Analiza Danych Studentów

Radosław Drogoś s24223

1. Wprowadzenie i opis danych

Celem jest znalezienie najlepszego modelu który przewiduje wartość score.

W zbiorze danych jest 4739 wierszy oraz 15 kolumn.

Kolumny numeryczne to: score, unemp, wage, distance, tuition, education.

Kolumny z danymi nominalnymi to: gender, ethnicity, fcollege, mcollege, home, urban, income, region.

W zbiorze danych brak kolumn z wartościami null.

Przykładowe wiersze danych:

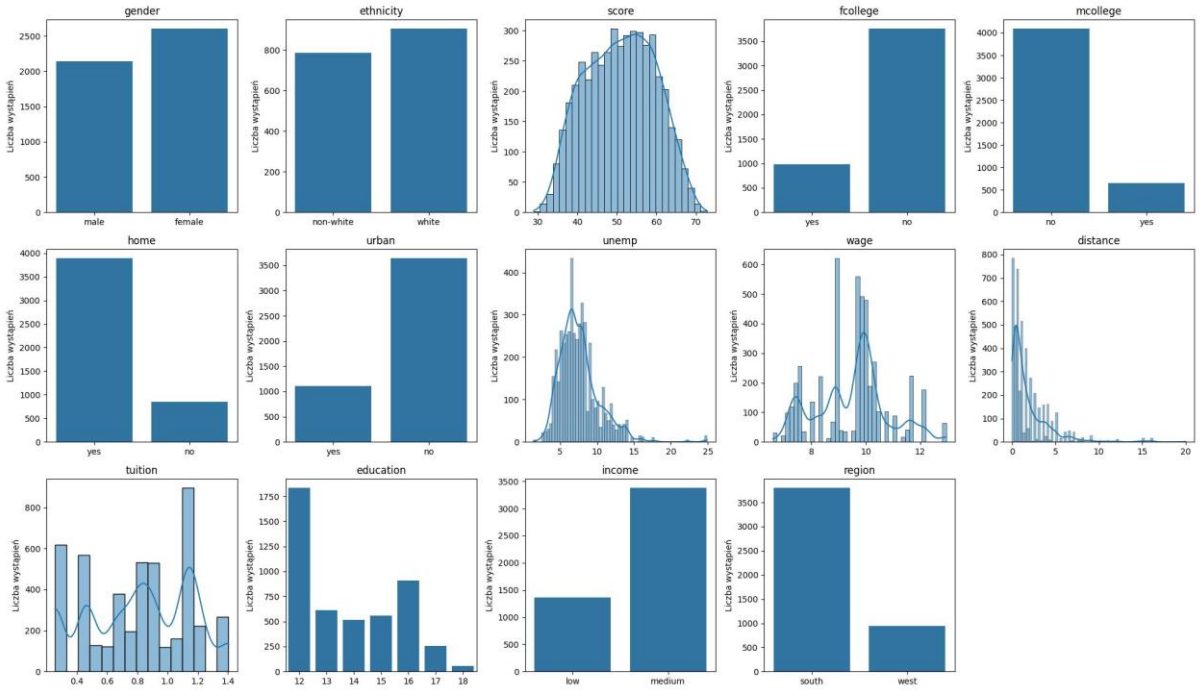
rownames	gender	ethnicity	score	...	tuition	education	income	region
1	1	2	39.1500	...	0.88915	0	0	0
2	0	2	48.8700	...	0.88915	0	1	0
3	1	2	48.7400	...	0.88915	0	1	0
4	1	0	40.4000	...	0.88915	0	1	0
5	0	2	40.4800	...	0.88915	1	1	0

Statystyki opisowe dla zmiennych:

