Raport - Analiza spożycia alkoholu wśród uczniów

Problem: Zbudowanie modeli do przewidywania ocen oraz spożycia alkoholu na podstawie danych ze zbioru.

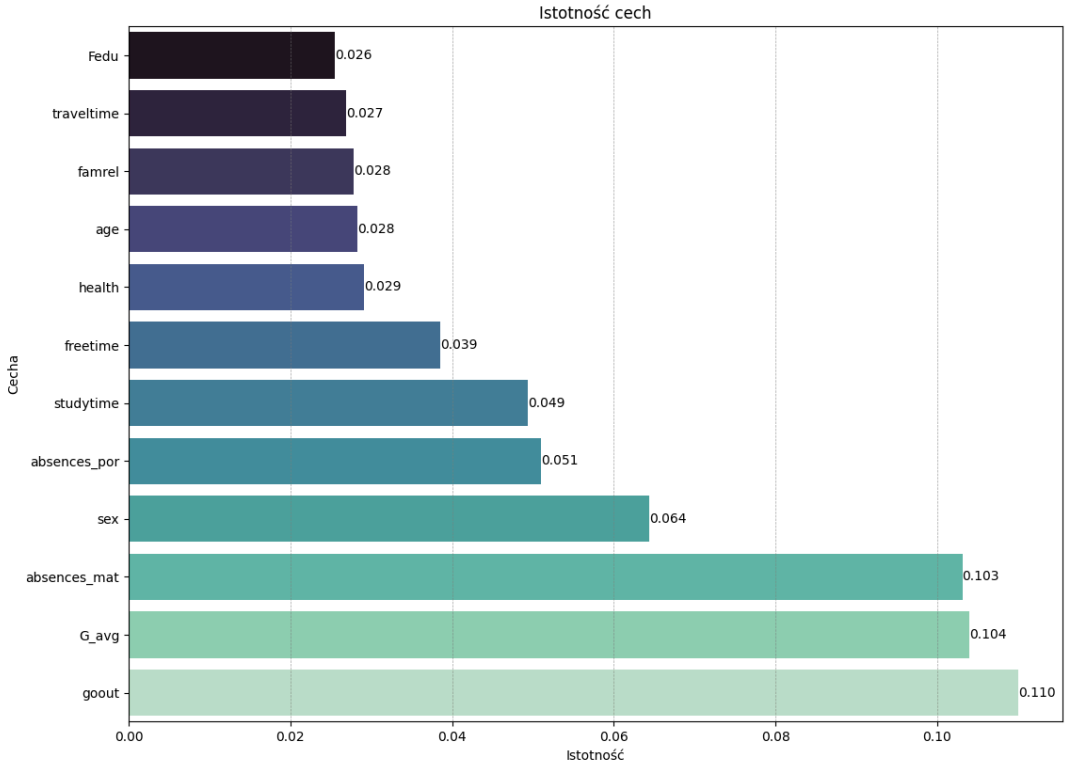
Zbiór danych: Wykorzystano zbiór danych z platformy kaggle. Dane zostały uzyskane w badaniu kursów matematyki i języka portugalskiego w szkole średniej. Zawierają wiele interesujących informacji społecznych, dotyczących płci oraz nauki dotyczącej uczniów.

Opis zastosowanych metod i kryterium wyboru najlepszej metody: Dobór hiperparametrów odbywał się drogą empiryczną. Posłużyliśmy się walidacją krzyżową, porównując wyniki modeli z różnymi wartościami hiperparametrów dla modelu lasu losowego (RandomForestClassifier) z pakietu sklearn.

Opis wyników działania rozwiązania: Kod do projektu znajduje się na githubie. Stworzyliśmy dwa modele:

1. Model przewidujący spożycie alkoholu przez uczniów na podstawie danych.
    - Dokładność modelu wyniosła 78%
    - Błąd średniokwadratowy 0,22
    - Najlepsze hiperparametry {'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'sqrt'}
        - “Max_features: sqrt” oznacza, że w jednym drzewie liczba cech brana pod uwagę jest pierwiastkiem liczby wszystkich cech
        - “min_samples_leaf: 1” oznacza, że minimalna liczba próbek potrzebna do bycia liściem węzła wynosi 1.
        - “min_samples_split: 5” oznacza, że minimalna liczba próbek do bycia węzłem wynosi 5.
        - “ n_estimators: 100” oznacza, że w lesie jest 100 drzew decyzyjnych.

- Istotność cech w modelu poniższa grafika. Najbardziej istotne są wyjścia z domu oraz średnia ocen.



2. Model przewidujący oceny na podstawie danych.
   - Dokładność modelu wyniosła 76%
   - Błąd średniokwadratowy 0,24
   - Najlepsze hiperparametry {'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'sqrt'}
     - “Max_features: sqrt” oznacza, że w jednym drzewie liczba cech brana pod uwagę jest pierwiastkiem liczby wszystkich cech
     - “min_samples_leaf: 2” oznacza, że minimalna liczba próbek potrzebna do bycia liściem węzła wynosi 2.
     - “min_samples_split: 5” oznacza, że minimalna liczba próbek do bycia węzłem wynosi 5.
     - “ n_estimators: 200” oznacza, że w lesie jest 200 drzew decyzyjnych.

- Istotność cech w modelu przedstawia poniższa grafika. Najbardziej istotne są niezaliczenia z matematyki oraz nieobecności na matematyce i portugalskim