Ćwiczenie 5. Klasyfikacja – indukcja drzew decyzyjnych

Wykonać następujące polecenia, używając zbiór danych i algorytm wskazany w pliku „przydziały.docx”.

1. (2 pkt) Przeprowadzić indukcję drzewa decyzyjnego bez użycia języka R dla małego zbioru danych wyodrębnionego ze zbioru oryginalnego. W celu utworzenia „małego zbioru danych” należy wybrać (inteligentnie!!!):
   - 3 lub 4 atrybuty jako zmienne opisujące,
   - 10 rerkordów jako zbiór trenujący,
   - 5 rekordów jako zbiór testowy.

   Należy pokazać indukcję drzewa decyzyjnego (podobnie jak to zostało zrobione w prezentacji). Narysować i przetestować utworzone drzewo.

2. (2 pkt) Wykorzystując cały zbiór danych przeprowadzić walidację krzyżową z użyciem k podzbiorów (za pomocą skryptu w języku R) dla różnych wartości parametru algorytmu: **`minCases`** (dla algorytmu C4.5) lub **`minsplit`** (dla algorytmu CART). Zastosować otrzymany model do klasyfikacji danych zarówno ze zbioru testowego jak i ze zbioru trenującego. Przedstawić otrzymane wyniki na wykresie liniowym, tzn. poziomy błędów (dla danych testowych i trenujących) względem wartości parametru algorytmu. Przedyskutować otrzymane wyniki. Skomentować nadmierne dopasowanie, jeżeli zostało zaobserwowane.

   Wartości parametrów:
   - liczba podzbiorów do walidacji krzyżowej k = 10;
   - parametr `minCases` dla algorytmu C4.5 powinien przyjmować wartości od 1 do 40
   - parametr `minsplit` dla algorytmu CART powinien przyjmować wartości od 2 to 80
   - pozostałe parametry, **`CF = 1.0`** (C4.5) i **`cp = 0.0`** (CART).

     ```
     C5.0(x=…,y=…,control=C5.0Control(minCases = …,CF=1.0))
     rpart(…, data = …, method =…, control=rpart.control(minsplit=… cp=0.0))
     ```

3. (1pkt) Dokonaj klasyfikacji nowych rekordów danych, podanych w pliku „new records.txt” przy użyciu drzewa decyzyjnego (utworzonego na podstawie całego zbioru danych) i algorytmu knn (działającego z całym zbiorem danych). Porównaj wyniki.