4.1
Dane `Cars93(MASS)` zawierają informację o samochodach. Dokładny opis danych znajduje się w pomocy `R`. Zmienna **Type** określa rodzaj samochodu: Small, Sporty, Compact, Midsize, Large, Van. Utworzyć nową zmienną grupującą o nazwie **Typ** przyjmującą cztery wartości:

- **Typ="D"** (duży) gdy **Type="Van"** lub **Type="Large"**
- **Typ="SR"** (średni) gdy **Type="Compact"** lub **Type="Medium"**
- **Typ="M"** (mały) gdy **Type="Small"**
- **Typ="SP"** (sportowy) gdy **Type="Sporty"**

Celem analizy jest identyfikacja rodzaju samochodu na podstawie zmiennych określających jego parametry.

a) Dopasować "duże" drzewo klasyfikacyjne używając zmiennych **Length**, **Weight**, **EngineSize**, **Horsepower**, **RPM** jako zmiennych objaśniających. Przyjąć czynnik złożoności `cp=0.0001` oraz parametr `minsplit=5` (minimalna liczba elementów, która musi być w węźle, aby jescze dokonywać w nim podziału).

- Wypisz strukturę otrzymanego drzewa.
- Wyrysuj wykres przedstawiający to drzewo.

b) Dopasować drzewo klasyfikacyjne używając tych samych parametrów co w punkcie (a), zastępując indeks Giniego (default) entropią (`parms=list(split="information")`). Porównaj wyniki.

c) Na podstawie drzewa zbudowanego w punkcie (a) dokonać predykcji dla obserwacji mającej wartości zmiennych równe wartościom średnich zmiennych ze zbioru na podstawie których skonsruowano drzewo.

d) Wybierz drzewo optymalne w oparciu o kryterium kosztu- złożoności, stosując regułę 1SE.

e) Dokonać oceny klasyfikatorów skonstruowanych na podstawie drzew z punktu (a) oraz punktu (d) szacując błąd klasyfikacji metodą walidacji krzyżowej typu "leave one out". Wybierz drzewo optymalne w oparciu o kryterium kosztu- złożoności, stosując regułę 1SE.

f) Wykorzystaj funkcję `tune.rpart (e1071)` do wyboru optymalnej wartości parametru `minsplit=5,10,15`.

4.2
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 funkcji

`lda`.

c) 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.

4.3
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) Dopasować "duże" drzewo klasyfikacyjne używając wszystkich zmiennych objaśniających. Przyjąć czynnik złożoności `cp=0.0001` oraz parametr `minsplit=5`.

- Wypisz strukturę otrzymanego drzewa.
- Wyrysuj wykres przedstawiający to drzewo

(zwróć uwagę na sposób kodowania zmiennych o wartościach dyskretnych w opisie drzewa).

b) Dokonaj zmiany wartości parametrów `cp=0.0001, 0.01, 0.5` oraz `minsplit=5,50`. Porównaj wyniki.

c) Wybierz drzewo optymalne w oparciu o kryterium kosztu- złożoności, stosując regułę 1SE.

4.4
Wczytaj dane *iris.data*. Dokładny opis danych znajduje się na stronie `http://archive.ics.uci.edu/ml/datasets/iris`.

a) Dopasować drzewo klasyfikacyjne używając wszystkich zmiennych objaśniających. Przyjąć ustawienia domyślne. Jakie są domyślne wartości parametrów: `minsplit` i `cp`?

- Wypisz strukturę otrzymanego drzewa.
- Wyrysuj wykres przedstawiający to drzewo

b) Oszacować błąd klasyfikacji dla drzewa z punktu (a) stosując walidację krzyżową typu "leave one out".

4.5
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`. Wypisz strukturę drzewa oraz sporządź wykres.

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 $x0$, 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 model liniowy. Porównaj sumę kwadratów rezyduów dla tego modelu z sumą kwadratów rezyduów dla "dużego" drzewa i drzewa przyciętego.

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