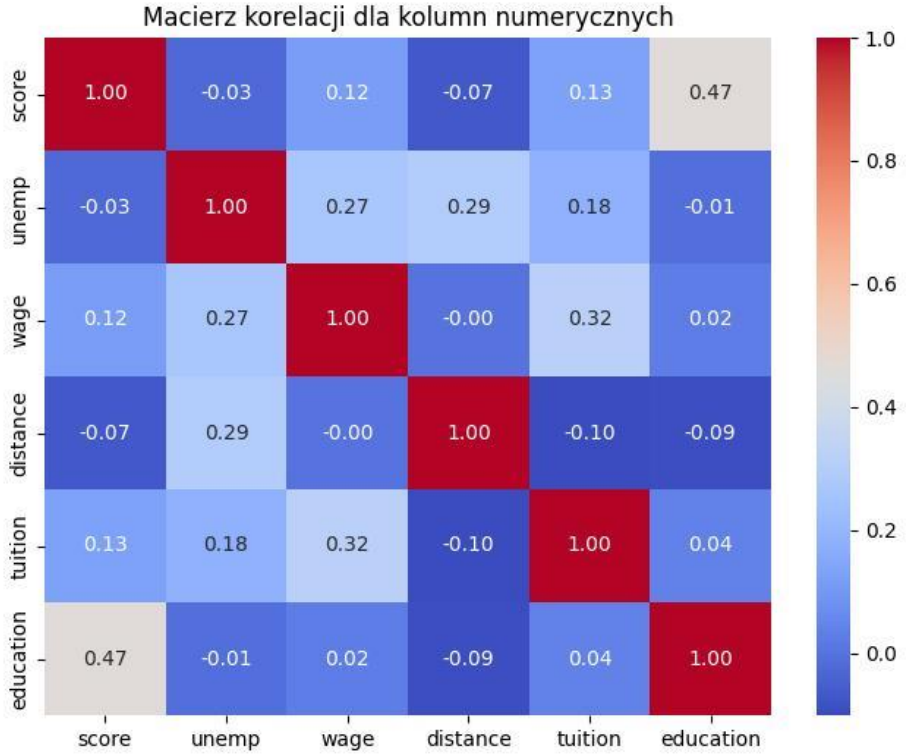
Nazwa Kolumny	rownames	gender	…	income	region
Średnia	3954.64	0.45	…	0.71	0.2
Odchylenie Std.	5953.83	0.50	…	0.45	0.4
Min	1	0	…	0	0
Q1	1185.5	0	…	0	0
Mediana	2370.0	0	…	1	0
Q3	3554.5	1	…	1	0
Max	37810	1	…	1	1

Wartości zmiennych są w większości kategoryczne lub binarne, co może wpływać na wybór algorytmów modelowania.

Histogramy dla danych:



Macierz korelacji dla danych numerycznych:



2. Przetwarzanie Danych:

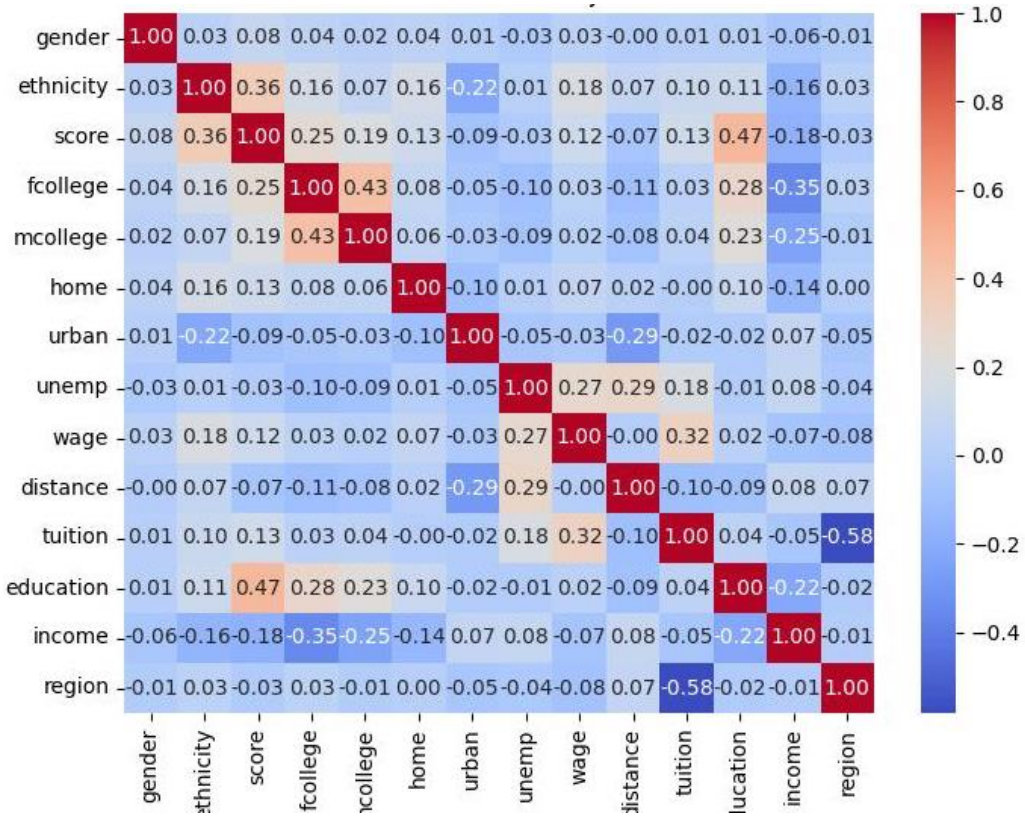
Usunięto zbędne kolumny, takie jak rownames, kolumny z dużą ilością braków.

