Eksploracja danych

true

2019-03-12

Spis treści

Wstęp **5**
O książce . . . 5
Zakres przedmiotu . . . 5
Zakres technik stosowanych w data mining . . . 5
Etapy eksploracji danych . . . 7

1 Import danych **9**
1.1 Przykład . . . 14

2 Przygotowanie danych **17**
2.1 Korekta zbioru danych . . . 17
2.2 Przykład . . . 26

3 Podział metod data mining **35**
3.1 Rodzaje wnioskowania . . . 35
3.2 Modele regresyjne . . . 36
3.3 Modele klasyfikacyjne . . . 37
3.4 Modele grupujące . . . 37

4 Drzewa decyzyjne **39**
4.1 Węzły i gałęzie . . . 39
4.2 Rodzaje reguł podziału . . . 39
4.3 Algorytm budowy drzewa . . . 41
4.4 Kryteria zatrzymania . . . 42
4.5 Reguły podziału . . . 42
4.6 Przycinanie drzewa decyzyjnego . . . 43
4.7 Obsługa braków danych . . . 45
4.8 Zalety i wady . . . 45
4.9 Przykład . . . 45
4.10 Inne algorytmy budowy drzew decyzyjnych implementowane w **R** . . . 54
4.11 Przykład . . . 54

5 Pochodne drzew decyzyjnych **61**
5.1 Bagging . . . 61
5.2 Lasy losowe . . . 71
5.3 Boosting . . . 72

Wstęp

O książce

Niniejsza książka powstała na bazie doświadczeń autora, a głównym jej celem jest przybliżenie czytelnikowi podstaw z dziedziny *Data mining* studentom kierunku *Matematyka* Politechniki Lubelskiej. Będzie łączyć w sobie zarówno treści teoretyczne związane z przedstawianymi etapami eksploracji danych i budową modeli, jak i praktyczne wskazówki dotczące budowy modeli w środowisku **R** (R Core Team, 2018). Podane zostaną również wskazówki, jak raportować wyniki analiz i jak dokonać właściwych ilustracji wyników. Bardzo użyteczny w napisaniu książki były pakiety programu R: **bookdown** (Xie, 2018a), **knitr** (Xie, 2018b) oraz pakiet **rmarkdown** (Allaire et al., 2018).

Zakres przedmiotu

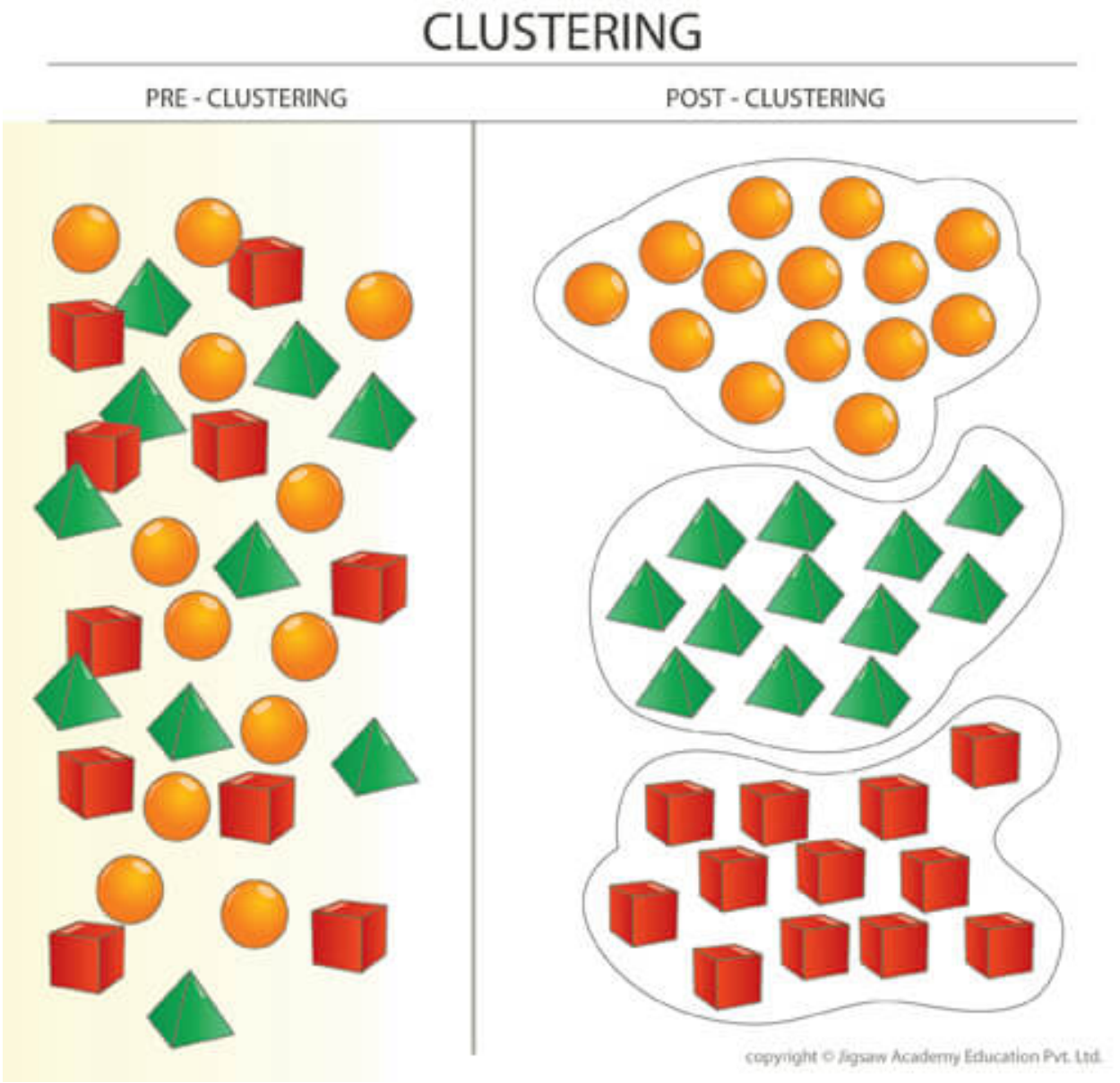
Przedmiot *Eksploracja danych* będzie obejmował swoim zakresem eksplorację i wizualizację danych oraz uczenie maszynowe. Eksploracja danych ma na celu pozyskiwanie i systematyzację wiedzy pochodzącej z danych. Odbywa się ona głównie przy użyciu technik statystycznych, rachunku prawdopodobieństwa i metod z zakresu baz danych. Natomiast uczenie maszynowe, to gałąź nauki (obejmuje nie tylko statystykę, choć to na niej się głównie opiera) dotyczącej budowy modeli zdolnych do rozpoznawania wzorców, przewidywania wartości i klasyfikacji obiektów. Data mining to szybko rosnaca grupa metod analizy danych rozwijana nie tylko przez statystyków ale również przez biologów, genetyków, cybernetyków, informatyków, ekonomistów, osoby pracujace nad rozpoznawaniem obrazów i wiele innych grup zawodowych. W dzisiejszych czasch trudno sobie wyobrazić życie bez sztucznej inteligencji. Towarzyszy ona nam w codziennym, życiu kiedy korzystamy z telefonów komórkowych, wyszukiwarek internetowych, robotów sprzątających, automatycznych samochodów, nawigacji czy gier komputerowych. Lista ta jest niepełna i stale się wydłuża.

href="https://twitter.com/i/status/1091069356367200256">January 31, 2019

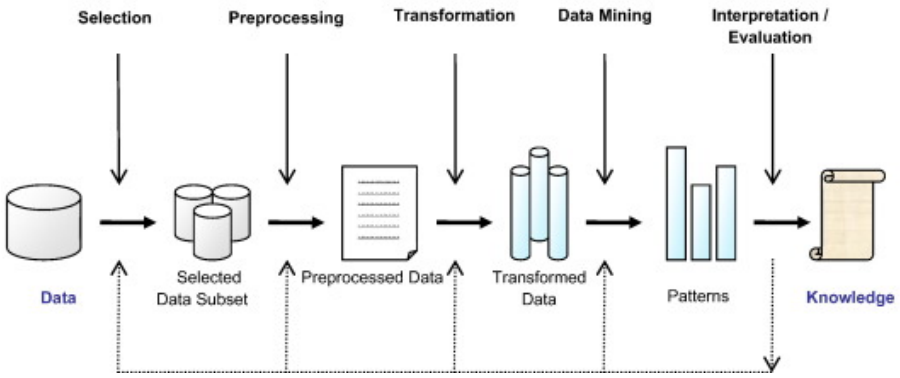
Zakres technik stosowanych w data mining

- statystyka opisowa
- wielowymiarowa analiza danych
- analiza szeregów czasowych
- analiza danych przestrzennych
- reguły asocjacji
- uczenie maszynowe[1], w tym:
  - klasyfikacja
  - predykcja
  - analiza skupień
  - *text mining*
- i wiele innych

[1] ang. *machine learning*



Rysunek 1: Przykład nienadzorowanego uczenia maszynowego. *Źródło:*https://analyticstraining.com/cluster-analysis-for-business/



Rysunek 2: Etapy eksploracji danych (Kavakiotis et al., 2017)

href="https://twitter.com/i/status/1097199751072690176">Ferbruary 17, 2019

Etapy eksploracji danych

1. Czyszczenie danych - polega na usuwaniu braków danych, usuwaniu stałych zmiennych, imputacji braków danych oraz przygotowaniu danych do dalszych analiz.
2. Integracja danych - łączenie danych pochodzących z różnych źródeł.
3. Selekcja danych - wybór z bazy tych danych, które są potrzebne do dalszych analiz.
4. Transformacja danych - przekształcenie i konsolidacja danych do postaci przydatnej do eksploracji.
5. Eksploracja danych - zastosowanie technik wymienionych wcześniej w celu odnalezienia wzorców[2] i zależności.
6. Ewaluacja modeli - ocena poprawności modeli oraz wzorców z nich uzyskanych.
7. Wizualizacja wyników - graficzne przedstawienie odkrytych wzorców.
8. Wdrażanie modeli - zastosowanie wyznaczonych wzorców.

[2]ang. *patterns*

Rozdział 1

Import danych

Środowisko **R** pozwala na import i export plików o różnych rozszerzeniach (`txt, csv, xls, xlsx, sav, xpt, dta`, itd.)[1]. W tym celu czasami trzeba zainstalować pakiety rozszerzające podstawowe możliwości R-a. Najnowsza[2] wersja programu RStudio (v. 1.1.463)[3] pozwala na wczytanie danych z popularnych źródeł za pomocą GUI.

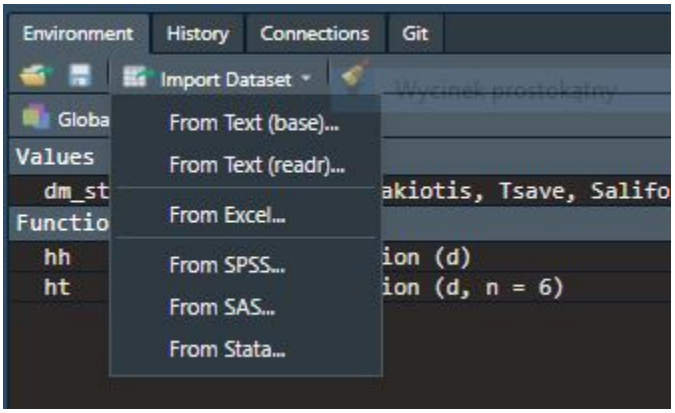
Jeśli dane są zapisane w trybie tekstowym (np. `txt`, `csv`), to wczytujemy je w następujący sposób

```
dane1 <- read.table("data/dane1.txt", header = T)
head(dane1)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
dane2 <- read.csv2("data/dane1.csv", header = T)
head(dane2)
```

[1] póki co nie jest mi znana funkcja pozwalająca na import plików programu Statistica
[2] na dzień 19.02.2019
[3] istnieją rownież nowsze wersje deweloperskie



Rysunek 1.1: Narzędzie do importu plików programu RStudio

	A	B	C	D	E	F
1	Długość kielic	Szerokość	Długość pł	Szerokość	Gatunki	
2	5,1	3,5	1,4	0,2	setosa	
3	4,9	3	1,4	0,2	setosa	
4	4,7	3,2	1,3	0,2	setosa	
5	4,6	3,1	1,5	0,2	setosa	
6	5	3,6	1,4	0,2	setosa	
7	5,4	3,9	1,7	0,4	setosa	
8	-	-	1,4	0,3	setosa	
9	5	3,4	1,5	0,2	BD	
10	4,4	2,9	1,4	0,2	setosa	
11	4,9	3,1	1,5	0,1	setosa	
12	5,4	3,7	1,5	0,2	setosa	
13	4,8	3,4	1,6	0,2	setosa	

Rysunek 1.2: Fragment pliku Excel

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
# funkcja pakietu readr wczytuje plik jako ramkę danych w formacie tibble
# pakiet readr jest częsią większego pakietu tidyverse,
# który został wczytany wczsniej
dane3 <- read_csv2("data/dane1.csv")
dane3
```

```
## # A tibble: 150 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl> <chr>
##  1          5.1         3.5          1.4         0.2 setosa
##  2          4.9         3            1.4         0.2 setosa
##  3          4.7         3.2          1.3         0.2 setosa
##  4          4.6         3.1          1.5         0.2 setosa
##  5          5           3.6          1.4         0.2 setosa
##  6          5.4         3.9          1.7         0.4 setosa
##  7          4.6         3.4          1.4         0.3 setosa
##  8          5           3.4          1.5         0.2 setosa
##  9          4.4         2.9          1.4         0.2 setosa
## 10          4.9         3.1          1.5         0.1 setosa
## # ... with 140 more rows
```

Jeśli dane są przechowywane w pliku Excel (np. `xlsx`), to importujemy je za pomocą funkcji `read_excel` pakietu `readxl`. Domyślnie jest wczytywany arkusz pierwszy ale jeśli zachodzi taka potrzeba, to można ustalić, który arkusz pliku Excel ma być wczytany za pomocą paramteru `sheet`, np. `sheet=3`, co oznacza, że zostanie wczytany trzeci arkusz pliku.

Ponieważ w pliku `dane1.xlsx` braki danych zostały zakodowane znakami `BD` oraz `-`, to należy ten fakt przekazać funkcji, aby poprawnie wczytać braki danych. W przeciwnym przypadku zmienne zawierające braki tak kodowane, będą wczytane jako zmienne znakowe.

```
library(readxl)
dane4 <- read_excel("data/dane1.xlsx", na = c("BD", "-"))
dane4
```

```
## # A tibble: 150 x 5
##    `Długość kielic~ `Szerokość kiel~ `Długość płatka` `Szerokość płat~
##               <dbl>            <dbl>            <dbl>            <dbl>
##  1              5.1              3.5              1.4              0.2
##  2              4.9              3                1.4              0.2
##  3              4.7              3.2              1.3              0.2
##  4              4.6              3.1              1.5              0.2
##  5              5                3.6              1.4              0.2
##  6              5.4              3.9              1.7              0.4
##  7             NA               NA                1.4              0.3
##  8              5                3.4              1.5              0.2
##  9              4.4              2.9              1.4              0.2
## 10              4.9              3.1              1.5              0.1
## # ... with 140 more rows, and 1 more variable: Gatunki <chr>
```

Istniej oczywiście jeszcze wiele innych fomatów danych, charakterystycznych dla programów, w których są traktowane jako domyślne.[4] W szczególny sposób należy zwrócić uwagę na pliki o rozszerzeniu `RData` lub `rda`[5] oraz pliki `rds`. Pliki `rda` służą do przechowywania obiektów programu `R`. Mogą to być pliki danych ale również obiekty graficzne (typu wyniki funkcji `ggplot`), modele (np. wynik funkcji `lm()`), zdefiniowane funkcje i wszystkie inne obiekty, które da się zapisać w środowisku `R`. Ponadto pliki `rda` pozawalają na zapisanie wilu obiektów w jednym pliku. Pliki o rozszerzeniu `rds` mają podobną funkcję z tym, że pozwalają na przechowywanie tylko jednego obiektu.

```
# wszystkie wczytane wcześniej pliki zapisuje w jednym pliku
save(dane1, dane2, dane3, dane4, file = "data/dane.rda")
# plik rda został zapisany
list.files(path = "data/")
```

```
## [1] "algae.csv"    "Analysis.txt" "dane.rda"     "dane1.csv"
## [5] "dane1.txt"    "dane1.xlsx"   "dane4.rds"    "dane4.sav"
```

```
# usuwam dane ze środowiska R
rm(dane1, dane2, dane3, dane4)
# sprawdzam co jest wczytane do R
ls()
```

```
## character(0)
```

```
# wczytuję plik rda
load("data/dane.rda")
# jeszcze raz sprawdzam co jest wczytane do R
ls()
```

```
## [1] "dane1" "dane2" "dane3" "dane4"
```

Zapisując obiekty jako oddzielne pliki, można przy wczytywaniu nadawać im nazwy.

