

### 6.1

Zbiór danych *SAheart.data* (South African Heart Disease) zawiera dane dotyczące zapadalności na zawał serca wśród mężczyzn pomiędzy 15 a 64 rokiem życia. Zmienna **chd** oznacza że wystąpił (wartość 1) lub nie wystąpił (wartość 0) zawał serca. Dokładny opis danych znajduje się w pliku *SAheart.info*.

- a) Dopasować drzewo klasyfikacyjne. Przyjąć czynnik złożoności  $cp=0.01$  oraz parametr  $minsplit=5$  (minimalna liczba elementów, która musi być w węźle, aby jeszcze dokonywać w nim podziału).
- b) Wyrysuj wykres przedstawiający to drzewo.
- c) Na podstawie drzewa zbudowanego w punkcie (a) dokonać predykcji dla obserwacji mającej wartości zmiennych równe wartościom średnich zmiennych ze zbioru na podstawie których skonstruowano drzewo.
- d) Wybierz drzewo optymalne w oparciu o kryterium kosztu- złożoności, stosując regułę 1SE.

### 6.2

Dane *earthquake.txt* dotyczą klasyfikacji wstrząsów na podstawie danych sejsmologicznych. Zmienna grupująca **popn** opisuje rodzaj wstrząsu: może to być trzęsienie ziemi (wartość *equake*) lub wybuch nuklearny (wartość *explosn*). Każdy wstrząs jest opisywany przez dwie zmienne objaśniające: **body** (magnituda fali głębokiej) i **surface** (magnituda fali powierzchniowej). Celem analizy jest identyfikacja rodzaju wstrząsu na podstawie zmiennych sejsmologicznych.

- a) Wykonać wykres rozproszenia dla zmiennych **body** i **surface**. Obiekty z klasy *equake* oznaczyć literą "Q", a obiekty z klasy *explosn* literą "X".
- b) Zaprezentować graficznie sposób w jaki dokonujemy klasyfikacji obiektów za pomocą drzewa klasyfikacyjnego. Rozważyć dwa przypadki: parametr  $minsplit=15$  oraz  $minsplit=5$ . Wartość parametru  $cp$  pozostawić jako domyślną.

### 6.3

Dane *fitness.txt* dotyczą parametrów wydolnościowych mężczyzn zmierzonych podczas biegu na 1.5 mili. W zbiorze znajdują się następujące zmienne:

- **Oxygen**- intensywność poboru tlenu (w ml na kg wagi ciała i minutę),
- **Age**- wiek (w latach),
- **Weight**- waga (w kg.),
- **RunTime**- czas przebiegnięcia 1.5 mili (w minutach),
- **RestPulse**- puls spoczynkowy,
- **RunPulse**- puls podczas biegu,
- **MaxPulse**- maksymalny puls podczas biegu.

Zmienną objaśniającą jest **Oxygen**.

- a) Dopasuj drzewo regresyjne używając wszystkich atrybutów. Przyjmij parametry:  $cp=0.01$ ,  $minsplit=2$ . Sporządź wykres przedstawiający drzewo.
- b) Na podstawie drzewa dopasowanego w punkcie (a) odpowiedz na pytanie dla jakiego biegacza

pobór tlenu jest oceniany jako największy?

c) Dokonaj prognozy na podstawie skonstruowanego drzewa wartości **Oxygen** dla obserwacji  $x_0$ , której współrzędne są równe medianom zmiennych ze zbioru danych. Odczytaj również wartość prognozowaną z wykresu drzewa.

d) Dokonaj wybory optymalnego poddrzewa stosując kryterium kosztu złożoności oraz regułę 1SE.

e) Dopasuj drzewo na podstawie dwóch zmiennych: **RunTime** oraz **Age** z parametrami  $cp=0.02$ ,  $minsplitlevel=2$ . Przedstaw graficznie predykcje zmiennej **Oxygen**.