Usunięto kolumny o niskiej wariancji.

Wyodrębniono cechy i zmienną docelową, przeskalowano je, aby modelowanie było bardziej wydajne.

Ze względu na to, iż w zbiorze występują w większości dane, które posiadają małą ilość różne wartości zostały one zamienione na postać liczbową tj. 0,1,2

Macierz korelacji:



3. Wybór i analiza danych modelu

Ze względu na to, iż przewidywane są wartości numeryczne, zostały wybrane modele regresyjne: Linear Regression, Random Forest oraz XGBoost.

Opis Wskaźników:

Mean Squared Error (MSE): Średni błąd kwadratowy - metryka używana do mierzenia średniego kwadratowego odchylenia między przewidywanymi a rzeczywistymi wartościami. Niższe wartości MSE wskazują na lepszą zgodność modelu z danymi.

R-squared (R^2): Miara, która pokazuje, jaka część wariancji zmiennej zależnej jest wyjaśniona przez zmienne niezależne. Wartość bliska 1 oznacza, że model dobrze dopasowuje się do danych, podczas gdy wartości bliższe 0 wskazują na słabe dopasowanie.

Mean Absolute Error (MAE): Średni błąd bezwzględny mierzy średnie bezwzględne różnice między przewidywanymi a rzeczywistymi wartościami. Mniejsza wartość MAE oznacza lepszą dokładność modelu.

Wyniki początkowe modeli:

	Linear Regression	Random Forest	XGBoost
MSE:	49.15	53.95	57.8
R^2:	0.35	0.29	0.24
MAE:	5.76	5.87	6.04

Tuning (Grid Search z 5-krotną Walidacją Krzyżową):

1. Random Forest
    a. najlepsze parametry: 'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 100
    b. najlepszy MSE (CV): 52.20387419987291
2. XGBoost
    a. Najlepsze parametry: 'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 50
    b. najlepszy MSE (CV): 50.38574997589963

Wyniki Modeli Po Tuningu:

	Linear Regression	Random Forest	XGBoost
MSE:	49.15	49.35	48.36
R^2:	0.35	0.35	0.36
MAE:	5.76	5.71	5.69

Ze względu na wyniki został wybrany XGBoost.

1. **Wnioski:**

Po przeprowadzonych analizach, model XGBoost został wybrany jako najlepszy do przewidywania ocen studentów. Jego wyniki, mierzone za pomocą błędu średniokwadratowego (MSE) i współczynnika determinacji (R^2), były lepsze niż w przypadku innych testowanych modeli, takich jak Random Forest i Linear Regression.

Mimo lepszych wyników, model XGBoost nie jest idealny. Niska wartość R^2 wskazuje, że dokładność przewidywań jest ograniczona. Wysoki MSE z kolei sugeruje, że prognozy modelu mogą znacznie różnić się od rzeczywistych ocen.

Możliwe przyczyny ograniczeń:

Brakujące dane: Model nie uwzględnia wszystkich czynników wpływających na ocenę studenta, takich jak zdolności, motywacja czy sytuacja życiowa.

Subiektywne czynniki: Ocena studenta jest w pewnym stopniu zależna od subiektywnych ocen nauczycieli, a także od losowych zdarzeń, które trudno przewidzieć.

Zbyt prosty model: Model opiera się głównie na danych historycznych i nie uwzględnia bardziej złożonych zależności między różnymi czynnikami.