```
rm(dane1, dane2, dane3)
ls()
```

[4] do ich wczytywania stosujemy funkcje pakietu `foreign`
[5] oznaczają to samo

```
## [1] "dane4"
```

```
saveRDS(dane4, file = "data/dane4.rds")
nowe_dane <- readRDS("data/dane4.rds")
nowe_dane
```

```
## # A tibble: 150 x 5
##    `Długość kielic~ `Szerokość kiel~ `Długość płatka` `Szerokość płat~
##               <dbl>            <dbl>            <dbl>            <dbl>
##  1              5.1              3.5              1.4              0.2
##  2              4.9              3                1.4              0.2
##  3              4.7              3.2              1.3              0.2
##  4              4.6              3.1              1.5              0.2
##  5              5                3.6              1.4              0.2
##  6              5.4              3.9              1.7              0.4
##  7             NA               NA                1.4              0.3
##  8              5                3.4              1.5              0.2
##  9              4.4              2.9              1.4              0.2
## 10              4.9              3.1              1.5              0.1
## # ... with 140 more rows, and 1 more variable: Gatunki <chr>
```

Oprócz wielu zalet takiego sposobu importu i eksportu danych jest jedna poważna wada, pliki te można odczytać jedynie za pomocą R. Osobiście polecam stosować do importu i eksportu danych plików w takich formatach, które mogą przeczytać wszyscy. Jak dotąd widać do importu różnych formatów danych potrzebujemy różnych funkcji, czasami nawet z różnych pakietów. Istnieje rozwiązanie tego poroblemu

```
library(rio)
dane1 <- import("data/dane1.txt")
head(dane1)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
dane2 <- import("data/dane1.csv", dec = ",")
# dane1.csv miały , jako znak rozdzielający cechę i mantysę liczb
# dlatego włączamy parametr dec
head(dane2)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
dane3 <- import("data/dane1.xlsx", na=c("BD","-"))
head(dane3)
```

```
##   Długość kielicha Szerokość kielicha Długość płatka Szerokość płatka
## 1              5.1                3.5            1.4               0.2
## 2              4.9                3.0            1.4               0.2
## 3              4.7                3.2            1.3               0.2
```

```
## 4             4.6              3.1           1.5             0.2
## 5             5.0              3.6           1.4             0.2
## 6             5.4              3.9           1.7             0.4
##   Gatunki
## 1  setosa
## 2  setosa
## 3  setosa
## 4  setosa
## 5  setosa
## 6  setosa
```

```
dane4 <- import("data/dane4.rds")
dane4
```

```
## # A tibble: 150 x 5
##    `Długość kielic~ `Szerokość kiel~ `Długość płatka` `Szerokość płat~
##               <dbl>            <dbl>            <dbl>            <dbl>
##  1              5.1              3.5              1.4              0.2
##  2              4.9              3                1.4              0.2
##  3              4.7              3.2              1.3              0.2
##  4              4.6              3.1              1.5              0.2
##  5              5                3.6              1.4              0.2
##  6              5.4              3.9              1.7              0.4
##  7             NA               NA                1.4              0.3
##  8              5                3.4              1.5              0.2
##  9              4.4              2.9              1.4              0.2
## 10              4.9              3.1              1.5              0.1
## # ... with 140 more rows, and 1 more variable: Gatunki <chr>
```

Lista możliwości jaką daje nam pakiet `rio` (hong Chan and Leeper, 2018) jest niemal nieograniczona:[6]

- Comma-separated data (.csv), using fread or, if fread = FALSE, read.table with row.names = FALSE and stringsAsFactors = FALSE
- Pipe-separated data (.psv), using fread or, if fread = FALSE, read.table with sep = '|', row.names = FALSE and stringsAsFactors = FALSE
- Tab-separated data (.tsv), using fread or, if fread = FALSE, read.table with row.names = FALSE and stringsAsFactors = FALSE
- SAS (.sas7bdat), using read_sas.
- SAS XPORT (.xpt), using read_xpt or, if haven = FALSE, read.xport.
- SPSS (.sav), using read_sav. If haven = FALSE, read.spss can be used.
- Stata (.dta), using read_dta. If haven = FALSE, read.dta can be used.
- SAS XPORT (.xpt), using read.xport.
- SPSS Portable Files (.por), using read_por.
- Excel (.xls and .xlsx), using read_excel. Use which to specify a sheet number. For .xlsx files, it is possible to set readxl = FALSE, so that read.xlsx can be used instead of readxl (the default).
- R syntax object (.R), using dget
- Saved R objects (.RData,.rda), using load for single-object .Rdata files. Use which to specify an object name for multi-object .Rdata files. This can be any R object (not just a data frame).
- Serialized R objects (.rds), using readRDS. This can be any R object (not just a data frame).
- Epiinfo (.rec), using read.epiinfo
- Minitab (.mtp), using read.mtp
- Systat (.syd), using read.systat
- "XBASE" database files (.dbf), using read.dbf
- Weka Attribute-Relation File Format (.arff), using read.arff

[6] fragment pliku `help` funkcji `import`

- Data Interchange Format (.dif), using read.DIF
- Fortran data (no recognized extension), using read.fortran
- Fixed-width format data (.fwf), using a faster version of read.fwf that requires a widths argument and by default in rio has stringsAsFactors = FALSE. If readr = TRUE, import will be performed using read_fwf, where widths should be: NULL, a vector of column widths, or the output of fwf_empty, fwf_widths, or fwf_positions.
- gzip comma-separated data (.csv.gz), using read.table with row.names = FALSE and stringsAsFactors = FALSE
- CSVY (CSV with a YAML metadata header) using read_csvy.
- Feather R/Python interchange format (.feather), using read_feather
- Fast storage (.fst), using read.fst
- JSON (.json), using fromJSON
- Matlab (.mat), using read.mat
- EViews (.wf1), using readEViews
- OpenDocument Spreadsheet (.ods), using read_ods. Use which to specify a sheet number.
- Single-table HTML documents (.html), using read_html. The data structure will only be read correctly if the HTML file can be converted to a list via as_list.
- Shallow XML documents (.xml), using read_xml. The data structure will only be read correctly if the XML file can be converted to a list via as_list.
- YAML (.yml), using yaml.load
- Clipboard import (on Windows and Mac OS), using read.table with row.names = FALSE
- Google Sheets, as Comma-separated data (.csv)

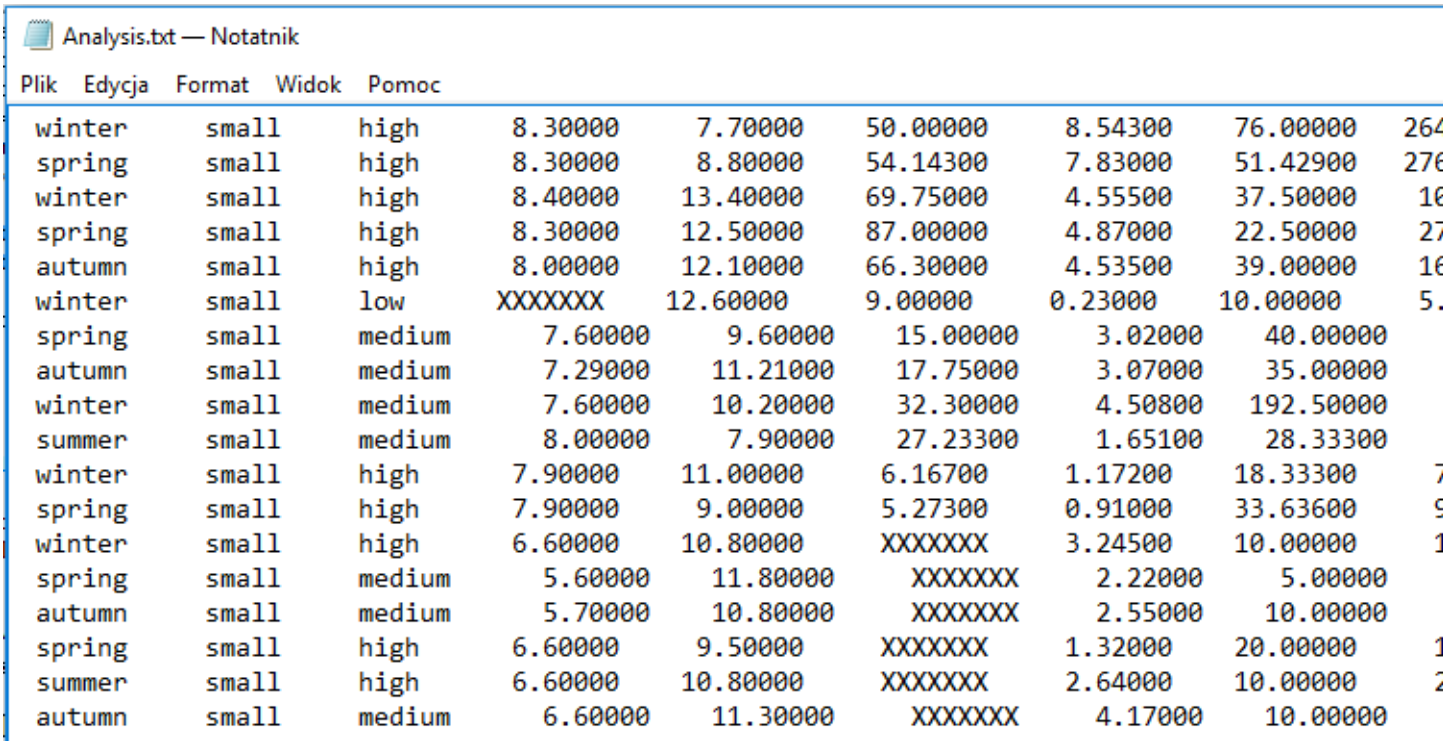
1.1 Przykład

Poniższa ilustracja przedstawia fragment pliku danych `Analysis.txt` zawierającego pewne błędy, które należy naprawić na etapie importu danych. Po pierwsze brakuje w nim nazw zmiennych[7]. Poszczególne kolumny nazywają się następująco: `season`, `size`, `speed`, `mxPH`, `mnO2`, `Cl`, `NO3`, `NH4`, `oPO4`, `PO4`, `Chla`, `a1`, `a2`, `a3`, `a4`, `a5`, `a6`, `a7`. Naszym zadaniem jest import tego pliku z jednoczesną obsługą braków[8] oraz nadaniem nagłówków kolumn. Plik `Analisis.txt` jest umieszczony w kagalogu `data/`. Z racji, że plik dotyczy glonów, to dane zapiszemy pod nazwą `algae`.

```
algae <- import('data/Analysis.txt', header=F,
                dec='.',
                col.names=c('season','size','speed','mxPH','mnO2','Cl',
                            'NO3','NH4','oPO4','PO4','Chla','a1','a2',
                            'a3','a4','a5','a6','a7'),
                na.strings=c('XXXXXXX'))
head(algae)
```

```
##   season  size  speed mxPH mnO2     Cl    NO3     NH4    oPO4     PO4 Chla
## 1 winter small medium 8.00  9.8 60.800  6.238 578.000 105.000 170.000 50.0
## 2 spring small medium 8.35  8.0 57.750  1.288 370.000 428.750 558.750  1.3
## 3 autumn small medium 8.10 11.4 40.020  5.330 346.667 125.667 187.057 15.6
## 4 spring small medium 8.07  4.8 77.364  2.302  98.182  61.182 138.700  1.4
## 5 autumn small medium 8.06  9.0 55.350 10.416 233.700  58.222  97.580 10.5
## 6 winter small   high 8.25 13.1 65.750  9.248 430.000  18.250  56.667 28.4
##     a1   a2   a3  a4   a5   a6  a7
## 1  0.0  0.0  0.0 0.0 34.2  8.3 0.0
## 2  1.4  7.6  4.8 1.9  6.7  0.0 2.1
## 3  3.3 53.6  1.9 0.0  0.0  0.0 9.7
## 4  3.1 41.0 18.9 0.0  1.4  0.0 1.4
```

[7] choć nie widać tego na rysunku
[8] braki danych są zakodowane przez XXXXXXX



Rysunek 1.3: Fragment pliku danych Analisis.txt

```
## 5  9.2  2.9  7.5 0.0  7.5  4.1 1.0
## 6 15.1 14.6  1.4 0.0 22.5 12.6 2.9
```

```
summary(algae)
```

```
##     season              size              speed               mxPH
##  Length:200         Length:200         Length:200         Min.   :5.600
##  Class :character   Class :character   Class :character   1st Qu.:7.700
##  Mode  :character   Mode  :character   Mode  :character   Median :8.060
##                                                           Mean   :8.012
##                                                           3rd Qu.:8.400
##                                                           Max.   :9.700
##                                                           NA's   :1
##       mnO2               Cl               NO3               NH4
##  Min.   : 1.500   Min.   :  0.222   Min.   : 0.050   Min.   :    5.00
##  1st Qu.: 7.725   1st Qu.: 10.981   1st Qu.: 1.296   1st Qu.:   38.33
##  Median : 9.800   Median : 32.730   Median : 2.675   Median :  103.17
##  Mean   : 9.118   Mean   : 43.636   Mean   : 3.282   Mean   :  501.30
##  3rd Qu.:10.800   3rd Qu.: 57.824   3rd Qu.: 4.446   3rd Qu.:  226.95
##  Max.   :13.400   Max.   :391.500   Max.   :45.650   Max.   :24064.00
##  NA's   :2        NA's   :10        NA's   :2        NA's   :2
##       oPO4              PO4               Chla               a1
##  Min.   :  1.00   Min.   :  1.00   Min.   :  0.200   Min.   : 0.00
##  1st Qu.: 15.70   1st Qu.: 41.38   1st Qu.:  2.000   1st Qu.: 1.50
##  Median : 40.15   Median :103.29   Median :  5.475   Median : 6.95
##  Mean   : 73.59   Mean   :137.88   Mean   : 13.971   Mean   :16.92
##  3rd Qu.: 99.33   3rd Qu.:213.75   3rd Qu.: 18.308   3rd Qu.:24.80
##  Max.   :564.60   Max.   :771.60   Max.   :110.456   Max.   :89.80
```

```
##  NA's   :2        NA's   :2        NA's   :12
##        a2               a3               a4               a5
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 3.000   Median : 1.550   Median : 0.000   Median : 1.900
##  Mean   : 7.458   Mean   : 4.309   Mean   : 1.992   Mean   : 5.064
##  3rd Qu.:11.375   3rd Qu.: 4.925   3rd Qu.: 2.400   3rd Qu.: 7.500
##  Max.   :72.600   Max.   :42.800   Max.   :44.600   Max.   :44.400
##
##        a6               a7
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 1.000
##  Mean   : 5.964   Mean   : 2.495
##  3rd Qu.: 6.925   3rd Qu.: 2.400
##  Max.   :77.600   Max.   :31.600
##
```

```
export(algae, file = "data/algae.csv")
```

Rozdział 2

Przygotowanie danych

2.1 Korekta zbioru danych

Dane, które importujemy z zewnętrznego źródła najczęściej nie spełniają formatów obowiązujących w **R**. Często zmienne zawierają niedopuszczalne znaki szczególne, odstępy w nazwach, powtórzone nazwy kolumn, nazwy zmiennych zaczynające się od liczby, czy puste wiersze lub kolumny. Przed przystąpieniem do analizy zbioru należy rozważyć ewentualne poprawki nazw zmiennych, czy usunięcie pustych kolumn i wierszy. Niektórych czynności można dokonać już na etapie importu danych, stosując pewne pakiety oraz nowe funkcjonalności środowiska **RStudio**. W większości przypadków uchroni nas to od żmudnego przekształcania typów zmiennych. Oczywiście wszystkie te czynności czyszczenia danych można również dokonać już po imporcie danych, za pomocą odpowiednich komend **R**.

```
## przykładowe niepożądane nazwy zmiennych
test_df <- as.data.frame(matrix(rnorm(18),ncol = 6))
names(test_df) <- c("hIgHlo", "REPEAT VALUE", "REPEAT VALUE",
                    "% successful (2009)",  "abc@!*", "")
test_df
```

```
##       hIgHlo REPEAT VALUE REPEAT VALUE % successful (2009)     abc@!*
## 1  0.5766562   -1.0119405   -2.2283192           0.9983100  0.4401545
## 2 -1.0209955   -1.9974551    0.5367048          -0.2359354 -0.7866444
## 3 -1.1346090    0.9599928    0.5118476          -1.1698186 -0.1328684
##
## 1  1.2451586
## 2 -0.2748401
## 3 -1.0657506
```

```
## do poprawy nazw zmiennych użyjemy funkcji make.names
names(test_df) <- make.names(names(test_df))
test_df
```

```
##       hIgHlo REPEAT.VALUE REPEAT.VALUE X..successful..2009.     abc...
## 1  0.5766562   -1.0119405   -2.2283192            0.9983100  0.4401545
## 2 -1.0209955   -1.9974551    0.5367048           -0.2359354 -0.7866444
## 3 -1.1346090    0.9599928    0.5118476           -1.1698186 -0.1328684
##            X
## 1  1.2451586
## 2 -0.2748401
## 3 -1.0657506
```

Efekt końcowy choć skuteczny to nie jest zadowalający. Czyszczenia nazw zmiennych można też dokonać stosując funkcję `clean_names` pakietu **janitor** (Firke, 2018). Pozwala on również na usuwanie pustych wierszy i kolumn, znajdowanie zduplikowanych rekordów, itp.

```
library(janitor)
test_df %>% # aby na stałe zmienić nazwy zmiennych trzeba podstawienia
    clean_names()
```

```
##     h_ig_hlo repeat_value repeat_value_2 x_successful_2009        abc
## 1  0.5766562   -1.0119405     -2.2283192         0.9983100  0.4401545
## 2 -1.0209955   -1.9974551      0.5367048        -0.2359354 -0.7866444
## 3 -1.1346090    0.9599928      0.5118476        -1.1698186 -0.1328684
##            x
## 1  1.2451586
## 2 -0.2748401
## 3 -1.0657506
```

```
# przykładowe dane
x <- data.frame(w1=c(1,4,2,NA),w2=c(NA,2,3,NA), w3=c(1,NA,1,NA))
x
```

```
##   w1 w2 w3
## 1  1 NA  1
## 2  4  2 NA
## 3  2  3  1
## 4 NA NA NA
```

```
x %>% remove_empty("rows")
```

```
##   w1 w2 w3
## 1  1 NA  1
## 2  4  2 NA
## 3  2  3  1
```

2.1.1 Identyfikacja braków danych

Zanim usuniemy jakiekolwiek braki w zbiorze, powinniśmy je najpierw zidentyfikować, określić ich charakter, a dopiero potem ewentualnie podjąć decyzję o uzupełnianiu braków.

