Lab3 – Analizator wyników – s23756

Raport z analizy i budowy modelu predykcyjnego

1. Wprowadzenie

Cel: Celem projektu jest zbudowanie modelu predykcyjnego, który przewiduje wartość zmiennej `score` na podstawie danych społeczno-ekonomicznych i demograficznych. Model ma za zadanie pomóc w lepszym zrozumieniu czynników wpływających na wynik `score` i ich znaczenie. W tym celu przeprowadzono szczegółową analizę danych, wybór modelu, trening i ocenę, a także wstępną optymalizację.

---

2. Eksploracja i wstępna analiza danych

2.1 Wczytanie i wstępna inspekcja danych

Dane zostały wczytane do środowiska analitycznego i poddane wstępnej inspekcji w celu zrozumienia ich struktury. Plik zawierał 15 kolumn oraz 4739 wierszy. Zmienna docelowa `score` jest zmienną liczbową, podczas gdy pozostałe kolumny obejmują zarówno zmienne kategoryczne, jak i liczbowe.

2.2 Analiza braków danych

Sprawdzono kompletność danych – wszystkie kolumny okazały się pełne, co oznacza, że w zbiorze nie ma brakujących wartości. Ten etap pozwala wyeliminować potencjalne błędy wynikające z braków danych i decydować o ewentualnej imputacji lub usunięciu brakujących danych.

2.3 Statystyki opisowe

Przeprowadzono analizę statystyczną, aby lepiej zrozumieć rozkład zmiennych liczbowych i sprawdzić, czy występują wartości odstające. Obliczone wartości średnie, odchylenia standardowe, kwartyle i zakresy umożliwiły zrozumienie ogólnej struktury danych, m.in. średnia wartość `score` wynosi 50.89, co wskazuje na pewne zróżnicowanie wyników.

2.4 Wizualizacja zmiennych

Dane wizualizowano w celu lepszego zrozumienia rozkładów i relacji pomiędzy zmiennymi. Wykresy wykazały m.in. że zmienna `distance` może mieć wpływ na `score`, co sugeruje, że bliższe odległości mogą wiązać się z wyższymi wynikami.

Wnioski: Analiza wstępna potwierdziła pełność danych oraz wskazała na istotne zróżnicowanie wartości zmiennej `score`. Potencjalnie istnieje korelacja pomiędzy `score` a zmiennymi takimi jak `distance`.

---

3. Inżynieria cech i przygotowanie danych

3.1 Przekształcenie danych kategorycznych

Zmiennym kategorycznym, takim jak `gender`, `ethnicity`, `income`, `region`, przypisano wartości numeryczne. Użyto techniki kodowania „one-hot", aby umożliwić modelowi prawidłowe rozpoznawanie ich wpływu bez zakłóceń związanych z wartościami kategorycznymi.

3.2 Podział danych

Dane zostały podzielone na zbiór treningowy (80%) i testowy (20%), co pozwala na ocenę modelu na danych niezależnych od tych, na których był trenowany.

3.3 Standaryzacja zmiennych

Przeprowadzono standaryzację zmiennych liczbowych, co umożliwiło ujednolicenie skali wartości i wyeliminowanie wpływu zmiennych o dużym zakresie na wynik końcowy.

Wnioski: Działania przygotowawcze poprawiły jakość danych, zapewniając lepszą spójność i większe możliwości predykcyjne modelu.

---

4. Wybór i trenowanie modelu

4.1 Wybór algorytmu

Na podstawie analizy danych zdecydowano się na wykorzystanie regresji liniowej – jest to model odpowiedni do przewidywania zmiennych ciągłych, a zmienna `score` dobrze pasuje do takiego podejścia.

4.2 Trening modelu

Model został przeszkolony na danych treningowych, co umożliwiło uchwycenie zależności pomiędzy zmiennymi wejściowymi a zmienną docelową `score`. Wykorzystano dane ze standaryzacją i odpowiednim zakodowaniem zmiennych kategorycznych, aby uzyskać lepsze wyniki.

Wnioski: Model regresji liniowej ujawnił podstawowe relacje między zmiennymi wejściowymi a `score`, co jest cenną wskazówką do dalszej pracy i optymalizacji.

---

5. Ocena i optymalizacja modelu

5.1 Metryki oceny

Model oceniono na zbiorze testowym, stosując następujące miary:

- **MAE** (średni błąd bezwzględny): 5.75, co wskazuje na przeciętne odchylenie przewidywanej wartości `score` od rzeczywistej wartości o około 5,75 punktu.
- **MSE** (średni błąd kwadratowy): 49.04, co wskazuje na większe błędy przy niektórych przewidywaniach.
- **R^2**: 0.35, co oznacza, że model wyjaśnia 35% zmienności w `score`.

5.2 Optymalizacja

Aby poprawić wyniki, przeprowadzono wstępną optymalizację, w tym walidację krzyżową i regulację hiperparametrów. Mimo to model wymaga dalszych udoskonaleń, aby uzyskać lepszą trafność predykcji.

Wnioski: Model regresji liniowej wykazuje umiarkowaną trafność, co uzasadnia potrzebę dalszej optymalizacji i potencjalnie bardziej zaawansowanych modeli predykcyjnych.

---

6. Wnioski końcowe

Opracowany model predykcyjny pozwala na wstępne przewidywanie wartości `score` na podstawie zidentyfikowanych cech demograficznych i społecznych. Analiza wykazała, że cechy te mają wpływ na `score`, jednak dokładność modelu można poprawić poprzez zastosowanie bardziej zaawansowanych technik modelowania.

Rekomendacje:

1. Przyszłe badania mogłyby obejmować bardziej złożone modele, jak np. lasy losowe czy sieci neuronowe.
2. Dalsza optymalizacja parametrów modelu może prowadzić do poprawy trafności przewidywań.
3. Przeprowadzenie dodatkowej analizy na większym zbiorze danych mogłoby zwiększyć wiarygodność wyników.