Lab3 – Analizator Wyników

S25098 Mikołaj Antoszewski

Wstęp

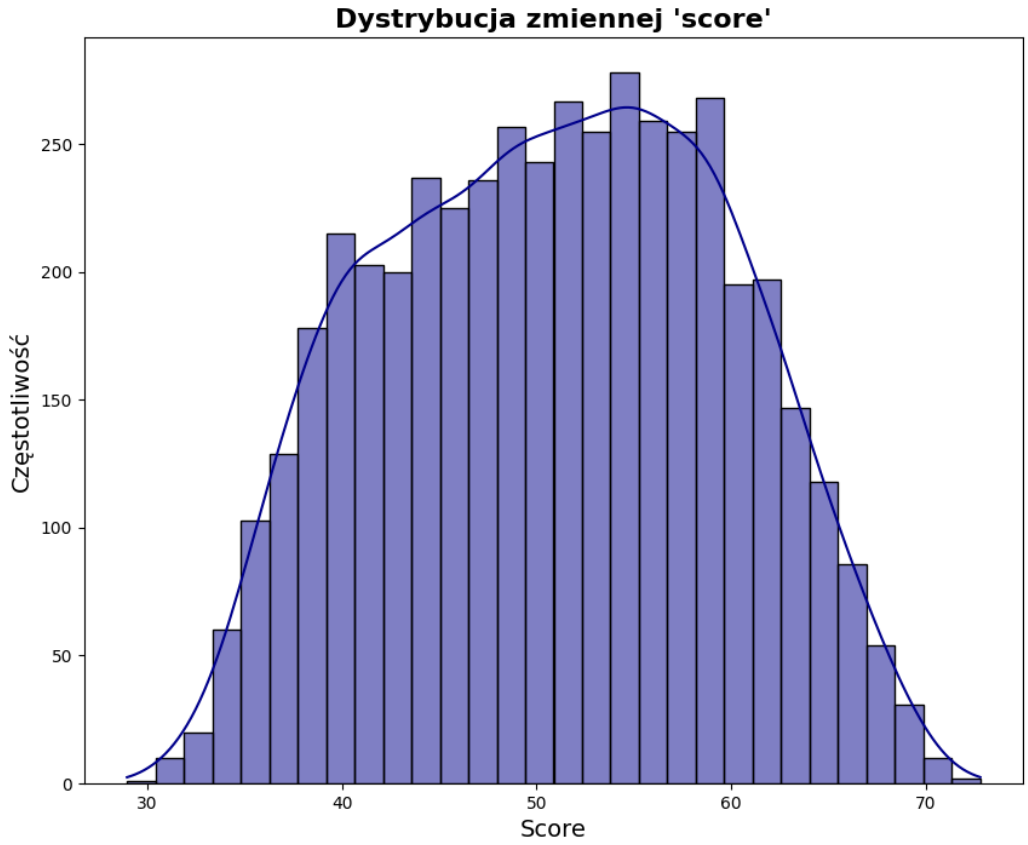
Projekt polega na stworzeniu modelu predykcyjnego, który ma na celu przewidywanie wyniku edukacyjnego (zmiennej **score**) na podstawie dostarczonego zbioru danych

Analiza danych

Dataset zawiera 4739 wierszy, 6 kolumn numerycznych i 8 kolumn kategorycznych Opisuje on dane dotyczące różnych osób, które mają na celu przewidywanie wyniku edukacyjnego (zmienna **score**) na podstawie różnych cech demograficznych i społecznych. Oto krótka charakterystyka poszczególnych zmiennych w zbiorze danych:

- **rownames** – Indeks wiersza w zbiorze danych, który identyfikuje poszczególne rekordy. Zostaje usunięty przed analizą.
- **gender** – Płeć osoby (np. male, female).
- **ethnicity** – Etniczność osoby, gdzie "other" odnosi się do innych grup etnicznych.
- **score** – Wynik edukacyjny, który jest zmienną zależną, której wartość ma zostać przewidziana przez model.
- **fcollege** – Wskazuje, czy osoba ma ukończoną szkołę wyższą.
- **mcollege** – Określa, czy matka osoby ma ukończoną szkołę wyższą.
- **home** – Wskazuje, czy osoba mieszka w domu.
- **urban** – Określa, czy osoba mieszka w obszarze miejskim.
- **unemp** – Liczba osób bez pracy w gospodarstwie domowym.
- **wage** – Wysokość wynagrodzenia osoby.
- **distance** – Odległość do najbliższej szkoły lub uczelni.
- **tuition** – Wysokość czesnego, które osoba płaci za edukację.
- **education** – Poziom wykształcenia osoby.
- **income** – Dochód gospodarstwa domowego.
- **region** – Region, w którym osoba mieszka.