```
algae <- rio::import("data/algae.csv")

# najprościej jest wywołać summary
summary(algae)
```

```
##     season              size              speed                mxPH
##  Length:200         Length:200         Length:200         Min.   :5.600
##  Class :character   Class :character   Class :character   1st Qu.:7.700
##  Mode  :character   Mode  :character   Mode  :character   Median :8.060
##                                                           Mean   :8.012
##                                                           3rd Qu.:8.400
##                                                           Max.   :9.700
##                                                           NA's   :1
##       mnO2              Cl               NO3              NH4
##  Min.   : 1.500   Min.   :  0.222   Min.   : 0.050   Min.   :    5.00
##  1st Qu.: 7.725   1st Qu.: 10.981   1st Qu.: 1.296   1st Qu.:   38.33
##  Median : 9.800   Median : 32.730   Median : 2.675   Median :  103.17
##  Mean   : 9.118   Mean   : 43.636   Mean   : 3.282   Mean   :  501.30
```

```
##  3rd Qu.:10.800   3rd Qu.: 57.824   3rd Qu.: 4.446   3rd Qu.:  226.95
##  Max.   :13.400   Max.   :391.500   Max.   :45.650   Max.   :24064.00
##  NA's   :2        NA's   :10        NA's   :2        NA's   :2
##       oPO4             PO4              Chla              a1
##  Min.   :  1.00   Min.   :  1.00   Min.   :  0.200   Min.   : 0.00
##  1st Qu.: 15.70   1st Qu.: 41.38   1st Qu.:  2.000   1st Qu.: 1.50
##  Median : 40.15   Median :103.29   Median :  5.475   Median : 6.95
##  Mean   : 73.59   Mean   :137.88   Mean   : 13.971   Mean   :16.92
##  3rd Qu.: 99.33   3rd Qu.:213.75   3rd Qu.: 18.308   3rd Qu.:24.80
##  Max.   :564.60   Max.   :771.60   Max.   :110.456   Max.   :89.80
##  NA's   :2        NA's   :2        NA's   :12
##        a2               a3               a4               a5
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 3.000   Median : 1.550   Median : 0.000   Median : 1.900
##  Mean   : 7.458   Mean   : 4.309   Mean   : 1.992   Mean   : 5.064
##  3rd Qu.:11.375   3rd Qu.: 4.925   3rd Qu.: 2.400   3rd Qu.: 7.500
##  Max.   :72.600   Max.   :42.800   Max.   :44.600   Max.   :44.400
##
##        a6               a7
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 1.000
##  Mean   : 5.964   Mean   : 2.495
##  3rd Qu.: 6.925   3rd Qu.: 2.400
##  Max.   :77.600   Max.   :31.600
##
```

```
## wyświetl niekompletne wiersze
algae[!complete.cases(algae),] %>% head()
```

```
##    season  size  speed mxPH mnO2   Cl   NO3 NH4 oPO4 PO4 Chla   a1  a2  a3
## 28 autumn small   high  6.8 11.1 9.00 0.630  20  4.0  NA  2.7 30.3 1.9 0.0
## 38 spring small   high  8.0   NA 1.45 0.810  10  2.5 3.0  0.3 75.8 0.0 0.0
## 48 winter small    low   NA 12.6 9.00 0.230  10  5.0 6.0  1.1 35.5 0.0 0.0
## 55 winter small   high  6.6 10.8   NA 3.245  10  1.0 6.5   NA 24.3 0.0 0.0
## 56 spring small medium  5.6 11.8   NA 2.220   5  1.0 1.0   NA 82.7 0.0 0.0
## 57 autumn small medium  5.7 10.8   NA 2.550  10  1.0 4.0   NA 16.8 4.6 3.9
##      a4  a5  a6  a7
## 28  0.0 2.1 1.4 2.1
## 38  0.0 0.0 0.0 0.0
## 48  0.0 0.0 0.0 0.0
## 55  0.0 0.0 0.0 0.0
## 56  0.0 0.0 0.0 0.0
## 57 11.5 0.0 0.0 0.0
```

```
## policz niekompletne wiersze
nrow(algae[!complete.cases(algae),])
```

```
## [1] 16
```

```
## sprawdzenie liczby braków w wierszach
apply(algae, 1, function(x) sum(is.na(x)))
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
##  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##   0   1   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##   2   2   2   2   2   2   2   6   1   0   0   0   0   0   0   0   0   0
##  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
##   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
## 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
## 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
##   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 199 200
##   6   0
```

Wiele ciekawych funkcji do eksploracji danych znajduje się w pakiecie **DMwR** (Torgo, 2013), który został przygotowany przy okazji publikacji książki *Data Mining with R*.

```
## poszukiwanie wierszy zawierających wiele braków
## w tym przypadku próg wyświetlania ustawiony jest na 0.2
## czyli 20% wszystkich kolumn
library(DMwR)
manyNAs(algae)
```

```
##  62 199
##  62 199
```

```
## tworzenie zbioru pozbawionego wierszy zawierających wiele braków
algae2 <- algae[-manyNAs(algae), ]

## sprawdzamy liczbę wybrakowanych wierszy które pozostały
nrow(algae2[!complete.cases(algae2),])
```

```
## [1] 14
```

```
## usuwamy wszystkie wiersze z brakami
algae3 <- na.omit(algae)

## wyświetl wiersze z brakami
algae3[!complete.cases(algae3),] %>% head()
```

```
##  [1] season size   speed  mxPH   mnO2   Cl     NO3    NH4    oPO4   PO4
## [11] Chla   a1     a2     a3     a4     a5     a6     a7
## <0 rows> (or 0-length row.names)
```

```
## liczba pozostałych wybrakowanych wierszy
nrow(algae3[!complete.cases(algae3),])
```

```
## [1] 0
```

```
## można oczywiście też ręcznie usuwać wiersze (nie polecam)
algae4 <- algae[-c(62,199),]
```

Można też zbudować funkcję, która będzie usuwała braki danych wg naszego upodobania.

```
## najpierw budujemy funkcję i ją kompilujemy aby R mógł ja stosować
## parametr prog ustala próg odcięcia wierszy
czysc.dane <- function(dt, prog = 0){
    licz.braki <- apply(dt, 1, function(x) sum(is.na(x)))
    czyste.dt <- dt[!(licz.braki/ncol(dt)>prog), ]
    return(czyste.dt)
}

## potem ją możemy stosować
algae4 <- czysc.dane(algae)
nrow(algae4[!complete.cases(algae4),])
```

```
## [1] 0
```

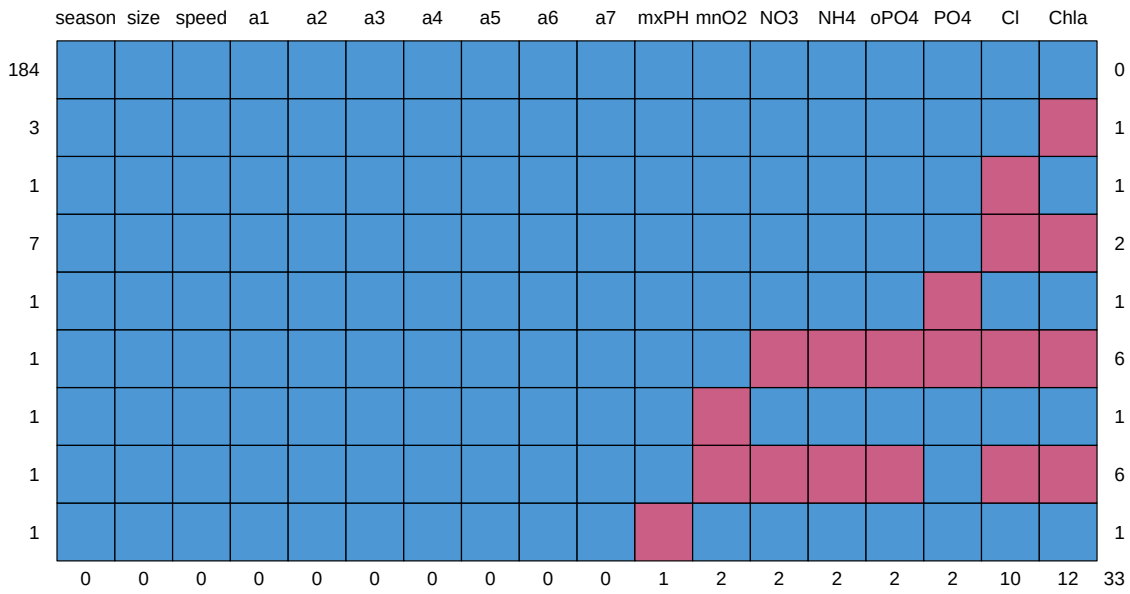
```
## czyścimy wiersze, których liczba braków przekracza 20% wszystkich kolumn
algae5 <- czysc.dane(algae, prog = 0.2)
nrow(algae5[!complete.cases(algae5),])
```

```
## [1] 14
```

Bardzo ciekawym narzędziem do znajdowania braków danych jest funkcja `md.pattern` pakietu **mice** (van Buuren and Groothuis-Oudshoorn, 2018). Wskazuje on ile braków występuje w ramach każdej zmiennej.

```
library(mice)
md.pattern(algae)
```

```
##     season size speed a1 a2 a3 a4 a5 a6 a7 mxPH mnO2 NO3 NH4 oPO4 PO4 Cl
## 184      1    1     1  1  1  1  1  1  1  1    1    1   1   1    1   1  1
## 3        1    1     1  1  1  1  1  1  1  1    1    1   1   1    1   1  1
## 1        1    1     1  1  1  1  1  1  1  1    1    1   1   1    1   1  0
## 7        1    1     1  1  1  1  1  1  1  1    1    1   1   1    1   1  0
## 1        1    1     1  1  1  1  1  1  1  1    1    1   1   1    1   0  1
## 1        1    1     1  1  1  1  1  1  1  1    1    1   0   0    0   0  0
## 1        1    1     1  1  1  1  1  1  1  1    1    0   1   1    1   1  1
## 1        1    1     1  1  1  1  1  1  1  1    1    0   0   0    0   1  0
## 1        1    1     1  1  1  1  1  1  1  1    0    1   1   1    1   1  1
##          0    0     0  0  0  0  0  0  0  0    1    2   2   2    2   2 10
##     Chla
## 184    1  0
## 3      0  1
## 1      1  1
## 7      0  2
## 1      1  1
## 1      0  6
## 1      1  1
## 1      0  6
## 1      1  1
##       12 33
```



Rysunek 2.1: Na czerwono zaznaczone są zmienne, które zwierają braki danych. Liczba w wierszu po lewej stronie wykresu wskazuje ile wierszy w bazie ma daną charakterystykę, a liczba po prawej oznacza ile zmiennych było *wybrakowanych*

2.1.2 Zastępowanie braków danych

Zastępowanie braków danych (zwane także *imputacją danych*) jest kolejnym etapem procesu przygotowania danych do analiz. Nie można jednak wyróżnić uniwersalnego sposobu zastępowania braków dla wszystkich możliwych sytuacji. Wśród statystyków panuje przekonanie, że w przypadku wystąpienia braków danych można zastosować trzy strategie:

- nic nie robić z brakami - co wydaje się niedorzeczne ale wcale takie nie jest, ponieważ istnieje wiele modeli statystycznych (np. drzewa decyzyjne), które świetnie radzą sobie w sytuacji braków danych. Niestety nie jest to sposób, który można stosować zawsze, ponieważ są również modele wymagające kompletności danych jak na przykład sieci neuronowe.
- usuwać braki wierszami[1] - to metoda, która jest stosowana domyślnie w przypadku kiedy twórca modelu nie zadecyduje o innym sposobie obsługi luk. Metoda ta ma swoją niewątpliwą zaletę w postaci jasnej i prostej procedury, ale szczególnie w przypadku niewielkich zbiorów może skutkować obciążeniem estymatorów. Nie wiemy bowiem jaka wartość faktycznie jest przypisana danej cesze. Jeśli jest to wartość bliska np. średniej, to nie wpłynie znacząco na obciążenie estymatora wartości oczekiwanej. W przypadku, gdy różni się ona znacznie od średniej tej cechy, to estymator może już wykazywać obciążenie. Jego wielkość zależy również od liczby usuniętych elementów. Nie jest zalecane usuwanie wielu wierszy ze zbioru danych i na podstawie okrojonego zbioru wyciąganie wniosków o populacji, ponieważ próba jest wówczas znacząco inna niż populacja. Dodatkowo jeśli estymatory są wyznaczane na podstawie zbioru wyraźnie mniej licznego, to precyzja estymatorów wyrażona wariancją spada. Reasumując, jeśli liczba wierszy z brakującymi danymi jest niewielka w stosunku do całego zbioru, to usuwanie wierszy jest sensownym rozwiązaniem.
- uzupełnianie braków - to procedura polegająca na zastępowaniu braków różnymi technikami. Jej niewątpliwą zaletą jest fakt posiadania kompletnych danych bez konieczności usuwania wierszy. Niestety wiąże się to również z pewnymi wadami. Zbiór posiadający wiele braków uzupełnianych nawet bardzo wyrafinowanymi metodami może cechować się zaniżoną wariancją poszczególnych cech oraz tzw. przeuczeniem[2].

1. Uzupełnianie średnią - braki w zakresie danej zmiennej uzupełniamy średnią tej zmiennej przypadków uzupełnionych.

```
algae[is.na(algae$mxPH), ]
```

```
##    season  size speed mxPH mnO2 Cl  NO3 NH4 oPO4 PO4 Chla   a1 a2 a3 a4 a5
## 48 winter small   low   NA 12.6  9 0.23  10    5   6  1.1 35.5  0  0  0  0
##    a6 a7
## 48  0  0
```

```
m <- mean(algae$mxPH, na.rm = T)
algae[is.na(algae$mxPH), "mxPH"] <- m
algae[is.na(algae$mxPH), ]
```

```
##  [1] season size   speed  mxPH   mnO2   Cl     NO3    NH4    oPO4   PO4
## [11] Chla   a1     a2     a3     a4     a5     a6     a7
## <0 rows> (or 0-length row.names)
```

2. Uzupełnianie medianą - braki w zakresie danej zmiennej uzupełniamy medianą tej zmiennej przypadków uzupełnionych.

```
algae %>% filter(is.na(Chla)) %>% head
```

```
##   season  size  speed mxPH mnO2 Cl   NO3 NH4 oPO4  PO4 Chla   a1  a2  a3
## 1 winter small   high  6.6 10.8 NA 3.245  10    1  6.5   NA 24.3 0.0 0.0
## 2 spring small medium  5.6 11.8 NA 2.220   5    1  1.0   NA 82.7 0.0 0.0
## 3 autumn small medium  5.7 10.8 NA 2.550  10    1  4.0   NA 16.8 4.6 3.9
```

[1] polega na usuwaniu wierszy zawierających braki
[2] więcej o zjawisku przeuczenia w dalszej części książki

```
## 4 spring small   high  6.6  9.5 NA 1.320  20    1  6.0   NA 46.8 0.0 0.0
## 5 summer small   high  6.6 10.8 NA 2.640  10    2 11.0   NA 46.9 0.0 0.0
## 6 autumn small medium  6.6 11.3 NA 4.170  10    1  6.0   NA 47.1 0.0 0.0
##     a4 a5  a6 a7
## 1  0.0  0 0.0  0
## 2  0.0  0 0.0  0
## 3 11.5  0 0.0  0
## 4 28.8  0 0.0  0
## 5 13.4  0 0.0  0
## 6  0.0  0 1.2  0
```

```
algae[is.na(algae$Chla), "Chla"] <- median(algae$Chla, na.rm = T)
```

3. Wypełnianie zmiennych typu wyliczeniowego, logicznego lub znakowego odbywa się najczęściej przez dobranie w miejsce brakującej wartości, elementu powtarzającego się najczęściej wśród obiektów obserwowanych. W pakiecie **DMwR** istnieje funkcja `centralImputation`, która wypełnia braki wartością centralną (w przypadku zmiennych typu liczbowego - medianą, a dla wartości logicznych, wyliczeniowych lub tekstowych - modą).

```
algae[48, "season"]
```

```
## [1] "winter"
```

```
algae[48, "season"] <- NA
algae.uzup <- centralImputation(algae)
algae.uzup[48,]
```

```
##    season  size speed     mxPH mnO2 Cl  NO3 NH4 oPO4 PO4 Chla   a1 a2 a3
## 48 winter small   low 8.011734 12.6  9 0.23  10    5   6  1.1 35.5  0  0
##    a4 a5 a6 a7
## 48  0  0  0  0
```

4. Jeszcze innym sposobem imputacji danych są algorytmy oparte o metodę k-najbliższych sąsiadów. Algorytm opiera się na prostej zasadzie, uzupełniania brakujących wartości medianą (w przypadku zmiennych ilościowych) lub modą (w przypadku zmiennych jakościowych) elementów, które są k-tymi najbliższymi sąsiadami w metryce

$$d(x,y) = \sqrt{\sum_{i=1}^{p} \delta_i(x_i, y_i)}, \tag{2.1}$$

gdzie δ_i jest odległością pomiędzy dwoma elementami ze względu na i-tą cech, określoną następująco

$$\delta_i(v_1, v_2) = \begin{cases} 1, & \text{jeśli zmienna jest jakościowa i } v_1 \neq v_2 \\ 0, & \text{jeśli zmienna jest jakościowa i } v_1 = v_2 \\ (v_1 - v_2)^2, & \text{jeśli zmienna jest ilościowa.} \end{cases} \tag{2.2}$$

Odległości są mierzone dla zmiennych standaryzowanych. Istnieje też odmiana z wagami, które maleją wraz ze wzrostem odległości pomiędzy sąsiadem a uzupełnianym elementem (np. $w(d) = \exp(d)$).

```
algae[48, ]
```

```
##    season  size speed     mxPH mnO2 Cl  NO3 NH4 oPO4 PO4 Chla   a1 a2 a3
## 48   <NA> small   low 8.011734 12.6  9 0.23  10    5   6  1.1 35.5  0  0
##    a4 a5 a6 a7
## 48  0  0  0  0
```

```
algae <- algae %>%
    mutate_if(is.character, as.factor)
algae.uzup <- knnImputation(algae, k = 5, scale = F, meth = "median")
algae.uzup[48,]
```

