Raport z budowy modelu predykcyjnego do przewidywania zmiennej *score* w CollegeDistance

Etap 1: Wstępna analiza danych

Dane zostały wczytane i poddane analizie w celu zidentyfikowania braków w danych oraz podstawowych rozkładów zmiennych.

Dla zmiennych numerycznych brakujące dane zostały uzupełnione medianą, co umożliwiło zachowanie pierwotnej struktury ich rozkładu bez wprowadzania skrajnych wartości.

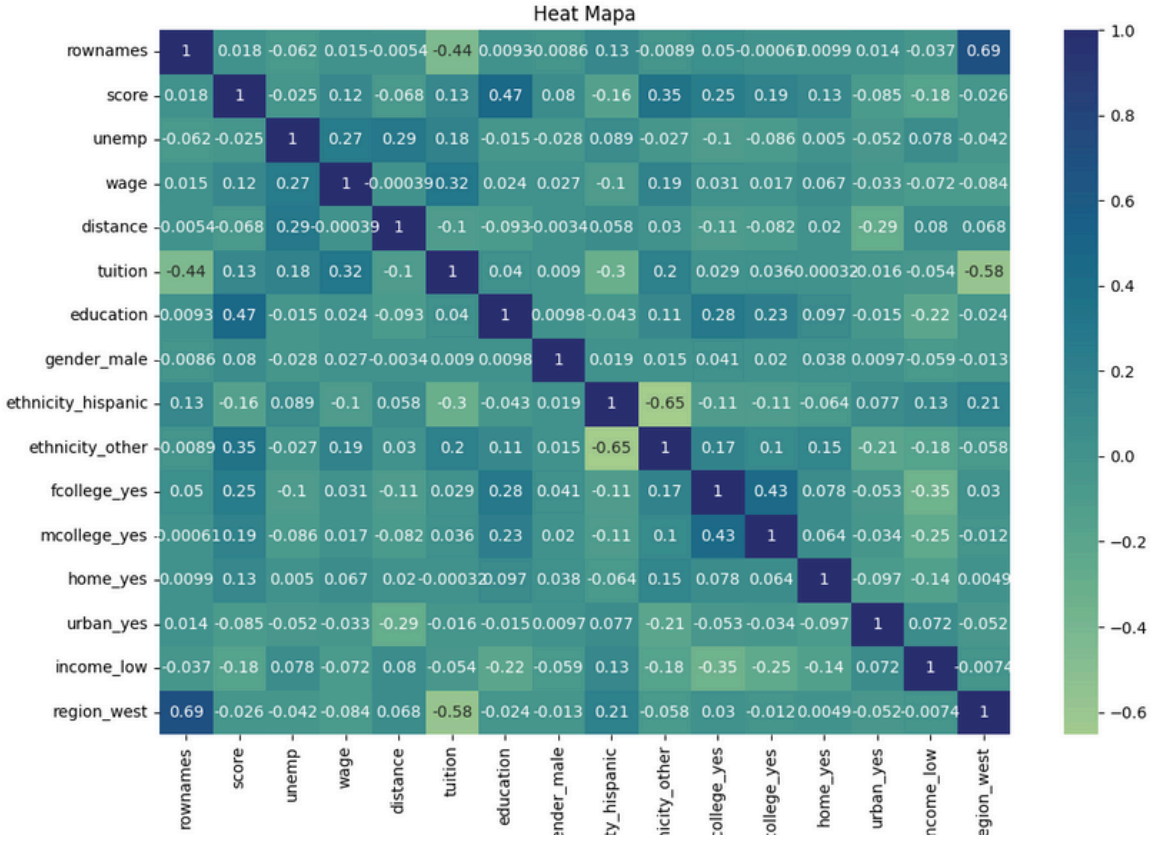
Dla zmiennych kategorycznych zastosowano imputację metodą najczęściej występującej wartości

