Praca domowa 1 - Wstęp do uczenia maszynowego

Oliwia Trzcińska

1 Cel projektu

Głównym celem projektu było zbadanie wpływu różnych parametrów modelu drzewa decyzyjnego na jego jakość predykcyjną.

2 Dane

Używamy zbioru danych (`X.csv`, `y.csv`). Dzielimy go na zbiory treningowy i testowy w proporcji $7:3$, ustawiając wartość parametru `random_state = 320584`.

3 Eksperyment

Eksperyment ma na celu dobór takich parametrów modelu drzewa decyzyjnego, które maksymalizują jego dokładność (przy ustalonym `random_state = 320584`). Aby znaleźć "najlepszy"model będziemy zmieniać parametry:

- kryterium podziału - `'gini'` lub `'entropy'`,
- głębokość drzewa - wartości od 5 do 19,
- minimalna liczba obserwacji w liściu - wartości 5, 10, 15, 20, 25, 30,
- liczba cech do rozważenia podczas szukania najlepszego podziału - `'auto'`, `'sqrt'` lub `'log2'`.

Tworzymy model z każdą kombinacją wartości wymienionych parametrów. Do obliczania dokładności używamy pięciokrotnej kroswalidacji na danych treningowych. Wyznaczamy średnią dokładność modelu i wybieramy te parametry, dla których wynik jest największy.

4 Analiza jakości predykcyjnej modelu

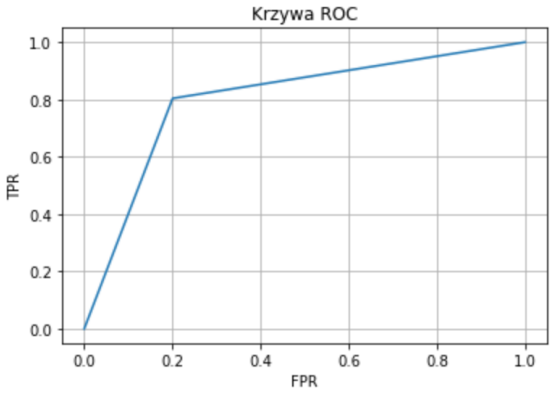
Po przeprowadzeniu wyżej opisanego eksperymentu podejmujemy decyzję o wyborze parametrów: `criterion = 'entropy'`, `max_depth = 16`, `min_samples_leaf = 5`, `max_features = 'auto'`, `random_state = 320584`.

Wykorzystując je trenujemy model na całym zbiorze treningowym. Następnie wykonujemy predykcję na zbiorze testowym i oceniamy jakość modelu.

Otrzymaliśmy następujące wyniki:

- macierz pomyłek: $\begin{bmatrix} 2408 & 604 \\ 586 & 2402 \end{bmatrix}$,
- dokładność: 0.8016666666666666,
- czułość: 0.8038821954484605,
- precyzja: 0.7990685296074518,

- krzywa ROC:



,

- wartość AUC: 0.8016754934745622.