```
##    season  size speed     mxPH mnO2 Cl  NO3 NH4 oPO4 PO4 Chla   a1 a2 a3
## 48 summer small   low 8.011734 12.6  9 0.23  10    5   6  1.1 35.5  0  0
##    a4 a5 a6 a7
## 48  0  0  0  0
```

Istnieją również dużo bardziej złożone algorytmy imputacji danych oparte na bardziej wyrafinowanych technikach, takich jak: predykcja modelami liniowymi, nieliniowymi, analiza dyskryminacyjna, drzewa klasyfikacyjne. Dwa najbardziej znane pakiety zawierające funkcje do imputacji w sposób złożony, to **Amelia** i **mice**.

Imputacja danych z zastosowaniem pakietu **mice** wymaga podjęcia kilku decyzji przed przystąpieniem do uzupełniania danych:

1. Czy dane są MAR (ang. *Missing At Random*) czy MNAR (ang. *Missing Not At Random*), co oznacza, że musimy się zastanowić jakie mogły być źródła braków danych, przypadkowe czy systematyczne?
2. Należy się zdecydować na formę imputacji, określając strukturę zależności pomiędzy cechami oraz rozkład błędu danej cechy?
3. Wybrać zbiór danych, który posłuży nam za predyktory w imputacji (nie mogą zawierać braków).
4. Określenie, które niepełne zmienne są funkcjami innych wybrakowanych zmiennych.
5. Określić w jakiej kolejności dane będą imputowane.
6. Określić parametry startowe imputacji (liczbę iteracji, warunek zbieżności).
7. Określić liczę imputowanych zbiorów.

Ad 1. Wyróżniamy następujące rodzaje braków danych:

- MCAR (ang. *Missing Completely At Random*) - z definicji to braki, których pojawienie się jest kompletnie losowe. Przykładowo gdy osoba poproszona o wypełnienie wieku w ankiecie będzie rzucać monetą czy wypełnić tą zmienną.
- MAR - oznacza, że obserwowane wartości i wybrakowane mają inne rozkłady ale da się je oszacować na podstawie danych obserwowanych. Przykładowo ciśnienie tętnicze u osób, które nie wypełniły tej wartości jest wyższe niż u osób, które wpisały swoje ciśnienie. Okazuje się, że osoby starsze z nadciśnieniem nie wypełniały ankiety w tym punkcie.
- MNAR - jeśli nie jest spełniony warunek MCAR i MAR, wówczas brak ma charakter nielosowy. Przykładowo respondenci osiągający wyższe zarobki sukcesywnie nie wypełniają pola "zarobki" i dodatkowo nie ma w ankiecie zmiennych, które pozwoliłyby nam ustalić, jakie to osoby.

Ad 2. Decyzja o algorytmie imputacji wynika bezpośrednio ze skali w jakiej jest mierzona dana zmienna. Ze względu na rodzaj cechy używać będziemy następujących metod:

Każdy z czterech typów danych ma swój domyślny algorytm przeznaczony do imputacji:

- zmienna ilościowa - `pmm`
- zmienna dychotomiczna (stany 0 lub 1) - `logreg`
- zmienna typu wyliczeniowego (nieuporządkowana) - `polyreg`
- zmienna typu wyliczeniowego (uporządkowana) - `polr`

Niewątpliwą zaletą metody `pmm` jest to, że wartości imputowane są ograniczone jedynie do obserwowanych wartości. Metody `norm` i `norm.nob` uzupełniają brakujące wartości w oparciu o model liniowy. Są one szybkie i efektywne w przypadku gdy reszty modelu są zbliżone rozkładem do normalności. Druga z tych technik nie bierze pod uwagę niepewności związanej z modelem imputującym. Metoda `2L.norm` opiera się na dwupoziomowym heterogenicznym modelu liniowym (skupienia są włączone jako efekt do modelu). Technika `polyreg` korzysta z funkcji `multinom` pakietu **nnet** tworzącej model wielomianowy. `polr` opiera się o proporcjonalny

model logitowy z pakietu **MASS**. `lda` to model dyskryminacyjny klasyfikujący obiekty na podstawie prawdopodobieństw *a posteriori*. Metoda `sample` zastępuje braki losowa wybranymi wartościami spośród wartości obserwowanych.

Ad 3. Do ustalenia predyktorów w modelu `mice` służy funkcja `predictorMatrix`. Po pierwsze wyświetla ona domyślny układ predyktorów włączanych do modelu. Można go dowolnie zmienić i podstawić do modelu imputującego dane parametrem `predictorMatrix`. Zera występujące w kolejnych wierszach macierzy predyktorów oznaczają pominięcie tej zmiennej przy imputacji innej zmiennej. Jeśli dodatkowo chcemy by jakaś zmienna nie była imputowana, to oprócz usunięcia jej z listy predyktorów, należy wymazać ją z listy metod predykcji (`method`).

Ogólne zalecenia co do tego jakie zmienne stosować jako predyktory jest takie, żeby brać ich jak najwięcej. Spowoduje to, że bardziej prawdopodobny staje się brak typu MAR a nie MNAR. Z drugiej jednak strony, nierzadko zbiory zawierają olbrzymią liczbę zmiennych i włączanie ich wszystkich do modelu imputującego nie będzie miało sensu.

Zalecenia doboru zmiennych są następujące:

- weź wszystkie te zmienne, które są włączane do modelu właściwego, czyli tego za pomocą którego chcesz poznać strukturę zależności;
- czasem do modelu imputującego należy też włączyć interakcje zmiennych z modelu właściwego;
- dodaj zmienne, które mogą mieć wpływ na wybrakowane cechy;
- włącz zmienne istotnie podnoszące poziom wyjaśnionej wariancji modelu;
- na koniec usuń te zmienne spośród predyktorów, które same zawierają zbyt wiele braków.

Ad 4-7. Decyzje podejmowane w tych punktach zależą istotnie od analizowanego zbioru i będą przedmiotem oddzielnych analiz w kontekście rozważanych zbiorów i zadań.

2.2 Przykład

Dokonamy imputacji zbioru `airquality` z wykorzystaniem pakietów **mice** i **VIM** (Templ et al., 2019)

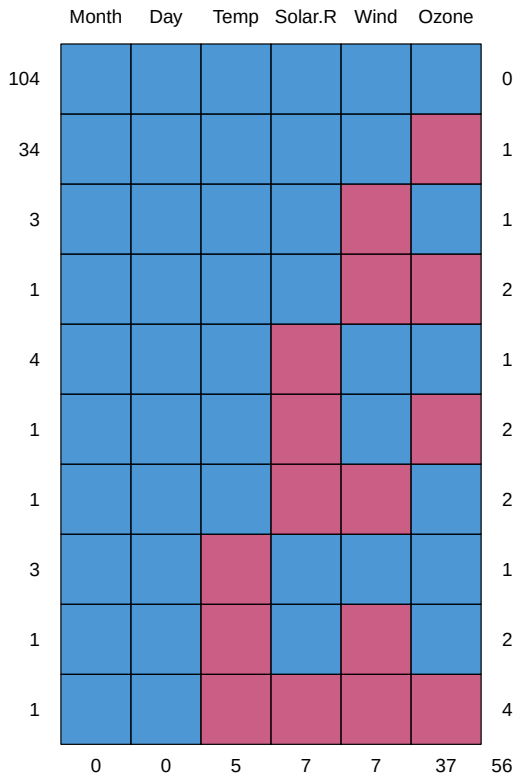
```
data <- airquality
summary(data)
```

```
##      Ozone           Solar.R          Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7                                       
##      Month            Day       
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0  
##
```

```
# tworzymy dodatkowe braki danych
data[4:10,3] <- rep(NA,7)
data[1:5,4] <- NA
summary(data)
```

```
##      Ozone           Solar.R          Wind             Temp
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :57.00
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:73.00
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.806   Mean   :78.28
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
##  NA's   :37       NA's   :7       NA's   :7        NA's   :5
##      Month            Day
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
##
```

```
md.pattern(data)
```

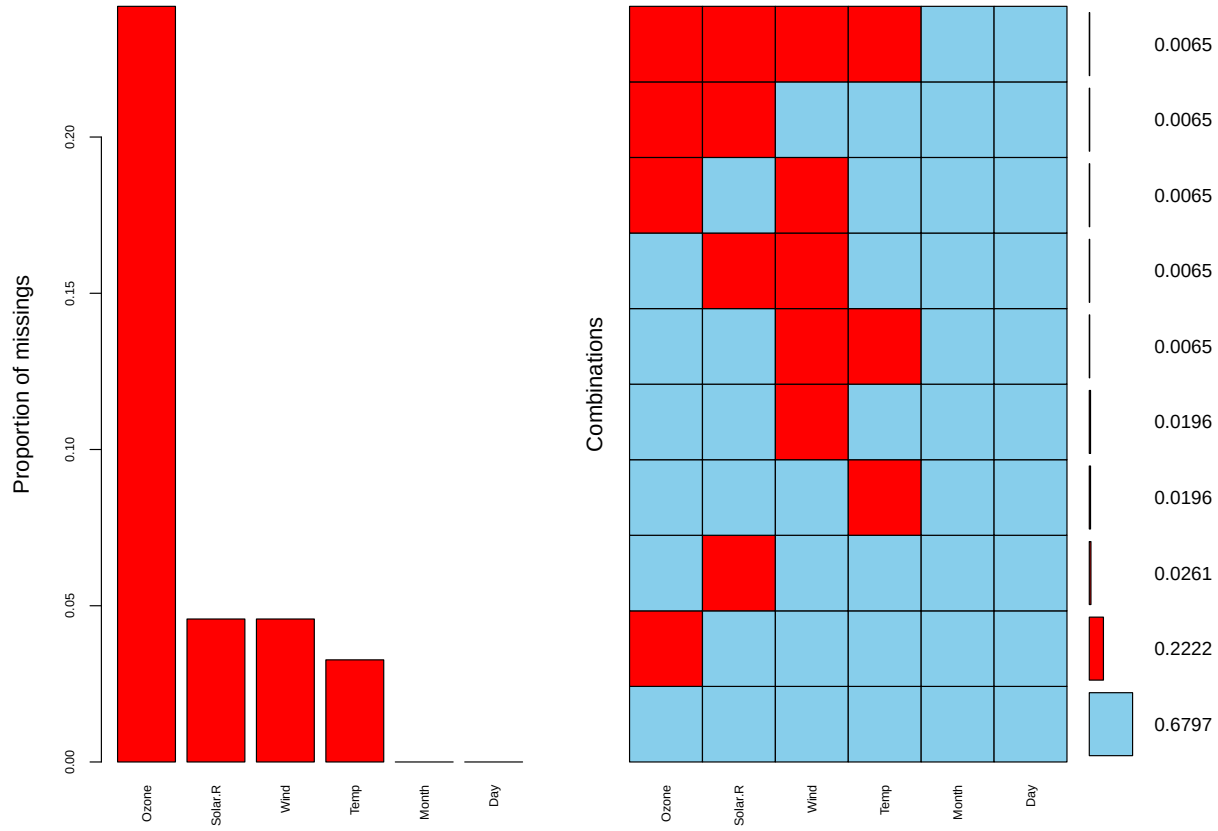



```
##     Month Day Temp Solar.R Wind Ozone
## 104     1   1    1       1    1     1  0
## 34      1   1    1       1    1     0  1
## 3       1   1    1       1    0     1  1
## 1       1   1    1       1    0     0  2
## 4       1   1    1       0    1     1  1
## 1       1   1    1       0    1     0  2
```

```
## 1        1    1    1        0    0      1   2
## 3        1    1    0        1    1      1   1
## 1        1    1    0        1    0      1   2
## 1        1    1    0        0    0      0   4
##          0    0    5        7    7     37  56
```

Do ilustracji braków danych można zastosować funkcje pakietu **VIM**.

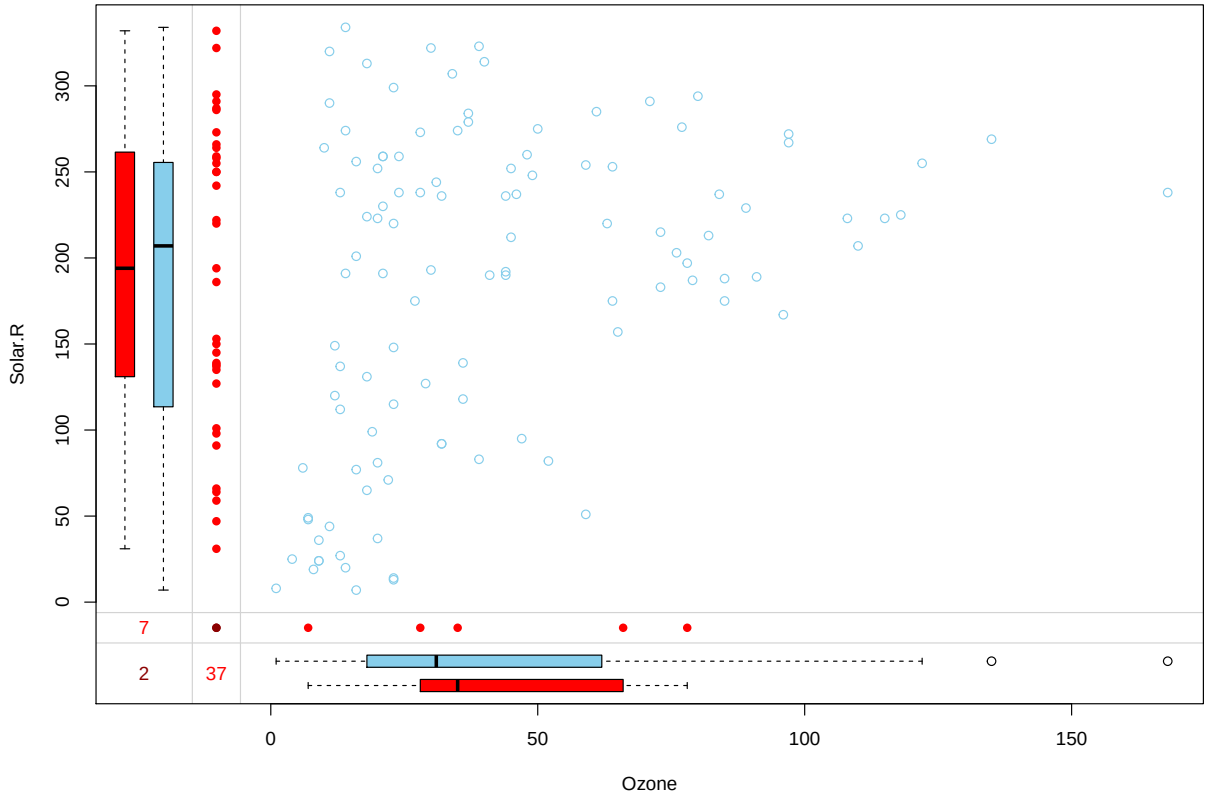
```
library(VIM)
aggr(data, numbers=TRUE,
     sortVars=TRUE,
     labels=names(data),
     cex.axis=.7)
```




```
##
##  Variables sorted by number of missings:
##  Variable       Count
##     Ozone 0.24183007
##   Solar.R 0.04575163
##      Wind 0.04575163
##      Temp 0.03267974
##     Month 0.00000000
##       Day 0.00000000
```

Tak przedstawia się wykres rozrzutu zmiennych `Ozone` i `Solar.R` z uwzględnieniem położenia braków danych.

```
marginplot(data[c(1,2)])
```



Dokonamy imputacji metodą pmm.

```
tempData <- mice(data,
                 maxit=50,
                 meth='pmm',
                 seed=44,
                 printFlag = F)
summary(tempData)
```

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##   Ozone Solar.R    Wind    Temp   Month     Day
##   "pmm"   "pmm"   "pmm"   "pmm"      ""      ""
## PredictorMatrix:
##         Ozone Solar.R Wind Temp Month Day
## Ozone       0       1    1    1     1   1
## Solar.R     1       0    1    1     1   1
## Wind        1       1    0    1     1   1
## Temp        1       1    1    0     1   1
## Month       1       1    1    1     0   1
## Day         1       1    1    1     1   0
```

Ponieważ, funkcja mice domyślnie dokonuje 5 kompletnych imputacji, możemy się przekonać jak bardzo

różnią się poszczególne imputacje i zdecydować się na jedną z nich.

```
head(tempData$imp$Ozone)
```

```
##     1  2  3  4  5
## 5  21 20  7 36 13
## 10 21 16 44 22 21
## 25 14 14 14  6  8
## 26 23 18  8 19 14
## 27 37 23 21  7  9
## 32 63 23  7 52 39
```

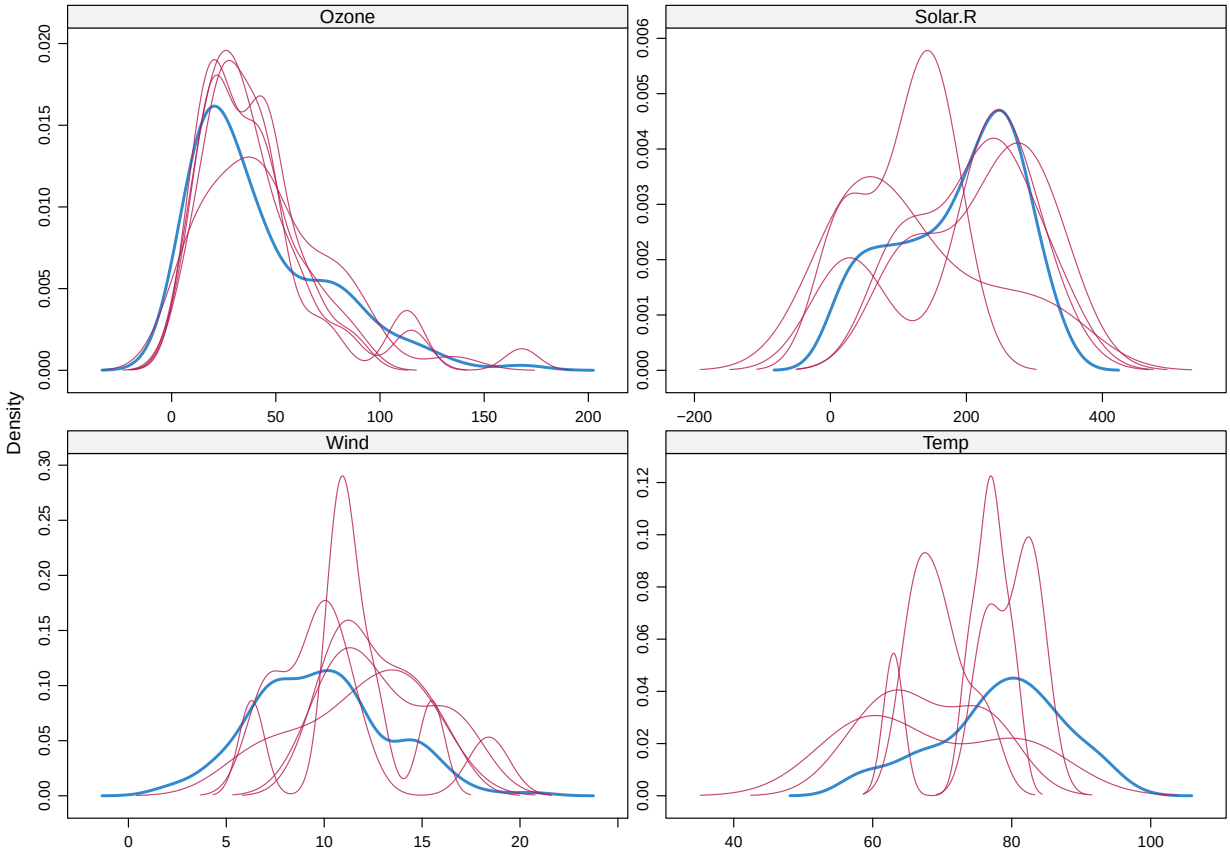
Ostatecznie imputacji dokonujemy wybierając jeden z zestawów danych uzupełniających (np. pierwszy).

```
completedData <- mice::complete(tempData, 1)
summary(completedData)
```

```
##      Ozone          Solar.R           Wind             Temp
##  Min.   :  1.0   Min.   :  7.0   Min.   : 1.700   Min.   :57.00
##  1st Qu.: 20.0   1st Qu.:115.0   1st Qu.: 7.400   1st Qu.:73.00
##  Median : 32.0   Median :212.0   Median : 9.700   Median :79.00
##  Mean   : 42.5   Mean   :187.9   Mean   : 9.931   Mean   :78.14
##  3rd Qu.: 59.0   3rd Qu.:259.0   3rd Qu.:11.500   3rd Qu.:85.00
##  Max.   :168.0   Max.   :334.0   Max.   :20.700   Max.   :97.00
##      Month            Day
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
```

Za pomocą funkcji pakietu `mice` możemy również przedstawić graficznie gdzie i jak zostały uzupełnione dane.

```
densityplot(tempData, ~Ozone+Solar.R+Wind+Temp)
```




```
stripplot(tempData, Ozone+Solar.R+Wind+Temp~.imp, pch = 20, cex = 1.2)
```

