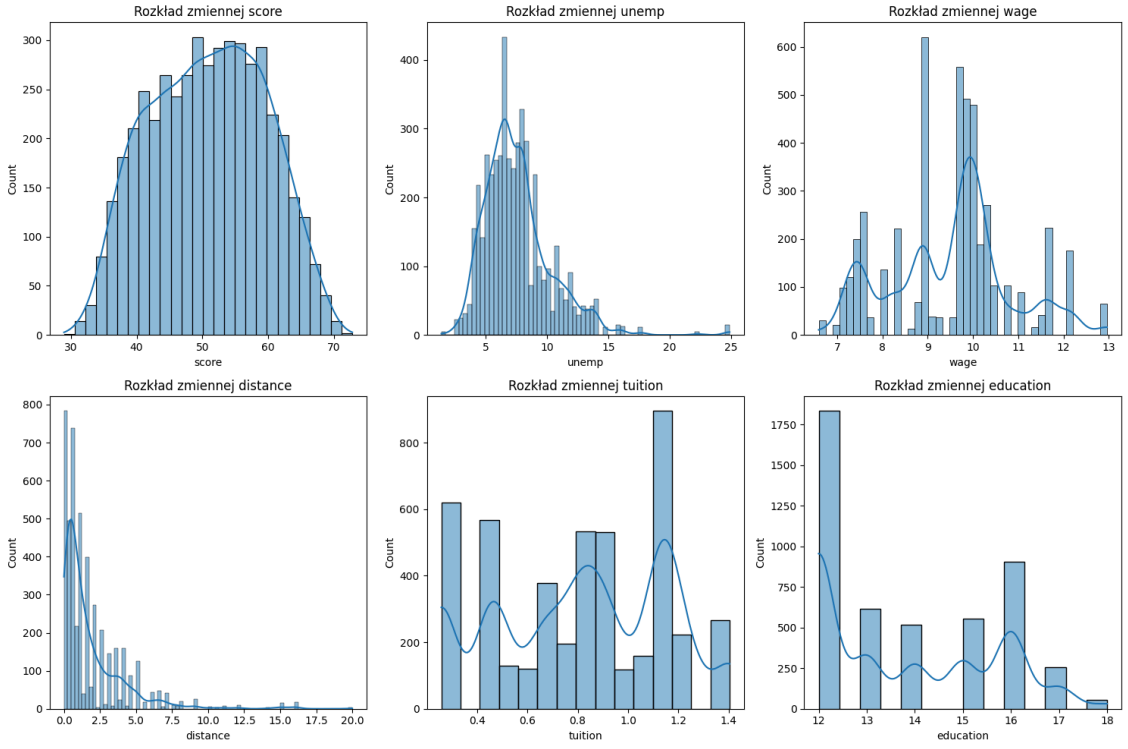
Raport: Budowa Modelu Predykcyjnego dla Przewidywania Wyniku Score