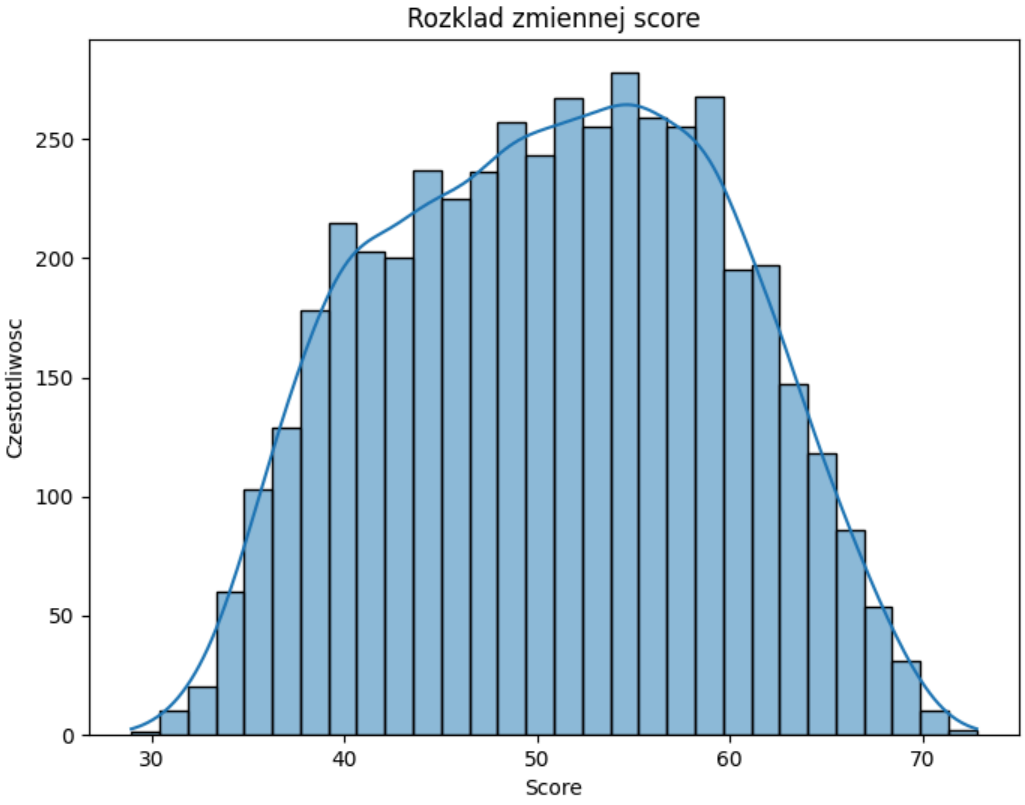
Przeprowadzono statystyczną analizę zmiennych, wygenerowano statystyki, takie jak średnia, mediana, kwartyle, rozkłady zmiennych itd.

```
Pierwsze 5 wierszy:
   rownames  gender ethnicity      score  ...  tuition education income region
0         1    male     other  39.150002  ...  0.88915        12   high  other
1         2  female     other  48.869999  ...  0.88915        12    low  other
2         3    male     other  48.740002  ...  0.88915        12    low  other
3         4    male      afam  40.400002  ...  0.88915        12    low  other
4         5  female     other  40.480000  ...  0.88915        13    low  other

[5 rows x 15 columns]

Ostatnie 5 wierszy:
      rownames gender ethnicity      score  ...  tuition education income region
4734      9391   male      afam  56.529999  ...  0.25751        13   high   west
4735      9401   male      afam  59.770000  ...  0.25751        15   high   west
4736      9411   male     other  43.169998  ...  0.25751        12   high   west
4737      9421   male      afam  49.970001  ...  0.25751        16   high   west
4738      9431   male      afam  53.410000  ...  0.25751        13   high   west

[5 rows x 15 columns]

Brakujace wartosci w danych kolumnach:
rownames     0
gender       0
ethnicity    0
score        0
fcollege     0
mcollege     0
home         0
urban        0
unemp        0
wage         0
distance     0
tuition      0
education    0
income       0
region       0
dtype: int64

Statystyki dla zmiennych numerycznych:
           rownames        score  ...      tuition    education
count   4739.000000  4739.000000  ...  4739.000000  4739.000000
mean    3954.638953    50.889029  ...     0.814608    13.807765
std     5953.827761     8.701910  ...     0.339504     1.789107
min        1.000000    28.950001  ...     0.257510    12.000000
25%     1185.500000    43.924999  ...     0.484990    12.000000
50%     2370.000000    51.189999  ...     0.824480    13.000000
75%     3554.500000    57.769999  ...     1.127020    16.000000
max    37810.000000    72.809998  ...     1.404160    18.000000
```

```
Rozklad zmiennej gender:
gender
female    2600
male      2139
Name: count, dtype: int64

Rozklad zmiennej ethnicity:
ethnicity
other       3050
hispanic     903
afam         786
Name: count, dtype: int64

Rozklad zmiennej income:
income
low     3374
high    1365
Name: count, dtype: int64

Brakujace wartosci po imputacji:
rownames     0
gender       0
ethnicity    0
score        0
fcollege     0
mcollege     0
home         0
urban        0
unemp        0
wage         0
distance     0
tuition      0
education    0
income       0
region       0
dtype: int64
```

