DATA MINING 2014 **Laboratorium 8 (METODY ŁĄCZENIA KLASYFIKATORÓW)**

8.1
Zbiór danych skoringowych *kredit.asc* zawiera informację o klientach banku. Zmienna **kredit** zawiera informację czy klient spłacił kredyt (wartość 1) czy nie spłacił kredytu (wartość 0). Każdy z klientów jest opisany w terminach punktów skoringowych. Dokładny opis danych znajduje się na stronie `http://www.ipipan.eu/~teisseyrep/DM/DANE/kredit.htm`. Podzielić zbiór danych na dwa zbiory: treningowy (wylosować 2/3 obserwacji) oraz testowy (pozostałe obserwacje).

a) Dopasować drzewo klasyfikacyjne na podstawie próby treningowej. Przyjąć domyślne wartości parametrów. Przedstawić graficznie drzewo i wypisać jego strukturę. Dokonać estymacji prawdopodobieństwa poprawnej klasyfikacji na próbie testowej.

b) Zaimplementuj metodę **bagging** dla drzew klasyfikacyjnych. Przyjąć domyślne wartości parametrów dla każdego drzewa. Rozważyć dwa przypadki: liczba pseudoprób jest równa 25 oraz 100. Dla każdgo przypadku obliczyć procent poprawnych klasyfikacji na zbiorze testowym i porównać wyniki.

8.2
Dane *pima-indians-diabetes.data* zwierają informację o zapadalności na cukrzycę wśród 768 Indianek w wieku co najmniej 21 lat, z plemienia Prima. Każda z Indianek jest opisywana 8 zmiennymi. Ostatnia zmienna w zbiorze przyjmuje dwie wartości: 1 oznacza że wystąpiła cukrzyca, natomiast 0 oznacza brak cukrzycy. Dokładny opis danych na stronie: `http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes`.
Podzielić zbiór danych na dwa zbiory: treningowy (wylosować 500 obserwacji) oraz testowy (pozostałe obserwacje).

a) Dopasować drzewo klasyfikacyjne na podstawie próby treningowej. Przyjąć wartości parametrów `cp=0.02`, `minsplit=5` . Przedstawić graficznie drzewo i wypisać jego strukturę. Dokonać estymacji prawdopodobieństwa poprawnej klasyfikacji na próbie testowej.

b) Dokonać innego podziału danych na część treningową i testową, wykonać polecenie z punktu (a) i porównać wyniki. Które zmienne w zbiorze możemy unać za istotne?

c) Skonstruować rodzinę klasyfikatorów na próbie treningowej używając metody **boosting** i algorytmu **AdaBoost**. Przyjąć następujące wartości parametrów: `cp=0.02`, `minsplit=5`, `maxdepth=3`. Niech liczba generowanych pseudoprób będzie równa 50. Obliczyć procent poprawnych klasyfikacji na zbiorze testowym i porównać wyniki z tymi otrzymanymi dla pojedynczego drzewa.

8.3
Dane w pliku *agaricus-lepiota.data* opisują różne rodzaje grzybów. Zbiór zawiera 8124 obserwacje oraz 23 atrybuty (dyskretne!). Zmienna grupująca **V1** przyjmuje dwie wartości: **V1="e"** (grzyb jadalny) oraz **V1="p"** (grzyb trujący lub niejadalny). Celem analizy jest modelowanie zależności cechy "przydatność do spożycia" od innych cech grzybów. Dokładny opis danych znajduje się na stronie `http://archive.ics.uci.edu/ml/datasets/mushroom`.

a) Dokonaj podziału zbioru danych na: próbę uczącą (50 procent losowo wybranych obserwacji) i testową (pozostałe obserwacje).

b) Porównaj działanie następujących metod szacując prawdopodobieństwo poprawnej klasyfikacji na zbiorze testowym.

1. pojedyncze drzewo klasyfikacyjne (dokonaj wyboru optymalnego drzewa stosując kryte-

rium kosztu-złożoności),

2. metoda Bagging (z parametrami domyślnymi),
3. metoda Boosting (z parametrami domyślnymi),
4. Las Losowy (z parametrami domyślnymi). Jakie są wartości podstawowych parametrów?

c) Powtórz punkt (b) dla innego podziału zbioru.

d) Wypisz strukturę wybranego drzewa "należacego do lasu losowego" używajac funkcji `getTree`.

e) Dla Lasu Losowego dokonaj oceny istotności zmiennych obliczając średnią zmianę indeksu Giniego dla każdej zmiennej (funkcje: `importance` oraz `varImpPlot`).