PROJEKT 1

Autorzy:
Kinga Florek, Michał Worsowicz

Inteligencja obliczeniowa w analizie danych cyfrowych

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

1 Krótki opis zestawu danych

Wybrany zestaw danych: Haberman's Survival Data Set
Algorytm: Lasy drzew decyzyjnych
Optymalizowany parametr: Sensitivity (czułość)

Opis zestawu danych:
Zbiór danych zawiera informacje o przypadkach z badań przeprowadzonych w latach 1958-1970 w szpitalu Billings na Uniwersytecie w Chicago na temat przeżywalności pacjentów, którzy przeszli operację raka piersi.

Liczba instancji: 306
Liczba cech: 4 (włącznie z atrybutami klasy)

Informacje o cechach:

1. Wiek pacjenta w momencie operacji (wartość numeryczna),
2. Rok operacji pacjenta (rok - 1900, wartość numeryczna),
3. Liczba pozytywnych wyników biopsji węzłów pachowych (wartość numeryczna),
4. Przeżywalność po operacji (atrybut klasy)
    - 1 = pacjent przeżył 5 lat lub dłużej,
    - 2 = pacjent zmarł w przeciągu 5 lat

Rodziaj problemu: Klasyfikacja
Brakujące dane: Nie

2 Krótki opis wybranej metody uczenia maszynowego

Metoda lasów drzew decyzyjnych polega na stworzeniu wielu indywidualnych drzew decyzyjnych, gdzie każde drzewo jest zbudowane na innym, losowym podzbiorze zbioru treningowego. Końcowa decyzja podejmowana jest na podstawie głosowania większościowego nad klasami, które wskazały poszczególne

drzewa decyzyjne.

Klasyfikator $RandomForestClassifier$ może przyjmować wiele parametrów, natomiast zdecydowaliśmy się na optymalizowanie parametru $n\_estimators$ oraz max_depth, które odpowiadają za ilość drzew w lesie i głębokość drzewa.

3 Sposób wyboru zbioru testowego

Podzieliliśmy nasz zbiór danych na zbiór treningowy i testowy w taki sposób, że 70% zbioru to zbiór treningowy, a 30% zbiór testowy. Dodatkowo przed podziałem zastosowaliśmy pomieszanie danych, poprzez zdefiniowanie $random\_state$ na 10. Podaliśmy liczbę całkowitą, która będzie użyta w generatorze liczb losowych podczas randomizacji.

4 Opis działania metody wyboru hiperparametrów

Użytą przez nas metodą wyboru optymalnych hiperparametrów jest Grid Search. Polega ona na doborze parametrów w ręcznie zdefiniowanych zakresach, których kombinacje zwizualizować można właśnie na siatce, w celu odnalezienia najlepszej ich kombinacji. W celu implementacji tej metody wykorzystujemy funkcję GridSearchCV (grid search cross validation), której argumenty określają:

- zbiór danych, który wykorzystujemy,
- parametry, które optymalizujemy,
- krotność walidacji krzyżowej (10 w naszym przypadku)