Wygenerowano Heat Mapę, oraz wykres dla rozkładu zmiennej **score**:


Heat Mapa
rownames 1 0.018 -0.062 0.015 -0.0054 -0.44 0.0093 -0.0086 0.13 -0.0089 0.05 -0.00061 0.0099 0.014 -0.037 0.69
score 0.018 1 -0.025 0.12 -0.068 0.13 0.47 0.08 -0.16 0.35 0.25 0.19 0.13 -0.085 -0.18 -0.026
unemp -0.062 -0.025 1 0.27 0.29 0.18 -0.015 -0.028 0.089 -0.027 -0.1 -0.086 0.005 -0.052 0.078 -0.042
wage 0.015 0.12 0.27 1 -0.00039 0.32 0.024 0.027 -0.1 0.19 0.031 0.017 0.067 -0.033 -0.072 -0.084
distance -0.0054 -0.068 0.29 -0.00039 1 -0.1 -0.093 -0.0034 0.058 0.03 -0.11 -0.082 0.02 -0.29 0.08 0.068
tuition -0.44 0.13 0.18 0.32 -0.1 1 0.04 0.009 -0.3 0.2 0.029 0.036 0.00032 0.016 -0.054 -0.58
education -0.0093 0.47 -0.015 0.024 -0.093 0.04 1 0.0098 -0.043 0.11 0.28 0.23 0.097 -0.015 -0.22 -0.024
gender_male -0.0086 0.08 -0.028 0.027 -0.0034 0.009 0.0098 1 0.019 0.015 0.041 0.02 0.038 0.0097 -0.059 -0.013
ethnicity_hispanic 0.13 -0.16 0.089 -0.1 0.058 -0.3 -0.043 0.019 1 -0.65 -0.11 -0.11 -0.064 0.077 0.13 0.21
ethnicity_other -0.0089 0.35 -0.027 0.19 0.03 0.2 0.11 0.015 -0.65 1 0.17 0.1 0.15 -0.21 -0.18 -0.058
fcollege_yes 0.05 0.25 -0.1 0.031 -0.11 0.029 0.28 0.041 -0.11 0.17 1 0.43 0.078 -0.053 -0.35 0.03
mcollege_yes -0.00061 0.19 -0.086 0.017 -0.082 0.036 0.23 0.02 -0.11 0.1 0.43 1 0.064 -0.034 -0.25 -0.012
home_yes -0.0099 0.13 0.005 0.067 0.02 -0.00032 0.097 0.038 -0.064 0.15 0.078 0.064 1 -0.097 -0.14 0.0049
urban_yes 0.014 -0.085 -0.052 -0.033 -0.29 -0.016 -0.015 0.0097 0.077 -0.21 -0.053 -0.034 -0.097 1 0.072 -0.052
income_low -0.037 -0.18 0.078 -0.072 0.08 -0.054 -0.22 -0.059 0.13 -0.18 -0.35 -0.25 -0.14 0.072 1 -0.0074
region_west 0.69 -0.026 -0.042 -0.084 0.068 -0.58 -0.024 -0.013 0.21 -0.058 0.03 -0.012 0.0049 -0.052 -0.0074 1
rownames score unemp wage distance tuition education gender_male ethnicity_hispanic ethnicity_other fcollege_yes mcollege_yes home_yes urban_yes income_low region_west
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6