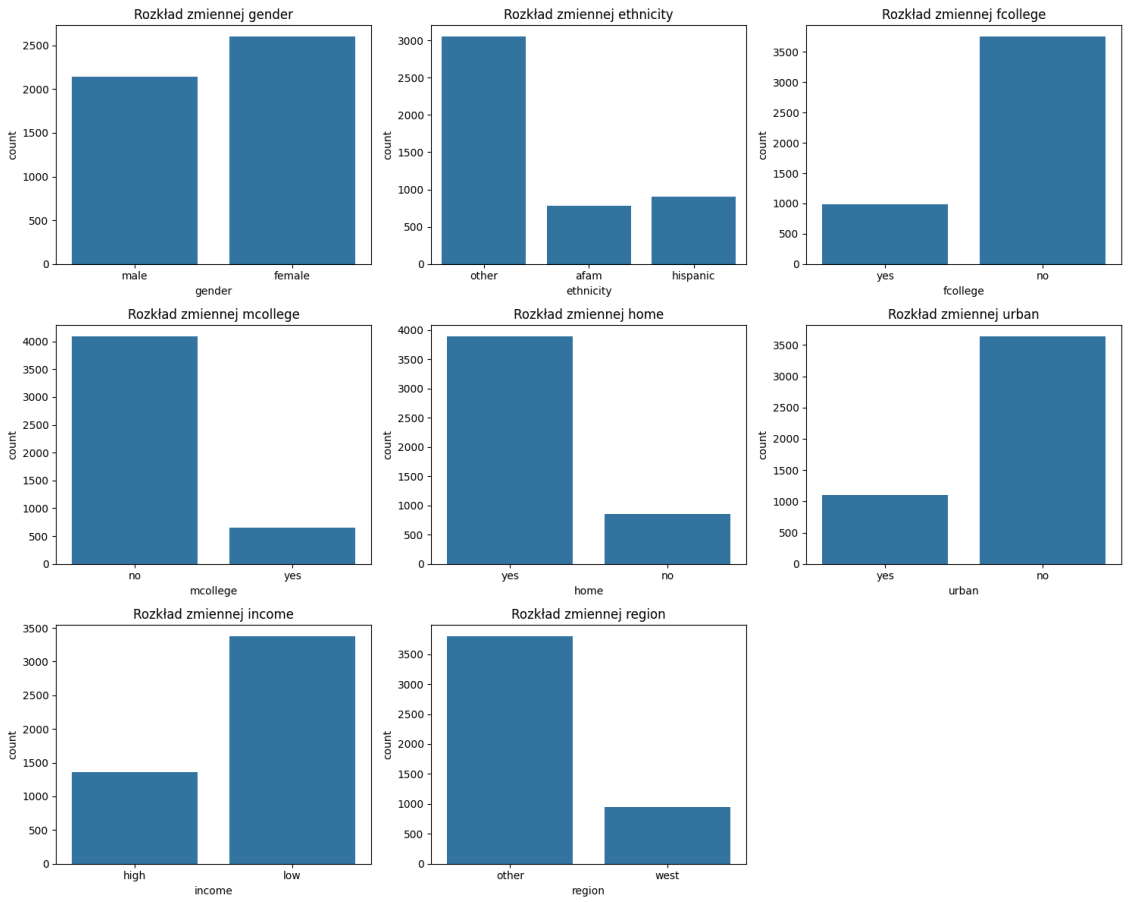
1. Cel projektu

Celem projektu było zbudowanie modelu predykcyjnego, który przewidywałby wartość zmiennej score na podstawie zestawu danych. W tym celu przeprowadzono analizę danych, inżynierię cech, wybór oraz ocenę jakości modelu.

2. Eksploracja i Wstępna Analiza Danych

Struktura Danych: Zestaw danych zawiera 15 kolumn i 4739 wierszy. Zmienna score jest zmienną docelową, a pozostałe kolumny są cechami predykcyjnymi.
Brakujące Wartości: Analiza wykazała brak brakujących wartości w danych, co pozwoliło uniknąć imputacji.
Statystyki Opisowe:
Średnia wartość score wynosi 50,89, z odchyleniem standardowym 8,7.
Zmienność w innych zmiennych liczbowych, takich jak unemp, wage, distance, i tuition, sugerowała potrzebę standaryzacji.
Wizualizacja: Histogramy i wykresy słupkowe pokazały rozkład zmiennych liczbowych i kategorycznych, ułatwiając identyfikację potencjalnych relacji między zmiennymi.





3. Inżynieria Cech i Przygotowanie Danych

Kategoryzacja: Przekształcono zmienne kategoryczne (gender, ethnicity, itp.) za pomocą LabelEncoder, co umożliwiło ich wykorzystanie w modelu.
Standaryzacja: Zastosowano standaryzację cech liczbowych (unemp, wage, distance, tuition, education), co umożliwia lepszą stabilność modelu.
Podział Zbioru Danych: Dane podzielono na zbiór treningowy (70%) i testowy (30%) dla oceny modelu.

4. Wybór i Trenowanie Modelu

Model Random Forest: Wybrano RandomForestRegressor jako główny model predykcyjny ze względu na jego zdolność do wychwytywania złożonych wzorców i dobrego działania z danymi zawierającymi zmienne kategoryczne oraz liczbowe.
Trenowanie Modelu: Model wytrenowano na zbiorze treningowym, a następnie oceniono jego jakość na zbiorze testowym.

5. Ocena i Optymalizacja Modelu

Dla oceny modelu zastosowano następujące metryki:
Mean Absolute Error (MAE): 0.69 – wskazuje na przeciętną bezwzględną różnicę między przewidywaniami a rzeczywistymi wartościami.
Mean Squared Error (MSE): 0.75 – podkreśla większe błędy predykcji (kwadrat błędu).

R-squared (R^2): 0.26 – sugeruje, że model wyjaśnia około 26% zmienności zmiennej score, co wskazuje na możliwość dalszej optymalizacji.

6. Podsumowanie i Wnioski

Model Random Forest zapewnił stabilne wyniki, ale istnieje przestrzeń do poprawy, w szczególności poprzez:

- Optymalizację hiperparametrów modelu (np. zwiększenie liczby drzew, testowanie głębokości drzewa).
- Eksperymenty z innymi modelami, jak regresja liniowa lub XGBoost (możliwy do przetestowania na większej mocy obliczeniowej).