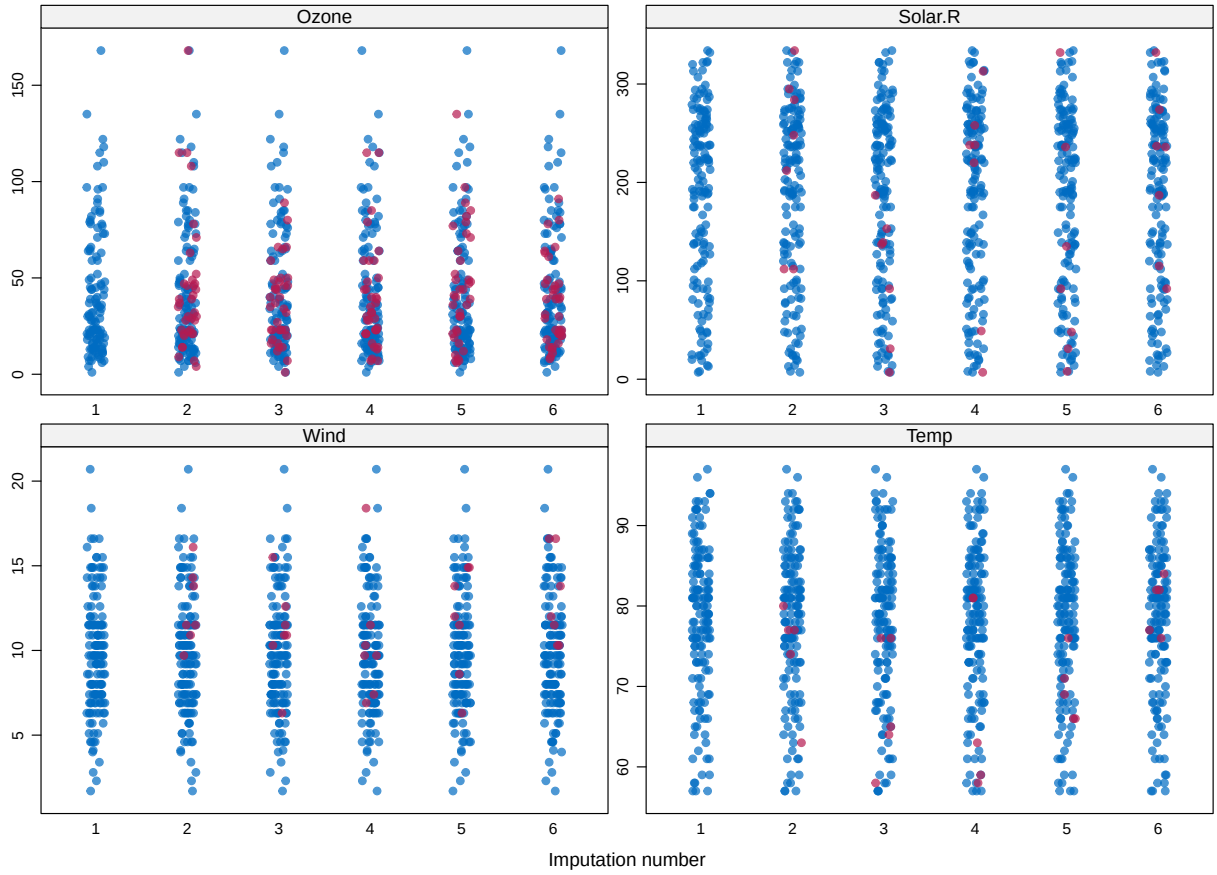

Ozone
Solar.R
Wind
Temp
Imputation number

Tabela 2.1: Zestaw metod imputacji danych stosowanych w pakiecie **mice**

method	type	description
pmm	any	Predictive.mean.matching
midastouch	any	Weighted predictive mean matching
sample	any	Random sample from observed values
cart	any	Classification and regression trees
rf	any	Random forest imputations
mean	numeric	Unconditional mean imputation
norm	numeric	Bayesian linear regression
norm.nob	numeric	Linear regression ignoring model error
norm.boot	numeric	Linear regression using bootstrap
norm.predict	numeric	Linear regression, predicted values
quadratic	numeric	Imputation of quadratic terms
ri	numeric	Random indicator for nonignorable data
logreg	binary	Logistic regression
logreg.boot	binary	Logistic regression with bootstrap
polr	ordered	Proportional odds model
polyreg	unordered	Polytomous logistic regression
lda	unordered	Linear discriminant analysis
2l.norm	numeric	Level-1 normal heteroscedastic
2l.lmer	numeric	Level-1 normal homoscedastic, lmer
2l.pan	numeric	Level-1 normal homoscedastic, pan
2l.bin	binary	Level-1 logistic, glmer
2lonly.mean	numeric	Level-2 class mean
2lonly.norm	numeric	Level-2 class normal
2lonly.pmm	any	Level-2 class predictive mean matching

Rozdział 3

Podział metod data mining

3.1 Rodzaje wnioskowania

Data mining to zestaw metod pozyskiwania wiedzy na podstawie danych. Ową wiedzę zdobywamy w procesie wnioskowania na podstawie modeli. Wnioskowanie możemy podzielić na dedukcyjne i indukcyjne. I tak z wnioskowaniem dedukcyjnym mamy do czynienia wówczas, gdy na podstawie obecnego stanu wiedzy potrafimy odpowiedzieć na postawione pytanie dotyczące nowej wiedzy, stosując reguły wnioskowania. O wnioskowaniem indukcyjnym powiemy, że jest to metoda pozyskiwania wiedzy na podstawie informacji ze zbioru uczącego. Znajduje ono szerokie zastosowanie w data mining i charakteryzuje się omylnością, ponieważ nawet najlepiej nauczony model na zbiorze uczącym nie zapewnia nam prawdziwości odpowiedzi w przypadku nowych danych, a jedynie je uprawdopodabnia. Esencją wnioskowania indukcyjnego w zakresie data mining, jest poszukiwanie na podstawie danych uczących modelu charakteryzującego się najlepszymi właściwościami predykcyjnymi i dającego się zastosować do zupełnie nowego zbioru danych.

Każdy proces uczenia z wykorzystaniem wnioskowania indukcyjnego składa się z następujących elementów.

3.1.1 Dziedzina

Dziedzina to zbiór wszystkich obiektów pozostających w zainteresowaniu badacza, będących przedmiotem wnioskowania, oznaczana najczęściej przez X. Przykładowo mogą to być zbiory osób, transakcji, urządzeń, instytucji, itp.

3.1.2 Obserwacja

Każdy element dziedziny $x \in X$ nazywamy obserwacją. Obserwacją nazywać będziemy zarówno rekordy danych ze zbioru uczącego, jak i ze zbioru testowego.

3.1.3 Atrybuty obserwacji

Każdy obiekt z dziedziny $x \in X$ można opisać zestawem cech (atrybutów), które w notacji matematycznej oznaczymy przez $a : X \to A$, gdzie A jest przestrzenią wartości atrybutów. Każda obserwacja x posiadająca k cech da się wyrazić wektorowo jako $(a_1(x), a_2(x), \ldots, a_k(x))$. Dla większości algorytmów uczenia maszynowego wyróżnia się trzy typy atrybutów:

- *nominalne* - posiadające skończoną liczbę stanów, które posiadają porządku;
- *porządkowe* - posiadające skończoną liczbę stanów z zachowaniem porządku;
- *ciągłe* - przyjmujące wartości numeryczne.

Często jeden z atrybutów spełnia specjalną rolę, ponieważ stanowi realizację cechy, którą traktujemy jako wyjściową (ang. *target value attribute*). W tym przypadku powiemy o **nadzorowanym uczeniu maszy-**

nowym. Jeśli zmiennej wyjściowej nie ma dziedzinie, to mówimy o **nienadzorowanym uczeniu maszynowym**.

3.1.4 Zbiór uczący

Zbiorem uczącym T (ang. *training set*) nazywamy podzbiór D dziedziny X (czyli $T \subseteq D \subseteq X$), gdzie zbiór D stanowi ogół dostępnych obserwacji z dziedziny X. Zbiór uczący zawiera informacje dotyczące badanego zjawiska, na podstawie których, dokonuje się doboru modelu, selekcji cech istotnych z punktu widzenia własności predykcyjnych lub jakości klasyfikacji, budowy modelu oraz optymalizacji jego parametrów. W przypadku uczenia z nauczycielem (nadzorowanego) zbiór T zawiera informację o wartościach atrybutów zmiennej wynikowej.

3.1.5 Zbiór testowy

Zbiór testowy T' (ang. *test set*) będący dopełnieniem zbioru uczącego do zbioru D, czyli $T' = D \setminus T$, stanowi zestaw danych służący do oceny poprawności modelu nadzorowanego. W przypadku metod nienadzorowanych raczej nie stosuje się zbiorów testowych.

3.1.6 Model

Model to narzędzie pozyskiwania wiedzy na podstawie zbioru uczącego. Nauczony model jest zbiorem reguł f, którego zadaniem jest oszacowanie wielkości wartości wynikowej lub odpowiednia klasyfikacja obiektów. W zadaniu grupowania obiektów (ang. *clustering task*), celem modelu jest podanie grup możliwie najbardziej jednorodnych przy zadanym zestawie zmiennych oraz ustalonej liczbie skupień (czasami wyznaczenie liczby skupień jest również częścią zadania stawianego przed modelem).

3.1.7 Jakość dopasowania modelu

Do oceny jakości dopasowania modelu wykorzystuje się, w zależności od zadania, wiele współczynników (np. dla zadań regresyjnych są to błąd średnio-kwadratowy - ang. *Mean Square Error* a dla zadań klasyfikacyjnych - trafność - ang. *Accuracy*). Możemy mówić dwóch rodzajach dopasowania modeli:

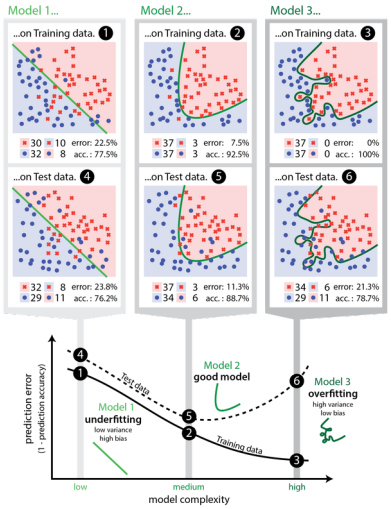
- poziom dopasowania na zbiorze uczącym
- poziom dopasowania na zbiorze testowym (oczywiście z punktu widzenia utylitarności modelu ten współczynnik jest ważniejszy).

W sytuacji, w której model wykazuje dobre charakterystyki jakości dopasowania na zbiorze uczącym ale słabe na testowym, mówimy o zjawisku przeuczenia modelu (ang. *overfitting*). Oznacza to, że model wskazuje predykcję poprawnie jedynie dla zbioru treningowego ale ma słaba własności generalizacyjne nowe przypadki danych. Takie model nie przedstawiają znaczącej wartości w odkrywaniu wiedzy w sposób indukcyjny.

Z drugiej strony parametry dopasowania modelu mogą pokazywać słabe dopasowanie, zarówno na zbiorze uczącym, jak i testowym. Wówczas również model nie jest użyteczny w pozyskiwaniu wiedzy na temat badanego zjawiska, a sytuację taką nazywamy niedouczeniem (ang. *underfitting*).

3.2 Modele regresyjne

Jednym z rodzajów zadań bazującym na wnioskowaniu indukcyjnym jest model regresyjny. Należy on do grupy metod nadzorowanych, których celem jest oszacowanie wartości cechy wyjściowej (która jest ilościowa) na podstawie zestawu predyktorów, które mogą być ilościowe i jakościowe. Uczenie takich modeli odbywa się poprzez optymalizację funkcji celu (np. MSE) na podstawie zbioru uczącego.



Rysunek 3.1: Przykłady niedoucznia (wykresy 1 i 4), poprawego modelu (2 i 5) i przeuczenia (3 i 6). Pierwszy wiersz wykresów pokazuje klasyfikację na podstawie modelu na zbiorze uczącym, a drugi na zbiorze testowym. Wykres na dole pokazuje związek pomiędzy złożonością modelu a wielkością błędu predykcji. *Źródło*: https://cambridgecoding.wordpress.com/2016/03/24/misleading-modelling-overfitting-cross-validation-and-the-bias-variance-trade-off/

3.3 Modele klasyfikacyjne

Podobnie jak modele regresyjne, modele klasyfikacyjne należą do grupy metod nadzorowanego uczenia maszynowego. Ich zadaniem jest właściwa klasyfikacja obiektów na podstawie wielkości predyktorów. Odpowiedzią modelu jest zawsze cecha typu jakościowego, natomiast predyktory mogą mieć dowolny typ. Wyróżnia się klasyfikację dwu i wielostanową. Lista modeli realizujących klasyfikację binarną jest nieco dłuższa niż w przypadku modeli z wielostanową cechą wynikową. Proces uczenia modelu klasyfikacyjnego również opiera się na optymalizacji funkcji celu. Tym razem są to zupełnie inne miary jakości dopasowania (np. trafność, czyli odsetek poprawnych klasyfikacji).

3.4 Modele grupujące

Bardzo szeroką gamę modeli nienadzorowanych stanowią metody analizy skupień. Ich zadaniem jest grupowanie obiektów w możliwie najbardziej jednorodne grupy, na podstawie wartości atrybutów poddanych analizie. Ponieważ są to metody “bez nauczyciela”, to ocena ich przydatności ma nieco inny charakter i choć istnieją różne wskaźniki jakości grupowania, to trudno tu o obiektywne wskazanie najlepszego rozwiązania.

Rozdział 4

Drzewa decyzyjne

Drzewo decyzyjne[1] jest strukturą hierarchiczną przedstawiającą model klasyfikacyjny lub regresyjny. Stosowane są szczególnie często wówczas, gdy funkcyjna postać związku pomiędzy predyktorami a zmienną wynikową jest nieznana lub ciężka do ustalenia. Każde drzewo decyzyjne składa się z korzenia (ang. *root*), węzłów (ang. *nodes*) i liści (ang. *leaves*). Korzeniem nazywamy początkowy węzeł drzewa, z którego poprzez podziały (ang. *splits*) powstają kolejne węzły potomne. Końcowe węzły, które nie podlegają podziałom nazywamy liśćmi, a linie łączące węzły nazywamy gałęziami (ang. *branches*).

Jeśli drzewo służy do zadań klasyfikacyjnych, to liście zawierają informację o tym, która klasa w danym ciągu podziałów jest najbardziej prawdopodobna. Natomiast jeśli drzewo jest regresyjne, to liście zawierają warunkowe miary tendencji centralnej (najczęściej średnią) wartości zmiennej wynikowej. Warunek stanowi szereg podziałów doprowadzający do danego węzła terminalnego (liścia). W obu przypadkach (klasyfikacji i regresji) drzewo "dąży" do takiego podziału by kolejne węzły, a co za tym idzie również liście, były ja najbardziej jednorodne ze względu na zmienną wynikową.

4.1 Węzły i gałęzie

Każdy podział rozdziela dziedzinę X na dwa lub więcej podobszarów dziedziny i wówczas każda obserwacja węzła nadrzędnego jest przyporządkowana węzłom potomnym. Każdy odchodzący węzeł potomny jest połączony gałęzią, która to wiąże się ściśle z możliwymi wynikami podziału. Każdy **n**-ty węzeł można opisać jako podzbiór dziedziny w następujący sposób

$$X_{\mathbf{n}} = \{x \in X | t_1(x) = r_1, t_2(x) = r_2, \ldots, t_k(x) = r_k\}, \tag{4.1}$$

gdzie $t_1, t_2, \ldots, t_k$ są podziałami, które przeprowadzają x w obszary $r_1, r_2, \ldots, r_k$. Przez

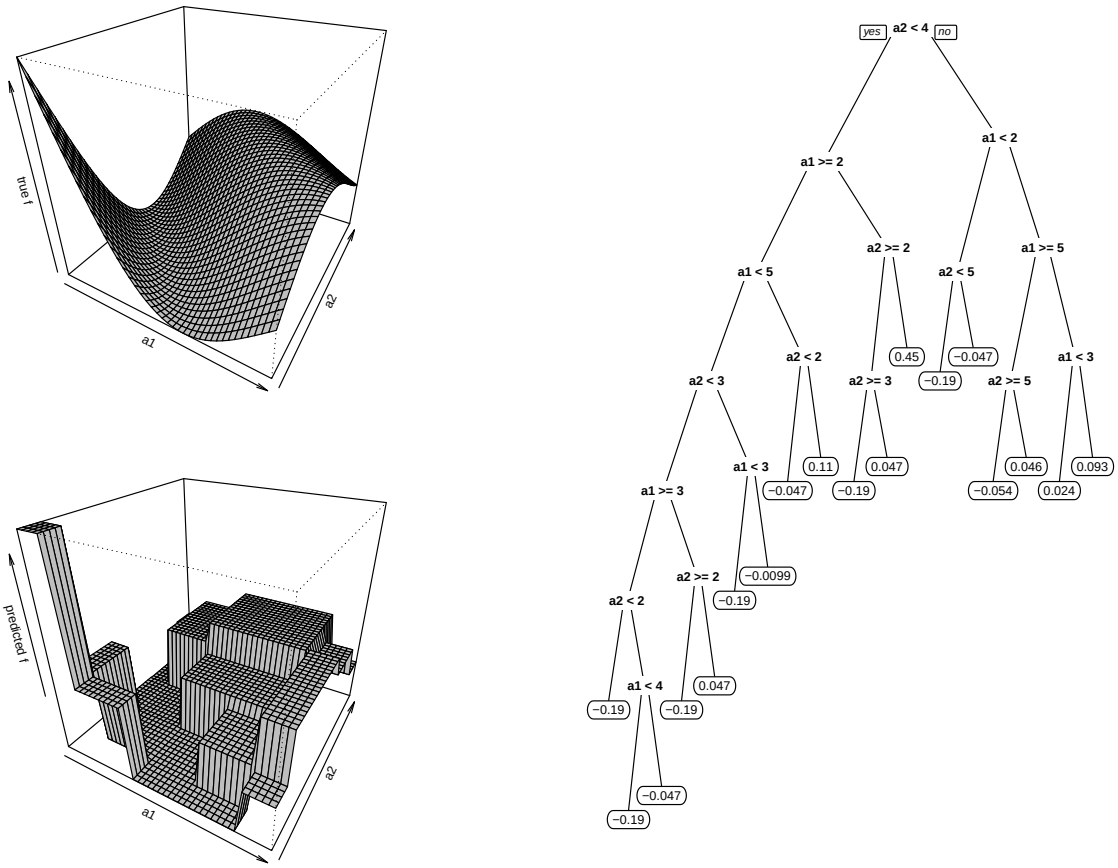
$$S_{\mathbf{n},t=r} = \{x \in S | t(x) = r\} \tag{4.2}$$

rozumiemy, że dokonano takiego ciągu podziałów zbioru S, że jego wartości znalazły się w **n**-tym węźle.

4.2 Rodzaje reguł podziału

Najczęściej występujące reguły podziału w drzewach decyzyjnych są jednowymiarowe, czyli warunek podziału jest generowany na podstawie jednego atrybutu. Istnieją podziały wielowymiarowe ale ze względu na złożoność obliczeniową są rzadziej stosowane.

[1] wyglądem przypomina odwrócone drzewo, stąd nazwa



Rysunek 4.1: Przykład działania drzewa regresyjnego. Wykes w lewym górnym rogu pokazuje prawdziwą zależność, wyres po prawej stronie jest ilustracją drzewa decyzyjnego, a wykres w lewym dolnym rogu pokazuje dyskretyzację przestrzeni dokonaną przez drzewo, a za razem sposób jego działania.

