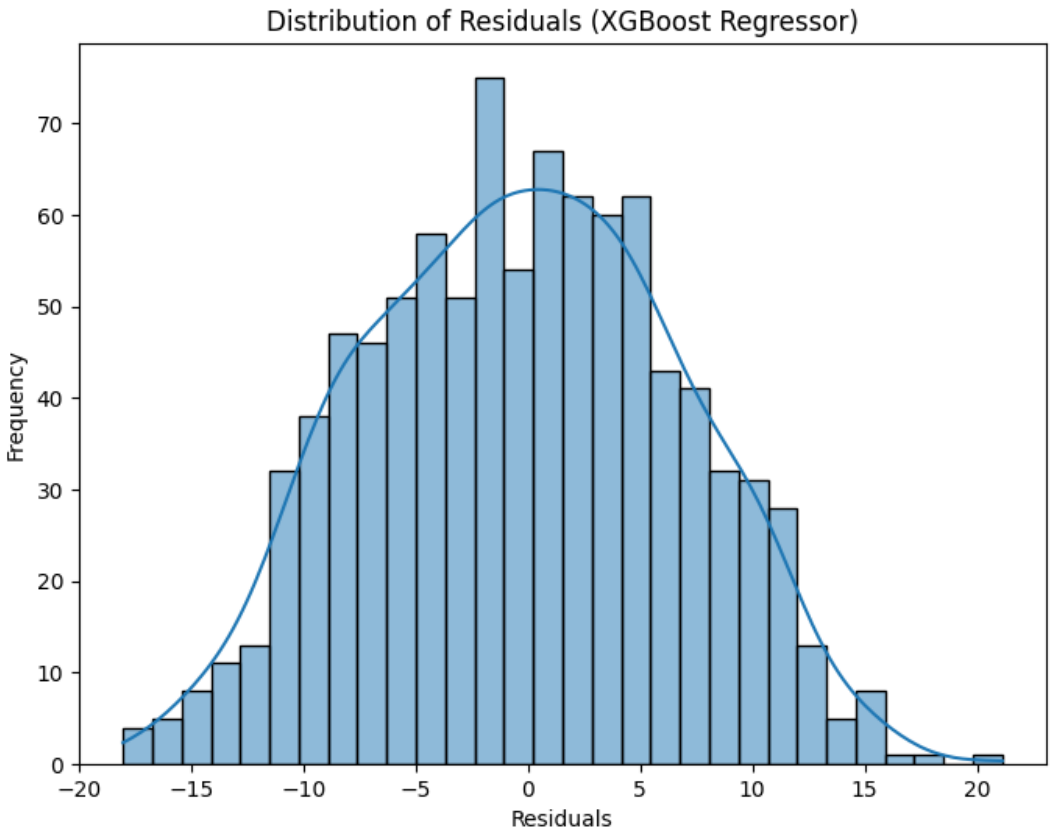


Wykres 2: „Feature Importance (XGBoost Regressor)”

Wykres znaczenia cech przedstawia wpływ poszczególnych zmiennych na przewidywania modelu XGBoost.

- **Oś Y** zawiera listę cech (zmiennych), które miały wpływ na wynik modelu, natomiast **oś X** reprezentuje ich znaczenie w przewidywaniach.
- Najważniejszą zmienną okazała się **education** (liczba lat edukacji), co sugeruje, że poziom wykształcenia ucznia miał najsilniejszy wpływ na wynik testu.
- Kolejną istotną zmienną była **ethnicity_other**, co może wskazywać na znaczenie pochodzenia etnicznego w kontekście wyników edukacyjnych.
- Wpływ miały również takie zmienne, jak **fcollege_yes** (czy ojciec ucznia ukończył studia) oraz **gender_male**, co może sugerować różnice między płciami i wykształceniem rodziców w przewidywaniu wyników testów.

Znaczenie tych zmiennych daje wgląd w czynniki, które najmocniej oddziałują na wyniki testów.