Poniżej przedstawiono rozkład zmiennej, którą będziemy prognozować, czyli **score**. Rozkład ten jest stosunkowo symetryczny z średnią wynoszącą 50,9.

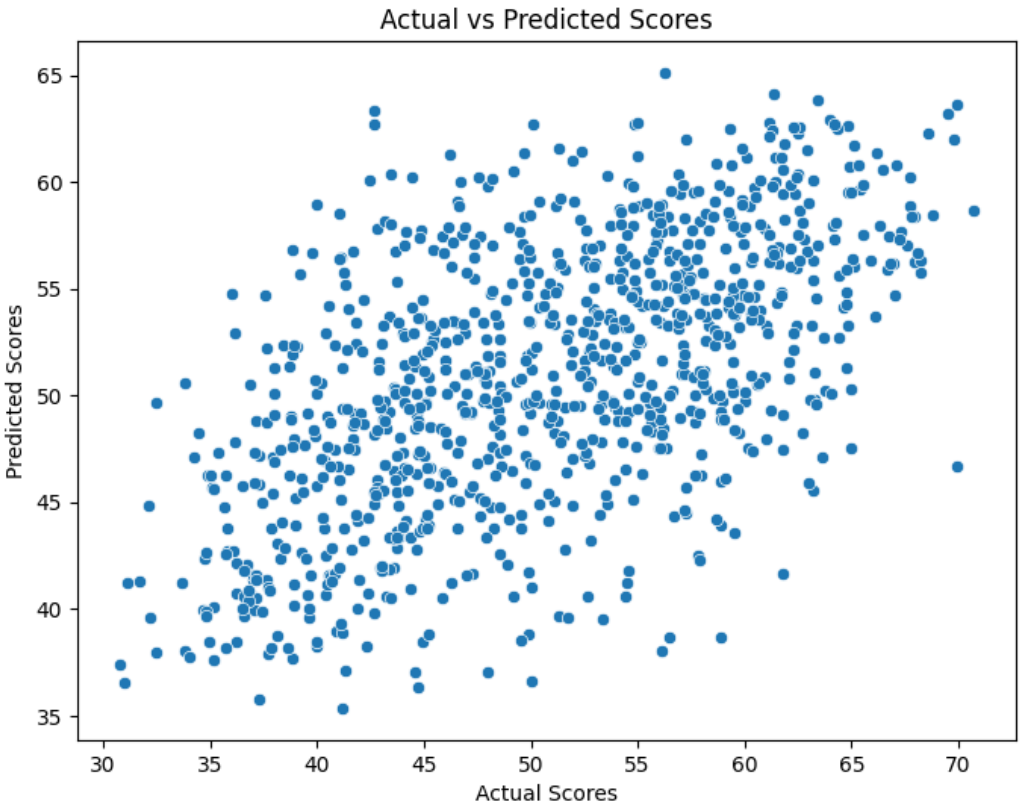


Przygotowanie danych:

1. **Wczytanie i czyszczenie danych:** Dane są wczytywane z pliku CSV, a brakujące wartości usuwane.
2. **Eksploracja danych:** Usuwane są nieistotne kolumny, a zmienne kategoryczne są kodowane. Tworzone są wykresy korelacji oraz rozkładu zmiennej **score**, co pozwala na lepsze zrozumienie danych.
3. **Przygotowanie danych:** Zmienna **score** jest wydzielana jako zmienna zależna, a pozostałe cechy są kodowane i skalowane. Dodawane są nowe cechy, jak iloczyn **distance** i **tuition**, oraz logarytmy dla **income** i **distance**. Dane są dzielone na zbiór treningowy i testowy.
4. **Zapis danych:** Przetworzone dane są zapisywane do plików CSV, co umożliwia ich dalsze wykorzystanie w modelowaniu.

Wnioski

Wykres służący do oceny dokładności modelu przedstawia porównanie **wyników rzeczywistych** z **przewidywanymi** w formie punktowej. Widoczna jest pozytywna korelacja, choć punkty są rozproszone, co wskazuje na pewne błędy modelu. Zakres wyników rzeczywistych to 30–70, a przewidywanych 35–65.



Regresja Liniowa:

Mean Squared Error (MSE): 49.01

R-squared (R^2): 0.35

Mean Absolute Error (MAE): 5.74

Las Losowy:

Mean Squared Error (MSE): 51.59

R-squared (R^2): 0.32

Mean Absolute Error (MAE): 5.75

Regresja Liniowa wydaje się lepszym modelem w tym przypadku, biorąc pod uwagę mniejsze MSE, MAE oraz wyższe R^2.