Rozklad zmiennej score
Czestotliwosc
0 50 100 150 200 250
30 40 50 60 70
Score

Wnioski z analizy: Zmienna docelowa **score**, wykazywała umiarkowaną zmienność w przedziale od 28.95 do 72.81. Pozostałe zmienne obejmowały różne aspekty socjalne (np. gender, income, region), które mogły wpływać na score. Zmienne **distance** i **wage** posiadają istotne odchylenia standardowe, co może sugerować ich potencjalny wpływ na zmienną **score.**

Etap 2: Inżynieria cech i przygotowanie danych

- W ramach tego etapu dokonano przetworzenia zmiennych kategorycznych na format numeryczny, stosując metodę One-Hot Encoding. Kodowanie zmiennych takich jak gender, income, oraz region umożliwiło modelowi wykorzystanie wszystkich informacji zawartych w tych kolumnach.
- Dane zostały podzielone na zbiór treningowy i testowy w proporcji 80/20
- Ostatecznie uzyskano zbiór treningowy o rozmiarze 3791 próbek oraz zbiór testowy o rozmiarze 948 próbek

Dzięki odpowiedniemu przygotowaniu danych zwiększono szanse na skuteczne trenowanie modelu

Etap 3: Wybór i trenowanie modelu

Jako model rozważono **regresję liniową** oraz **lasy losowe**. Regresja liniowa, jako metoda oparta na prostym modelowaniu liniowych zależności, została użyta jako model bazowy.

Model lasów losowych, znany z wysokiej skuteczności przy analizie danych o złożonych relacjach, został wybrany do oceny i porównania z regresją liniową.

Skuteczność tych modeli oceniono przy pomocy metryk takich jak średni błąd kwadratowy (MSE), współczynnik determinacji (R^2), oraz średni błąd absolutny (MAE).

Wyniki:

- **Regresja liniowa**:
  - MSE: 49.04
  - R^2: 0.353
  - MAE: 5.75
- **Lasy losowe**:
  - MSE: 52.10
  - R^2: 0.313
  - MAE: 5.76

Wniosek: Regresja liniowa uzyskała trochę lepsze wyniki niż lasy losowe, co sugeruje, że zmienna score może być w pewnym stopniu liniowo zależna od pozostałych cech, jednak niska wartość R^2 wskazuje na ograniczoną dokładność predykcji.

Etap 4: Ocena i optymalizacja modelu

Wynik dla algorytmu lasów losowych nie był satysfakcjonujący, więc wprowadzono optymalizację:

- Przeprowadzono tuning hiperparametrów z użyciem GridSearch, testując różne wartości głębokości drzewa (max_depth), liczby cech (max_features), oraz liczby estymatorów (estimators).
- Optymalizacja miała na celu poprawienie dokładności modelu i zmniejszenie wartości błędu przewidywania.

Wyniki po optymalizacji:

- **Najlepsze hiperparametry**:
  - max_depth: 10
  - max_features: 'sqrt'
  - n_estimators: 150
- **Ostateczne wyniki lasów losowych**:
  - MSE: 47.82
  - R^2: 0.369
  - MAE: 5.70

Wnioski końcowe:

Optymalizacja modelu lasów losowych tuningiem hiperparametrów sprawiła, że radził on sobie lepiej na zbiorze testowym. Zapewnia on stosunkowo najlepsze wyniki i umożliwia umiarkowanie dokładne przewidywanie zmiennej **score.**