4.2.1 Podziały dla atrybutów ze skali nominalnej

Istnieją dwa typy reguł podziału dla skali nominalnej:

- oparte na wartości atrybutu (ang. *value based*) - wówczas funkcja testowa przyjmuje postać $t(x) = a(x)$, czyli podział generują wartości atrybutu;
- oparte na równości (ang. *equality based*) - gdzie funkcja testowa jest zdefiniowana jako

$$t(x) = \begin{cases} 1, & \text{gdy } a(x) = \nu \\ 0, & \text{w przeciwnym przypadku,} \end{cases} \tag{4.3}$$

  gdzie $\nu \in A$ i A jest zbiorem możliwych wartości a. W tym przypadku podział jest dychotomiczny, albo obiekt ma wartość atrybutu równą ν, albo go nie ma.

4.2.2 Podziały dla atrybutów ze skali ciągłej

Reguły podziału stosowane do skali ciągłej, to:

- oparta na nierównościach (ang. *inequality based*) - zdefiniowana jako

$$t(x) = \begin{cases} 1, & \text{gdy } a(x) \leq \nu \\ 0, & \text{w przeciwnym przypadku} \end{cases} \tag{4.4}$$

  gdzie $\nu \in A$;
- przedziałowa (ang. *interval based*) - zdefiniowana jako

$$t(x) = \begin{cases} 1, & \text{gdy } a(x) \in I_1 \\ 2, & \text{gdy } a(x) \in I_2 \\ \vdots & \\ k, & \text{gdy } a(x) \in I_k \end{cases} \tag{4.5}$$

  gdzie $I_1, I_2, \ldots, I_k \subset A$ stanowią rozłączny podział (przedziałami) przeciwdziedziny A.

4.2.3 Podziały dla atrybutów ze skali porządkowej

Podziały te mogą wykorzystywać oba wcześniej wspomniane typy, w zależności od potrzeb.

4.3 Algorytm budowy drzewa

1. stwórz początkowy węzeł (korzeń) i oznacz go jako *otwarty*;
2. przypisz wszystkie możliwe rekordy do węzła początkowego;
3. **dopóki** istnieją otwarte węzły **wykonuj**:
   - wybierz węzeł **n**, wyznacz potrzebne statystyki opisowe zmiennej zależnej dla tego węzła i przypisz wartość docelową;
   - **jeśli** kryterium zatrzymania podziału jest spełnione dla węzła n, **to** oznacz go za **zamknięty**;
   - **w przeciwnym przypadku** wybierz podział r elementów węzła **n**, i dla każdego podzbioru podziału stwórz węzeł niższego rzędu (potomka) $\mathbf{n}_r$ oraz oznacz go jako *otwarty*;
   - następnie przypisz wszystkie przypadki generowane podziałem r do odpowiednich węzłów potomków $\mathbf{n}_r$;
   - oznacza węzeł **n** jako *zamknięty*.

Sposób przypisywania wartości docelowej wiąże się ściśle z rodzajem drzewa. W drzewach regresyjnych chodzi o wyliczenie średniej lub mediany dla obserwacji ujętych w danym węźle. Natomiast w przypadku

drzewa klasyfikacyjnego, wyznacza się wartości prawdopodobieństw przynależności obserwacji znajdującej się w danym węźle do poszczególnych klas

$$\P(d|\mathbf{n}) = \P_{T_\mathbf{n}}(d) = \frac{|T^d_\mathbf{n}|}{|T_\mathbf{n}|}, \tag{4.6}$$

gdzie $T_\mathbf{n}$ oznaczają obserwacje zbioru uczącego znajdujące się w węźle $\mathbf{n}$, a $T^d_\mathbf{n}$ oznacza dodatkowo podzbiór zbioru uczącego w $\mathbf{n}$ węźle, które należą do klasy d. Oczywiście klasyfikacja na podstawie otrzymanych prawdopodobieństw w danym węźle jest dokonana przez wybór klasy charakteryzującej się najwyższym prawdopodobieństwem.

4.4 Kryteria zatrzymania

Kryterium zatrzymania jest warunkiem, który decyduje o tym, że dany węzeł uznajemy za zamknięty i nie dokonujemy dalszego jego podziału. Wyróżniamy następujące kryteria zatrzymania:

- jednorodność węzła - w przypadku drzewa klasyfikacyjnego może zdarzyć się sytuacja, że wszystkie obserwacje węzła będą pochodziły z jednej klasy. Wówczas nie ma sensu dokonywać dalszego podziału węzła;
- węzeł jest pusty - zbiór przypisanych obserwacji zbioru uczącego do $\mathbf{n}$-tego węzła jest pusty;
- brak reguł podziału - wszystkie reguły podziału zostały wykorzystane, zatem nie da się stworzyć potomnych węzłów, które charakteryzowałyby się większą homogenicznością;

Warunki ujęte w pierwszych dwóch kryteriach mogą być nieco złagodzone, poprzez zatrzymanie podziałów wówczas, gdy prawdopodobieństwo przynależenia do pewnej klasy przekroczy ustalony próg lub gdy liczebność węzła spadnie poniżej ustalonej wartości.

W literaturze tematu istnieje jeszcze jedno często stosowane kryterium zatrzymania oparte na wielkości drzewa. Węzeł potomny ustala się jako zamknięty, gdy długość ścieżki dojścia do nie go przekroczy ustaloną wartość.

4.5 Reguły podziału

Ważnym elementem algorytmu tworzenia drzewa regresyjnego jest *reguła podziału*. Dobierana jest w taki sposób aby zmaksymalizować zdolności generalizacyjne drzewa. Złożoność drzewa mierzona jest najczęściej przeciętną liczbą podziałów potrzebnych do dotarcia do liścia zaczynając od korzenia. Liście są najczęściej tworzone wówczas gdy dyspersja wartości wynikowej jest stosunkowo mała lub węzeł zawiera w miarę homogeniczne obserwacje ze względu na przynależność do klasy zmiennej wynikowej. W przypadku drzew regresyjnych zmienność na poziomie węzłów jest dobrą miarą służącą do definiowania podziału w węźle. I tak, jeśli pewien podział generuje nam stosunkowo małe dyspersje wartości docelowych w węzłach potomnych, to można ten podział uznać za właściwy. Jeśli T_n oznacza zbiór rekordów należących do węzła n, a $T_{n,t=r}$ są podzbiorami generowanymi przez podział r w węzłach potomnych dla n, to dyspersję wartości docelowej f będziemy oznaczali następująco

$$\mathrm{disp}_{T_{n,t=r}}(f). \tag{4.7}$$

Regułę podziału możemy określać poprzez minimalizację średniej ważonej dyspersji wartości docelowej następującej postaci

$$\mathrm{disp}_n(f|t) = \sum_{r\in R_t} \frac{|T_{n,t=r}|}{|T_n|}\,\mathrm{disp}_{T_{n,t=r}}(f), \tag{4.8}$$

gdzie $|\ |$ oznacza moc zbioru, a R_t zbiór wszystkich możliwych wartości reguły podziału. Czasami wygodniej będzie maksymalizować przyrost dyspersji (lub spadek)

$$\triangle\,\mathrm{disp}_n(f|t) = \mathrm{disp}_n(f) - \sum_{r\in R_t} \frac{|T_{n,t=r}|}{|T_n|}\,\mathrm{disp}_{T_{n,t=r}}(f). \tag{4.9}$$

Miarą heterogeniczności węzłów ze względu na zmienną wynikową (ang. *impurity*) w drzewach klasyfikacyjnych, która pozwala na tworzenie kolejnych podziałów węzła, są najczęściej wskaźnik Gini'ego i entropia (Breiman et al., 2017).

Entropią podzbioru uczącego w węźle $\mathbf{n}$, wyznaczamy wg wzoru

$$E_{T_\mathbf{n}}(c|t) = \sum_{x \in R_t} \frac{|T_{\mathbf{n},t=r}|}{|T_\mathbf{n}|} E_{T_{\mathbf{n},t=r}}(c), \tag{4.10}$$

gdzie t jest podziałem (kandydatem), r potencjalnym wynikiem podziału t, c jest oznaczeniem klasy zmiennej wynikowej, a

$$E_{T_{\mathbf{n},t=r}}(c) = \sum_{d \in C} -¶_{T_{\mathbf{n},t=r}}(c = d) \log ¶_{T_{\mathbf{n},t=r}}(c = d), \tag{4.11}$$

przy czym

$$¶_{T_{\mathbf{n},t=r}}(c = d) = ¶_{T_\mathbf{n}}(c = d|t = r). \tag{4.12}$$

Podobnie definiuje się indeks Gini'ego

$$Gi_{T_\mathbf{n}}(c|t) = \sum_{x \in R_t} \frac{|T_{\mathbf{n},t=r}|}{|T_\mathbf{n}|} Gi_{T_{\mathbf{n},t=r}}(c), \tag{4.13}$$

gdzie

$$Gi_{T_{\mathbf{n},t=r}}(c) = \sum_{d \in C} ¶_{T_{\mathbf{n},t=r}}(c = d) \cdot (1 - ¶_{T_{\mathbf{n},t=r}}(c = d)) = 1 - \sum_{d \in C} ¶^2_{T_{\mathbf{n},t=r}}(c = d). \tag{4.14}$$

Dla tak zdefiniowanych miar "nieczystości" węzłów, podziału dokonujemy w taki sposób, aby zminimalizować współczynnik Gini'ego lub entropię. Im niższe miary nieczystości, tym bardziej obserwacje znajdujące się w węźle są monokulturą[2]. Nierzadko korzysta się również z współczynnika przyrostu informacji (ang. *information gain*)

$$\Delta E_{T_\mathbf{n}}(c|t) = E_{T_\mathbf{n}}(c) - E_{T_\mathbf{n}}(c|t). \tag{4.15}$$

Istnieje również jego odpowiednik dla indeksu Gini'ego. W obu przypadkach optymalnego podziału szukamy poprzez maksymalizację przyrostu informacji.

4.6 Przycinanie drzewa decyzyjnego

Uczenie drzewa decyzyjnego wiąże się z ryzykiem przeuczenia modelu (podobnie jak to się ma w przypadku innych modeli predykcyjnych). Wcześniej przytoczone reguły zatrzymania (np. głębokość drzewa czy zatrzymanie przy osiągnięciu jednorodności na zadanym poziomie) pomagają kontrolować poziom generalizacji drzewa ale czasami będzie dodatkowo potrzebne przycięcie drzewa, czyli usunięcie pewnych podziałów, a co za tym idzie, również liści (węzłów).

4.6.1 Przycinanie redukujące błąd

Jedną ze strategii przycinania drzewa jest przycinanie redukujące błąd (ang. *reduced error pruning*). Polega ono na porównaniu błędów (najczęściej używana jest miara odsetka błędnych klasyfikacji lub MSE) liścia $\mathbf{l}$ i węzła do którego drzewo przycinamy $\mathbf{n}$ na całkiem nowym zbiorze uczącym R. Niech $e_R(\mathbf{l})$ i $e_R(\mathbf{n})$ oznaczają odpowiednio błędy na zbiorze R liścia i węzła. Przez błąd węzła rozumiemy błąd pod-drzewa o korzeniu $\mathbf{n}$. Wówczas jeśli zachodzi warunek

$$e_R(\mathbf{l}) \leq e_R(\mathbf{n}), \tag{4.16}$$

to zaleca się zastąpić węzeł $\mathbf{n}$ liściem $\mathbf{l}$.

[2] prawie wszystkie są w jednej klasie

4.6.2 Przycinanie minimalizujące błąd

Przycinanie minimalizujące błąd opiera się na spostrzeżeniu, że błąd drzewa przyciętego charakteryzuje się zbyt pesymistyczną oceną i dlatego wymaga korekty. Węzeł drzewa klasyfikacyjnego $\mathbf{n}$ zastępujemy liściem $\mathbf{l}$, jeśli

$$\hat{e}_T(\mathbf{l}) \leq \hat{e}_T(\mathbf{n}), \tag{4.17}$$

gdzie

$$\hat{e}_T(\mathbf{n}) = \sum_{\mathbf{n}' \in N(\mathbf{n})} \frac{|T_{\mathbf{n}'}|}{|T_{\mathbf{n}}|} \hat{e}_T(\mathbf{n}'), \tag{4.18}$$

a $N(\mathbf{n})$ jest zbiorem wszystkich możliwych węzłów potomnych węzła $\mathbf{n}$ i

$$\hat{e}_T(\mathbf{l}) = 1 - \frac{|\{x \in T_{\mathbf{l}} | c(x) = d_{\mathbf{l}}\}| + mp}{|T_{\mathbf{l}}| + m}, \tag{4.19}$$

gdzie p jest prawdopodobieństwem przynależności do klasy $d_{\mathbf{l}}$ ustalona na podstawie zewnętrznej wiedzy (gdy jej nie posiadamy przyjmujemy $p = 1/|C|$).

W przypadku drzewa regresyjnego znajdujemy wiele analogii, ponieważ jeśli dla pewnego zbioru rekordów T spełniony jest warunek

$$\mathrm{mse}_T(\mathbf{l}) \leq \mathrm{mse}_T(\mathbf{n}), \tag{4.20}$$

gdzie $\mathbf{l}$ i $\mathbf{n}$ oznaczają odpowiednio liść i węzeł, to wówczas zastępujemy węzeł $\mathbf{n}$ przez liść $\mathbf{l}$.

Estymatory wyznaczone na podstawie niewielkiej próby, mogą być obarczone znaczącym błędem. Wyliczanie błędu średnio-kwadratowego dla podzbioru nowych wartości może się charakteryzować takim obciążeniem. Dlatego stosuje się statystyki opisowe z poprawką, której pochodzenie może mieć trzy źródła: wiedza merytoryczna na temat szukanej wartości, założeń modelu lub na podstawie wyliczeń opartych o cały zbiór wartości.

Skorygowany estymator błędu średnio-kwadratowego ma następującą postać

$$\widehat{\mathrm{mse}}_T(\mathbf{l}) = \frac{\sum_{x \in T} (f(x) - m_{\mathbf{l},m,m_0}(f))^2 + mS_0^2}{|T_{\mathbf{l}}| + m}, \tag{4.21}$$

gdzie

$$m_{\mathbf{l},m,m_0}(f) = \frac{\sum_{x \in T_{\mathbf{l}}} f(x) + mm_0}{|T_{\mathbf{l}}| + m}, \tag{4.22}$$

a m_0 i S_0^2 są średnią i wariancją wyznaczonymi na całej próbie uczącej. Błąd średnio-kwadratowy węzła $\mathbf{n}$ ma postać

$$\widehat{\mathrm{mse}}_T(\mathbf{n}) = \sum_{\mathbf{n}' \in N(\mathbf{n})} \frac{|T_{\mathbf{n}'}|}{|T_{\mathbf{n}}|} \widehat{\mathrm{mse}}_T(\mathbf{n}'). \tag{4.23}$$

Wówczas kryterium podcięcia można zapisać w następujący sposób

$$\widehat{\mathrm{mse}}_T(\mathbf{l}) \leq \widehat{\mathrm{mse}}_T(\mathbf{n}) \tag{4.24}$$

4.6.3 Przycinanie ze względu na współczynnik złożoności drzewa

Przycinanie ze względu na współczynnik złożoności drzewa (ang. *cost-complexity pruning*) polega na wprowadzeniu "kary" za zwiększoną złożoność drzewa. Drzewa klasyfikacyjne przycinamy gdy spełniony jest warunek

$$e_T(\mathbf{l}) \leq e_T(\mathbf{n}) + \alpha C(\mathbf{n}), \tag{4.25}$$

gdzie $C(\mathbf{n})$ oznacza złożoność drzewa mierzoną liczbą liści, a α parametrem wagi kary za złożoność drzewa.

Wspomniane kryterium przycięcia dla drzew regresyjnych bazuje na względnym błędzie średniokwadratowym (ang. *relative square error*), czyli

$$\widehat{\mathrm{rse}}_T(\mathbf{n}) = \frac{|T|\widehat{\mathrm{mse}}_T(\mathbf{n})}{(|T|-1)S_T^2(f)}, \tag{4.26}$$

gdzie T oznacza podzbiór X, S_T^2 wariancję na zbiorze T. Wówczas kryterium podcięcia wygląda następująco

$$\widehat{\mathrm{rse}}_T(\mathbf{l}) \leq \widehat{\mathrm{rse}}_T(\mathbf{n}) + \alpha C(\mathbf{n}). \tag{4.27}$$

4.7 Obsługa braków danych

Drzewa decyzyjne wyjątkowo dobrze radzą sobie z obsługa zbiorów z brakami. Stosowane są głównie dwie strategie:

- udziałów obserwacji (ang. *fractional instances*) - rozważane są wszystkie możliwe podziały dla brakującej obserwacji i przypisywana jest im odpowiednia waga lub prawdopodobieństwo, w oparciu o zaobserwowany rozkład znanych obserwacji. Te same wagi są stosowane do predykcji wartości na podstawie drzewa z brakami danych.
- podziałów zastępczych (ang. *surrogate splits*) - jeśli wynik podziału nie może być ustalony dla obserwacji z brakami, to używany jest podział zastępczy (pierwszy), jeśli i ten nie może zostać ustalony, to stosuje się kolejny. Kolejne podziały zastępcze są generowane tak, aby wynik podziału możliwie najbardziej przypominał podział właściwy.

4.8 Zalety i wady

4.8.1 Zalety

- łatwe w interpretacji;
- nie wymagają żmudnego przygotowania danych (brak standaryzacji, wprowadzania zmiennych binarnych, dopuszcza występowanie braków danych);
- działa na obu typach zmiennych - jakościowych i ilościowych;
- dopuszcza nieliniowość związku między zmienną wynikową a predyktorami;
- odporny na odstępstwa od założeń;
- pozwala na obsługę dużych zbiorów danych.

4.8.2 Wady

- brak jawnej postaci zależności;
- zależność struktury drzewa od użytego algorytmu;
- przegrywa jakością predykcji z innymi metodami nadzorowanego uczenia maszynowego.

4.9 Przykład

Przykładem zastosowania drzew decyzyjnych będzie klasyfikacja irysów na podstawie długości i szerokości kielicha i płatka.

```
library(tidyverse)
library(rpart) # pakiet do tworzenia drzew typu CART
library(rpart.plot) # pakiet do rysowania drzew
```

Każde zadanie ucznia maszynowego zaczynamy od czyszczenia danych i odpowiedniego ich przygotowania ale w tym przypadku skupimy się jedynie na budowie, optymalizacji i ewaluacji modelu.

