Projekt Zaliczeniowy - wytyczne

Dane

Dane pochodzą ze strony (http://archive.ics.uci.edu/ml/datasets/Adult); jest to publicznie dostępny zbiór danych wykorzystywany w projektach machine-learning

Cel

Opracowanie modelu (regresja logistyczna), który na podstawie danych przewidzi przychód osoby (poniżej 50k rocznie lub powyżej 50k rocznie)

Zadania

1. W kolumnie type_employer:
    1. zastąpić wpisy "Federal-gov" i "Local-gov" wpisem "SL-gov"
    2. zastąpić wpisy "Self-emp-inc" i "Self-emp-not-inc" wpisem "self-emp"
2. W kolumnie "marital" zredukować liczbę wpisów do trzech (Married;Not-Married;Never-Married)
3. Zmniejszyć liczbę wpisów w kolumnie country (np. grupowanie przez kontynenty? inne podejście?)
4. Zastąpić wpisy "?" na wartości NA
5. Usunąć wiersze zawierające wpisy NA
6. Podzielić dane na zestaw testowy i uczący
7. zbudować model za pomocą funkcji (glm)
8. dopracować model za pomocą funkcji step
9. Przetestować model na danych testowych (tablica pomyłek)
10. Obliczyć F1-score dla opracowanego modelu