Dokumentacja LAB 3 – s24667

1.Eksploracja i wstępna analiza danych

CollegeDistance zawiera 4739 wierszy, 7 kolumn numerycznych i 8 kolumn kategorycznych. W całym zbiorze danych nie występują żadne braki. W kolumnie ‘distance’ pojawiają się wartości 0, aczkolwiek można to zinterpretować, że te osoby zakwaterowane są na terenie uczelni.

2.Inżynieria cech i przygotowanie danych

Wykonano następujące działania w celu przygotowania danych:

- Kategoryzacja: wykorzystano OneHotEncoder w celu przekształcenia wartości kategorycznych w wartości binarne

- Imputacja: w przypadku wystąpienia brakujących wartości następuje ich uzupełnienie:

    -dla wartości numerycznych, uzupełniane są średnią wartości w kolumnie

    -dla wartości kategorycznych, uzupełniane są najczęściej pojawiającą się kategorią

- Standaryzacja: dokonano standaryzacji danych numerycznych przy pomocy StandardScaler

- Podział danych na zbiór treningowy i testowy

3.Wybór i trenowanie modelu

RANDOM FOREST - Wybrany został model lasu losowego, dobrze radzi sobie z danymi o wysokiej złożoności oraz nieliniowością.

4.Ocena i optymalizacja modelu

Uzyskane wyniki dla zbioru testowego:
MAE = 5.8995341170264295
R^2 = 0.2542072511350574

W celu polepszenia uzyskanych wyników, została wykonana optymalizacja modelu. Przeprowadzono tunowanie hiperparametrów i walidację krzyżową przy pomocy GridSearchCV

Uzyskane wyniki po optymalizacji modelu:
MAE = 5.6972939478006
R^2 = 0.329420189183869