4.9.1 Podział zbioru na próbę uczącą i testową

```
set.seed(44)
dt.train <- iris %>%
    sample_frac(size = 0.7)
dt.test <- setdiff(iris, dt.train)
str(dt.train)
```

```
## 'data.frame':    105 obs. of  5 variables:
##  $ Sepal.Length: num  6.4 4.4 6.6 5.4 5 5.4 5.6 4.4 5.4 6.1 ...
##  $ Sepal.Width : num  2.7 3.2 3 3 3.6 3.4 2.9 2.9 3.9 2.9 ...
##  $ Petal.Length: num  5.3 1.3 4.4 4.5 1.4 1.7 3.6 1.4 1.3 4.7 ...
##  $ Petal.Width : num  1.9 0.2 1.4 1.5 0.2 0.2 1.3 0.2 0.4 1.4 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 3 1 2 2 1 1 2 1 1 2 ...
```

```
str(dt.test)
```

```
## 'data.frame':    45 obs. of  5 variables:
##  $ Sepal.Length: num  4.7 4.6 5.4 4.8 5.8 5.1 5.1 5.1 5 5.2 ...
##  $ Sepal.Width : num  3.2 3.1 3.9 3.4 4 3.8 3.7 3.3 3 3.5 ...
##  $ Petal.Length: num  1.3 1.5 1.7 1.6 1.2 1.5 1.5 1.7 1.6 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.4 0.2 0.2 0.3 0.4 0.5 0.2 0.2 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

4.9.2 Budowa drzewa

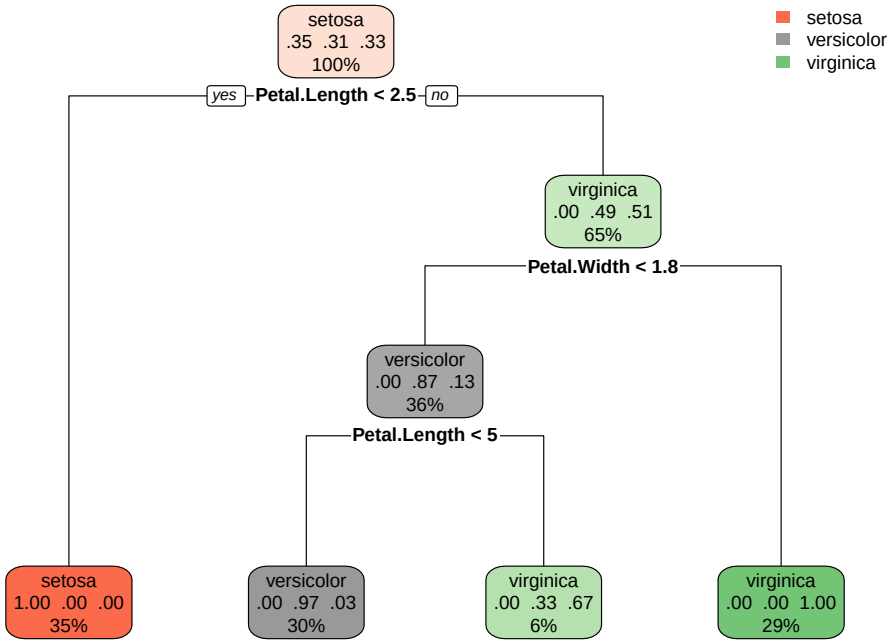
Budowy drzewa dokonujemy za pomocą funkcji `rpart` pakietu **rpart** (Therneau and Atkinson, 2018) stosując zapis formuły zależności. Drzewo zostanie zbudowane z uwzględnieniem kilku kryteriów zatrzymania:

- minimalna liczebność węzła, który może zostać podzielony to 10 - ze względu na małą liczebność zbioru uczącego;
- minimalna liczebność liścia to 5 - aby nie dopuścić do przeuczenia modelu;
- maksymalna głębokość drzewa to 4 - aby nie dopuścić do przeuczenia modelu.

```
mod.rpart <- rpart(Species~., data = dt.train,
                   control = rpart.control(minsplit = 10,
                                           minbucket = 5,
                                           maxdepth = 4))
summary(mod.rpart)
```

```
## Call:
## rpart(formula = Species ~ ., data = dt.train, control = rpart.control(minsplit = 10,
##     minbucket = 5, maxdepth = 4))
##   n= 105
##
##           CP nsplit  rel error    xerror       xstd
## 1 0.51470588      0 1.00000000 1.1764706 0.06418173
## 2 0.41176471      1 0.48529412 0.6617647 0.07457243
## 3 0.02941176      2 0.07352941 0.1029412 0.03758880
## 4 0.01000000      3 0.04411765 0.1029412 0.03758880
##
## Variable importance
##  Petal.Width Petal.Length Sepal.Length  Sepal.Width
##           35           33           21           12
##
## Node number 1: 105 observations,    complexity param=0.5147059
```

```
##   predicted class=setosa      expected loss=0.647619  P(node) =1
##     class counts:    37    33    35
##    probabilities: 0.352 0.314 0.333
##   left son=2 (37 obs) right son=3 (68 obs)
##   Primary splits:
##       Petal.Length < 2.45 to the left,  improve=35.95322, (0 missing)
##       Petal.Width  < 0.8  to the left,  improve=35.95322, (0 missing)
##       Sepal.Length < 5.45 to the left,  improve=25.39467, (0 missing)
##       Sepal.Width  < 3.35 to the right, improve=12.69596, (0 missing)
##   Surrogate splits:
##       Petal.Width  < 0.8  to the left,  agree=1.000, adj=1.000, (0 split)
##       Sepal.Length < 5.45 to the left,  agree=0.924, adj=0.784, (0 split)
##       Sepal.Width  < 3.35 to the right, agree=0.819, adj=0.486, (0 split)
##
## Node number 2: 37 observations
##   predicted class=setosa      expected loss=0  P(node) =0.352381
##     class counts:    37     0     0
##    probabilities: 1.000 0.000 0.000
##
## Node number 3: 68 observations,    complexity param=0.4117647
##   predicted class=virginica   expected loss=0.4852941  P(node) =0.647619
##     class counts:     0    33    35
##    probabilities: 0.000 0.485 0.515
##   left son=6 (38 obs) right son=7 (30 obs)
##   Primary splits:
##       Petal.Width  < 1.75 to the left,  improve=25.286380, (0 missing)
##       Petal.Length < 4.75 to the left,  improve=24.879360, (0 missing)
##       Sepal.Length < 5.75 to the left,  improve= 6.713875, (0 missing)
##       Sepal.Width  < 3.25 to the left,  improve= 1.336180, (0 missing)
##   Surrogate splits:
##       Petal.Length < 4.75 to the left,  agree=0.882, adj=0.733, (0 split)
##       Sepal.Length < 6.15 to the left,  agree=0.721, adj=0.367, (0 split)
##       Sepal.Width  < 3.15 to the left,  agree=0.618, adj=0.133, (0 split)
##
## Node number 6: 38 observations,    complexity param=0.02941176
##   predicted class=versicolor  expected loss=0.1315789  P(node) =0.3619048
##     class counts:     0    33     5
##    probabilities: 0.000 0.868 0.132
##   left son=12 (32 obs) right son=13 (6 obs)
##   Primary splits:
##       Petal.Length < 4.95 to the left,  improve=4.0800440, (0 missing)
##       Petal.Width  < 1.45 to the left,  improve=1.2257490, (0 missing)
##       Sepal.Width  < 2.65 to the right, improve=0.6168705, (0 missing)
##       Sepal.Length < 5.95 to the left,  improve=0.4736842, (0 missing)
##   Surrogate splits:
##       Petal.Width < 1.55 to the left,  agree=0.868, adj=0.167, (0 split)
##
## Node number 7: 30 observations
##   predicted class=virginica   expected loss=0  P(node) =0.2857143
##     class counts:     0     0    30
##    probabilities: 0.000 0.000 1.000
##
## Node number 12: 32 observations
##   predicted class=versicolor  expected loss=0.03125  P(node) =0.3047619
```



Rysunek 4.2: Obraz drzewa klasyfikacyjnego.

```
##     class counts:     0    31     1
##    probabilities: 0.000 0.969 0.031
##
## Node number 13: 6 observations
##   predicted class=virginica   expected loss=0.3333333  P(node) =0.05714286
##     class counts:     0     2     4
##    probabilities: 0.000 0.333 0.667
```

```
rpart.plot(mod.rpart)
```

Powyższy wykres przedstawia strukturę drzewa klasyfikacyjnego. Kolorami są oznaczone klasy, które w danym węźle dominują. Nasycenie barwy decyduje o sile tej dominacji. W każdym węźle podana jest klasa, do której najprawdopodobniej należą jego obserwacje. Ponadto podane są proporcje przynależności do klas zmiennej wynikowej oraz procent obserwacji zbioru uczącego należących do danego węzła. Pod każdym węzłem podana jest reguła podziału.

4.9.3 Przycinanie drzewa

Zanim przystąpimy do przycinania drzewa należy sprawdzić, jakie są zdolności generalizacyjne modelu. Oceny tej dokonujemy najczęściej sprawdzając macierz klasyfikacji.

```
pred.prob <- predict(mod.rpart,
                     newdata = dt.test)
```

```
pred.prob[10:20,]
```

```
##    setosa versicolor virginica
## 10      1    0.00000   0.00000
## 11      1    0.00000   0.00000
## 12      1    0.00000   0.00000
## 13      1    0.00000   0.00000
## 14      0    0.96875   0.03125
## 15      0    0.96875   0.03125
## 16      0    0.96875   0.03125
## 17      0    0.96875   0.03125
## 18      0    0.96875   0.03125
## 19      0    0.96875   0.03125
## 20      0    0.00000   1.00000
```

```
pred.class <- predict(mod.rpart,
                      newdata = dt.test,
                      type = "class")
pred.class
```

```
##          1          2          3          4          5          6
##     setosa     setosa     setosa     setosa     setosa     setosa
##          7          8          9         10         11         12
##     setosa     setosa     setosa     setosa     setosa     setosa
##         13         14         15         16         17         18
##     setosa versicolor versicolor versicolor versicolor versicolor
##         19         20         21         22         23         24
## versicolor  virginica versicolor versicolor versicolor versicolor
##         25         26         27         28         29         30
## versicolor versicolor versicolor versicolor versicolor versicolor
##         31         32         33         34         35         36
##  virginica  virginica  virginica  virginica  virginica  virginica
##         37         38         39         40         41         42
##  virginica  virginica  virginica  virginica  virginica  virginica
##         43         44         45
##  virginica  virginica  virginica
## Levels: setosa versicolor virginica
```

```
tab <- table(predykcja = pred.class, obserwacja = dt.test$Species)
tab
```

```
##             obserwacja
## predykcja    setosa versicolor virginica
##   setosa         13          0         0
##   versicolor      0         16         0
##   virginica       0          1        15
```

Jak widać z powyższej tabeli, model całkiem dobrze radzi sobie z poprawną klasyfikacją obserwacji do odpowiednich kategorii. Tylko jedna obserwacja została błędnie zaklasyfikowana.

W dalszej kolejności sprawdzimy, czy nie jest konieczne przycięcie drzewa. Jednym z kryteriów przycinania drzewa jest przycinanie ze względu na złożoność drzewa. W tym przypadku jest wyrażony parametrem `cp`. Istnieje powszechnie stosowana reguła jednego odchylenia standardowego, która mówi, że drzewo należy przyciąć wówczas, gdy błąd oszacowany na podstawie sprawdzianu krzyżowego (`xerror`), pierwszy raz zejdzie poniżej poziomu wyznaczonego przez najniższą wartość błędu powiększonego o odchylenie standardowe tego błędu (`xstd`). Na podstawie poniższej tabeli można ustalić, że poziomem odcięcia jest wartość

$0.10294 + 0.037589 = 0.140529$. Pierwszy raz błąd przyjmuje wartość mniejszą od 0.140529 po drugim podziale (`nsplit=2`). Temu poziomowi odpowiada `cp` o wartości 0.029412 i to jest złożoność drzewa, którą powinniśmy przyjąć do przycięcia drzewa.

```
printcp(mod.rpart)
```

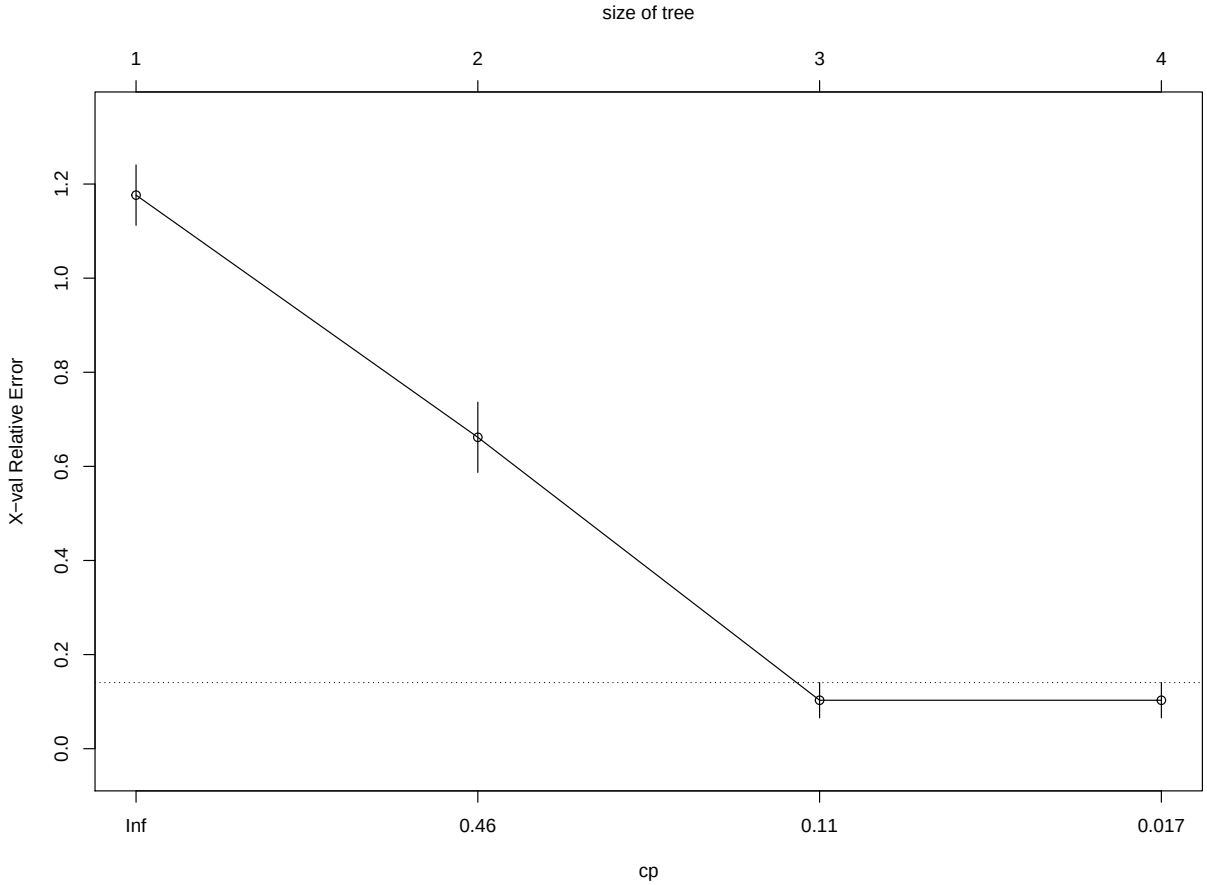
```
##
## Classification tree:
## rpart(formula = Species ~ ., data = dt.train, control = rpart.control(minsplit = 10,
##     minbucket = 5, maxdepth = 4))
##
## Variables actually used in tree construction:
## [1] Petal.Length Petal.Width
##
## Root node error: 68/105 = 0.64762
##
## n= 105
##
##         CP nsplit rel error  xerror     xstd
## 1 0.514706      0  1.000000 1.17647 0.064182
## 2 0.411765      1  0.485294 0.66176 0.074572
## 3 0.029412      2  0.073529 0.10294 0.037589
## 4 0.010000      3  0.044118 0.10294 0.037589
```

```
plotcp(mod.rpart)
```

Przycięte drzewo wygląda następująco:

```
mod.rpart2 <- prune(mod.rpart, cp = 0.029412)
summary(mod.rpart2)
```

```
## Call:
## rpart(formula = Species ~ ., data = dt.train, control = rpart.control(minsplit = 10,
##     minbucket = 5, maxdepth = 4))
##   n= 105
##
##          CP nsplit  rel error    xerror       xstd
## 1 0.5147059      0 1.00000000 1.1764706 0.06418173
## 2 0.4117647      1 0.48529412 0.6617647 0.07457243
## 3 0.0294120      2 0.07352941 0.1029412 0.03758880
##
## Variable importance
##  Petal.Width Petal.Length Sepal.Length  Sepal.Width
##           35           31           22           12
##
## Node number 1: 105 observations,    complexity param=0.5147059
##   predicted class=setosa      expected loss=0.647619  P(node) =1
##     class counts:    37    33    35
##    probabilities: 0.352 0.314 0.333
##   left son=2 (37 obs) right son=3 (68 obs)
##   Primary splits:
##       Petal.Length < 2.45 to the left,  improve=35.95322, (0 missing)
##       Petal.Width  < 0.8  to the left,  improve=35.95322, (0 missing)
##       Sepal.Length < 5.45 to the left,  improve=25.39467, (0 missing)
##       Sepal.Width  < 3.35 to the right, improve=12.69596, (0 missing)
##   Surrogate splits:
```



Rysunek 4.3: Na wykresie błędów punkt odcięcia zaznaczony jest linią przerywaną

```
##       Petal.Width  < 0.8  to the left,  agree=1.000, adj=1.000, (0 split)
##       Sepal.Length < 5.45 to the left,  agree=0.924, adj=0.784, (0 split)
##       Sepal.Width  < 3.35 to the right, agree=0.819, adj=0.486, (0 split)
##
## Node number 2: 37 observations
##   predicted class=setosa      expected loss=0  P(node) =0.352381
##     class counts:    37     0     0
##    probabilities: 1.000 0.000 0.000
##
## Node number 3: 68 observations,    complexity param=0.4117647
##   predicted class=virginica   expected loss=0.4852941  P(node) =0.647619
##     class counts:     0    33    35
##    probabilities: 0.000 0.485 0.515
##   left son=6 (38 obs) right son=7 (30 obs)
##   Primary splits:
##       Petal.Width  < 1.75 to the left,  improve=25.286380, (0 missing)
##       Petal.Length < 4.75 to the left,  improve=24.879360, (0 missing)
##       Sepal.Length < 5.75 to the left,  improve= 6.713875, (0 missing)
##       Sepal.Width  < 3.25 to the left,  improve= 1.336180, (0 missing)
##   Surrogate splits:
##       Petal.Length < 4.75 to the left,  agree=0.882, adj=0.733, (0 split)
##       Sepal.Length < 6.15 to the left,  agree=0.721, adj=0.367, (0 split)
##       Sepal.Width  < 3.15 to the left,  agree=0.618, adj=0.133, (0 split)
##
## Node number 6: 38 observations
##   predicted class=versicolor  expected loss=0.1315789  P(node) =0.3619048
##     class counts:     0    33     5
##    probabilities: 0.000 0.868 0.132
##
## Node number 7: 30 observations
##   predicted class=virginica   expected loss=0  P(node) =0.2857143
##     class counts:     0     0    30
##    probabilities: 0.000 0.000 1.000
```

```
rpart.plot(mod.rpart2)
```

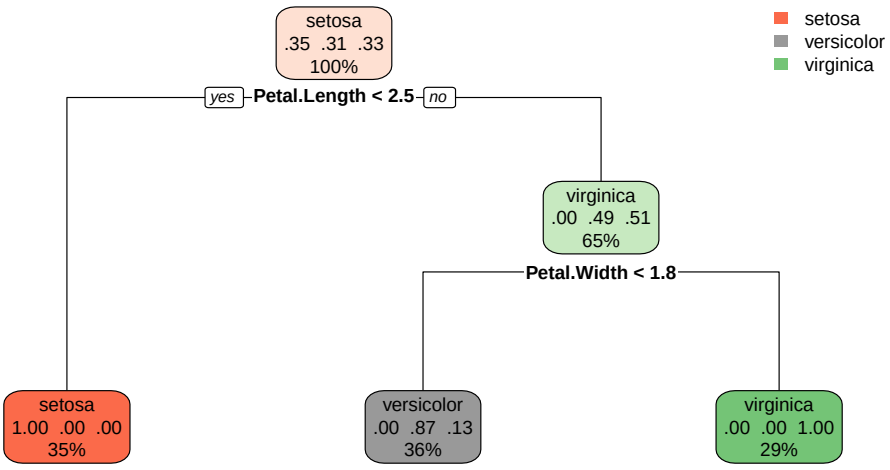
4.9.4 Ocena dopasowania modelu

Na koniec budowy modelu należy sprawdzić jego jakość na zbiorze testowym.