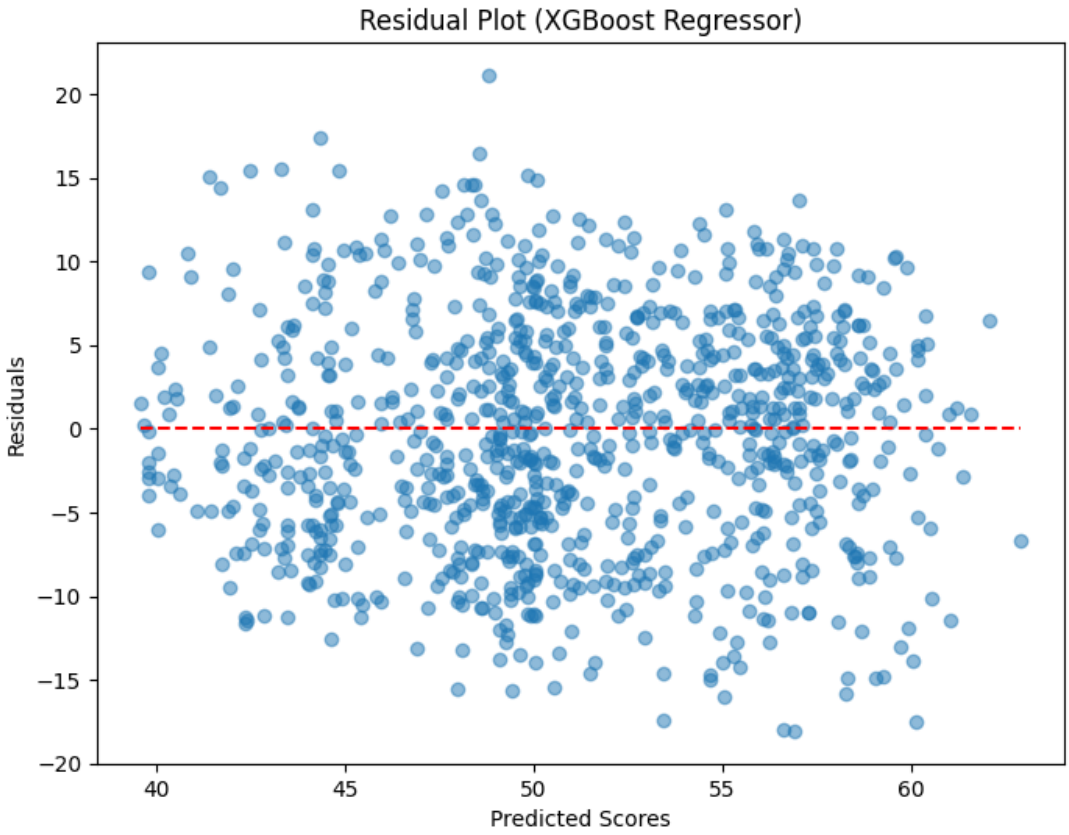


Wykres 3: „Distribution of Residuals (XGBoost Regressor)”

Wykres przedstawia rozkład reszt modelu XGBoost, czyli różnic między rzeczywistymi a przewidywanymi wynikami.

- **Oś X** przedstawia wartości reszt, a **oś Y** – częstotliwość występowania tych reszt.
- Kształt rozkładu reszt przypomina krzywą normalną, co sugeruje, że model popełnia błędy w sposób stosunkowo losowy.
- Większość reszt oscyluje wokół zera, co jest pozytywnym sygnałem, ale występują też pewne odchylenia w obu kierunkach, co świadczy o tym, że

model ma tendencję do lekkiego przeszacowania i niedoszacowania wyników w pewnych przypadkach.



Wykres 4: „Residual Plot (XGBoost Regressor)”

Wykres reszt ilustruje rozkład błędów w zależności od przewidywanych wyników.

- **Oś X** reprezentuje przewidywane wyniki testów, a **oś Y** pokazuje wartość reszt (różnice między rzeczywistymi a przewidywanymi wynikami).
- Czerwona przerywana linia wskazuje na poziom zera, który reprezentuje idealne dopasowanie.

- Rozkład punktów wokół tej linii jest dość równomierny, co oznacza, że model nie ma wyraźnych błędów systematycznych. Widać jednak pewne wzorce rozproszenia, co sugeruje, że model nie w pełni radzi sobie z bardziej skrajnymi wartościami wyników.

5.2. Podsumowanie

Model XGBoost, mimo iż nie osiągnął bardzo wysokiej wartości R^2, okazał się najskuteczniejszy spośród analizowanych modeli. Wartości reszt są rozproszone w sposób zbliżony do normalnego, co wskazuje na losowe błędy, jednak model wykazuje pewne ograniczenia w przewidywaniu wartości ekstremalnych. Wyniki te wskazują, że istnieją istotne czynniki, takie jak poziom edukacji, pochodzenie etniczne i wykształcenie rodziców, które mają znaczący wpływ na wyniki testów.