```
pred.class2 <- predict(mod.rpart2,
                       newdata = dt.test,
                       type = "class")
tab2 <- table(predykcja = pred.class2, obserwacja = dt.test$Species)
tab2
```

```
##             obserwacja
## predykcja    setosa versicolor virginica
##   setosa         13          0         0
##   versicolor      0         16         0
##   virginica       0          1        15
```

Mimo przycięcia drzewa, klasyfikacja pozostaje na niezmienionym poziomie. Odsetek poprawnych klasyfikacji możemy oszacować za pomocą



Rysunek 4.4: Drzewo klasyfikacyjne po przycięciu

```
round(sum(diag(tab2))/sum(tab2)*100,1)
```

```
## [1] 97.8
```

4.10 Inne algorytmy budowy drzew decyzyjnych implementowane w R

Oprócz najbardziej znanego algorytmu CART implementowanego w postaci funkcji pakietu **rpart**, istnieją również inne algorytmy, które znalzały swoje implementacje w R. Są to:

- *CHAID*[3] - algorytm przeznaczony do budowy drzew klafyfikacyjnych, gdzie zarówno zmienna wynikowa, jak i zmienne niezależne muszą być ze skali jakościowej. Główną różnicą w stosunku do drzew typu CART jest sposób budowy podziałów, oparty na teście niezależności χ^2 Pearsona. Wyboru reguły podziału dokonuje się poprzez testowanie niezależności zmiennej niezależnej z predyktorami. Reguła o największej wartości statystyki χ^2 jest stosowana w pierwszej kolejności. Implementacja tego algorytmu znajduje się w pakiecie **CHAID**[4] (funkcja do tworzenia drzewa o tej samej nazwie `chaid`) (Team, 2015).
- *Ctree*[5] - algorytm zbliżony zasadą działania do CHAID, ponieważ również wykorzystuje testowanie do wyboru reguły podziału. Różni się jednak tym, że może być stosowany do zmiennych dowolnego typu. Implementację R-ową można znaleźć w pakietach **party** (Hothorn et al., 2006) lub **partykit** (Hothorn and Zeileis, 2015) - funkcją do tworzenia modelu jest `ctree`.
- *C4.5* - algorytm stworzony przez Quinlan (1993) w oparciu, o również jego autorstwa, algorytm ID3. W dużym uproszczeniu, dobór reguł podziału odbywa się na podstawie przyrostu informacji (patrz Reguły podziału). W przeciwieństwie do pierwotnego algorytmu ID3, C4.5 nie raczej nie przeucza drzew. Implementacja R-owa znajduje się w pakiecie **RWeka** (Hornik et al., 2009) - funkcja do budowy drzewa to `J48`.
- *C5.0* - kolejny algorytm autorstwa Kuhn and Quinlan (2018) jest usprawnieniem algorytmu C4.5, generującym mniejsze drzewa automatycznie przycinane na podstawie złożności drzewa. Jest szybszy od poprzednika i pozwala na zastosowanie metody *boosting*[6]. Implementacja R-owa znajduje się w pakiecie *C50*, a funkcja do budowy drzewa to `C5.0`.

4.11 Przykład

W celu porównania wyników klasyfikacji na podstawie drzew decyzyjnych o różnych algorytmach, zostaną nauczone modele w oparciu o funkcje `ctree`, `J48` i `C5.0` dla tego samego zestawu danych co w przykładzie wcześniejszym.

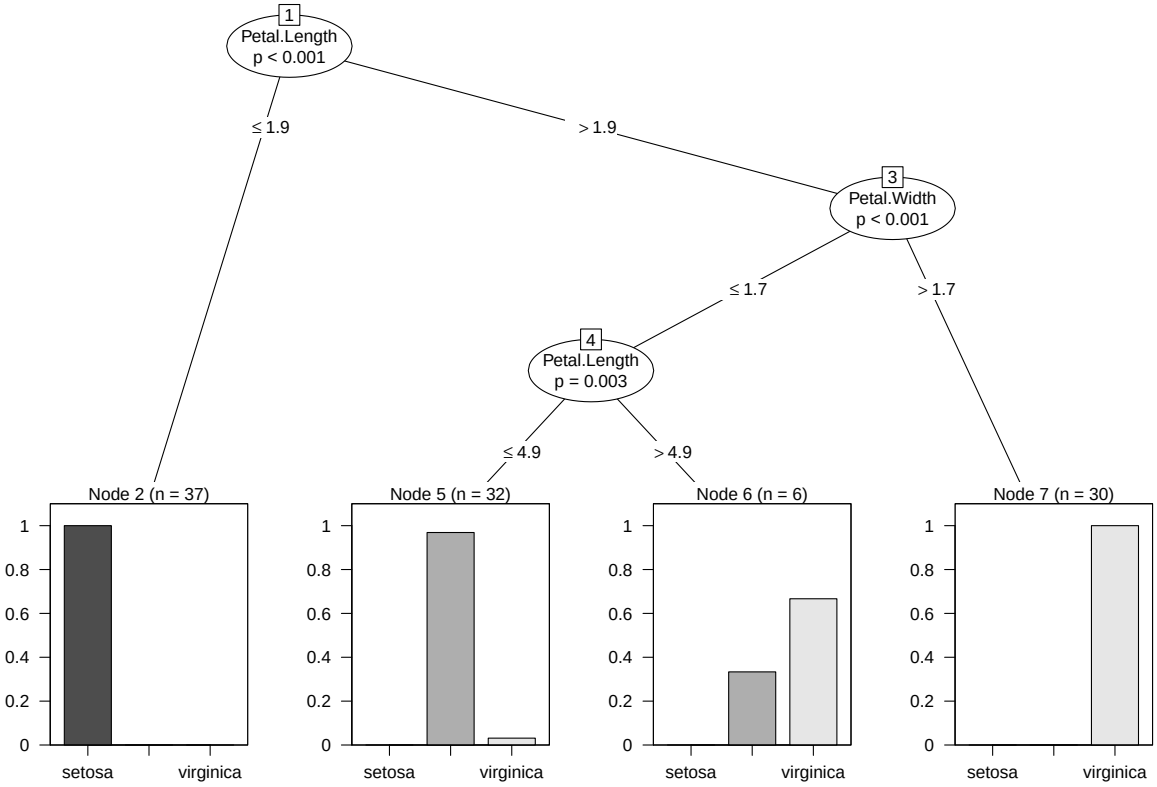
4.11.1 `ctree`

Na początek ustalamy parametry ograniczające rozrozt drzewa podobne jak w poprzednim przykładzie.

```
library(partykit)
tree2 <- ctree(Species~., data = dt.train,
               control = ctree_control(minsplit = 10,
                                       minbucket = 5,
                                       maxdepth = 4))
tree2
```

```
##
```

[3] Chi-square automatic interaction detection
[4] brak w oficjalnej dystrybucji CRAN
[5] Conditional Inference Trees
[6] budowa klasyfikatora w oparciu o proces iteracyjny, w którym kolejne w kolejnych iteracjach budowane są proste drzewa i przypisywane są im wagi - im gorszy klasyfikator, tym większa waga - po to aby nauczyć drzewo klasyfikować "trudne" przypadki



Rysunek 4.5: Wykres drzewa decyzyjnego zbudowanego metodą ctree

```
## Model formula:
## Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
##
## Fitted party:
## [1] root
## |   [2] Petal.Length <= 1.9: setosa (n = 37, err = 0.0%)
## |   [3] Petal.Length > 1.9
## |   |   [4] Petal.Width <= 1.7
## |   |   |   [5] Petal.Length <= 4.9: versicolor (n = 32, err = 3.1%)
## |   |   |   [6] Petal.Length > 4.9: virginica (n = 6, err = 33.3%)
## |   |   [7] Petal.Width > 1.7: virginica (n = 30, err = 0.0%)
##
## Number of inner nodes:    3
## Number of terminal nodes: 4
```

```
plot(tree2)
```

Wydaje się, że drzewo nie jest optymalne, ponieważ w węźle 6 obserwacje z grup `versicolor` i `virginica` są nieco pomieszane. Ostateczne oceny dokonujemy na podstawie próby testowej.

```
pred2 <- predict(tree2, newdata = dt.test)
tab <- table(predykcja = pred2, obserwacja = dt.test$Species)
tab
```

```
##             obserwacja
## predykcja    setosa versicolor virginica
```

```
##   setosa         13          0         0
##   versicolor      0         16         0
##   virginica       0          1        15
```

Dopiero ocena jakości klasyfikacji na podstawie próby testowej pokazuje, że model zbudowany za pomocą `ctree` daje podobną precyzję jak `rpart` przycięty.

4.11.2 J48

W tym przypadku model sam poszukuje optymalnego rozwiazania przycinając się automatycznie.

```
library(RWeka)
tree3 <- J48(Species~., data = dt.train)
tree3
```

```
## J48 pruned tree
## ------------------
##
## Petal.Width <= 0.6: setosa (37.0)
## Petal.Width > 0.6
## |   Petal.Width <= 1.7
## |   |   Petal.Length <= 4.9: versicolor (32.0/1.0)
## |   |   Petal.Length > 4.9
## |   |   |   Petal.Width <= 1.5: virginica (3.0)
## |   |   |   Petal.Width > 1.5: versicolor (3.0/1.0)
## |   Petal.Width > 1.7: virginica (30.0)
##
## Number of Leaves  :  5
##
## Size of the tree :   9
```

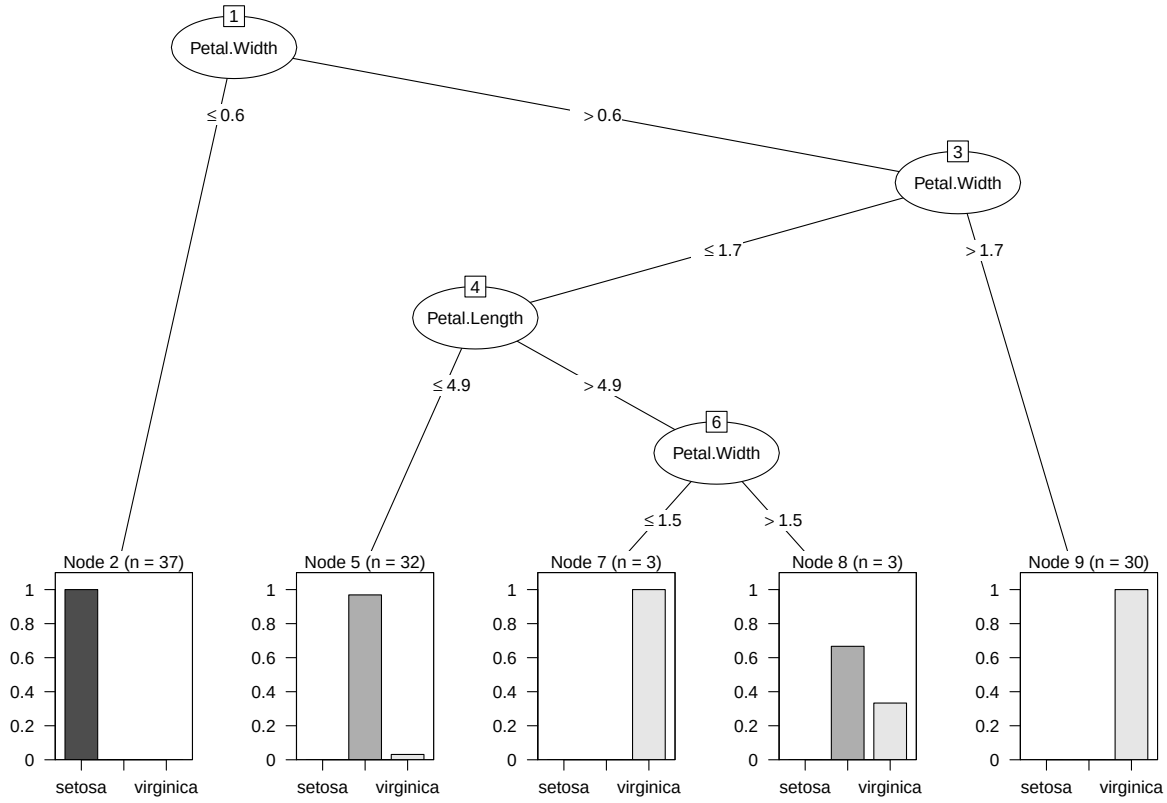
```
plot(tree3)
```

Drzewo jest nieco bardziej rozbudowane niż `tree2` i `mod.rpart2`.

```
summary(tree3)
```

```
##
## === Summary ===
##
## Correctly Classified Instances         103               98.0952 %
## Incorrectly Classified Instances         2                1.9048 %
## Kappa statistic                          0.9714
## Mean absolute error                      0.0208
## Root mean squared error                  0.1019
## Relative absolute error                  4.6776 %
## Root relative squared error             21.628  %
## Total Number of Instances              105
##
## === Confusion Matrix ===
##
##   a  b  c   <-- classified as
##  37  0  0 |  a = setosa
##   0 33  0 |  b = versicolor
##   0  2 33 |  c = virginica
```

Podsumowanie dopasowania drzewa na próbie uczącej jest bardzo dobre, bo poprawnych klasyfikacji jest



Rysunek 4.6: Wykres drzewa decyzyjnego zbudowanego metodą J48

ponad 98%. Oceny dopasowania i tak dokonujemy na zbiorze testowym.

```
pred3 <- predict(tree3, newdata = dt.test)
tab <- table(predykcja = pred3, obserwacja = dt.test$Species)
tab
```

```
##             obserwacja
## predykcja    setosa versicolor virginica
##   setosa         13          0         0
##   versicolor      0         16         0
##   virginica       0          1        15
```

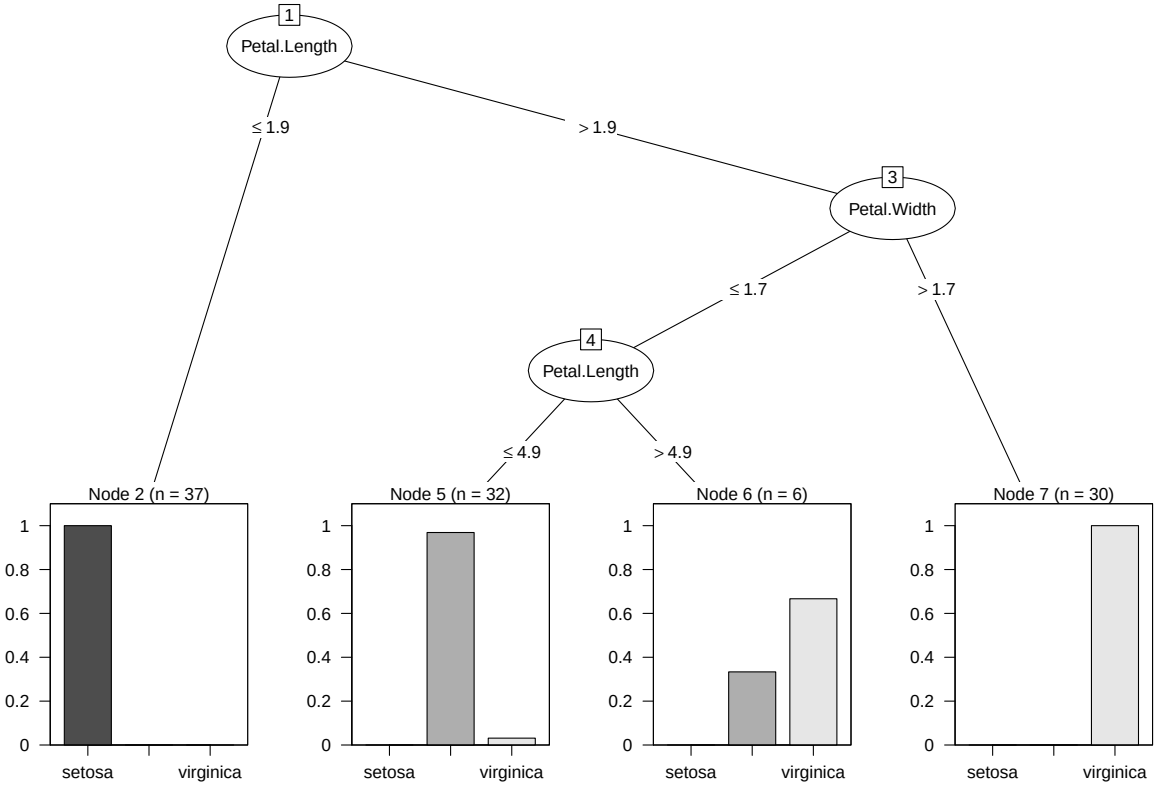
Otrzymujemy identyczną macierz klasyfikacji jak w poprzednich przypadkach.

4.11.3 C50

Tym razem również nie trzeba ustawiać parametrów drzewa, ponieważ algorytm działa tak aby zapobiec rozrostowi drzewa przy jednoczesnej wysokiej poprawności klasyfikacji.

```
library(C50)
tree4 <- C5.0(Species~., data = dt.train)
summary(tree4)
```

```
##
## Call:
## C5.0.formula(formula = Species ~ ., data = dt.train)
##
##
## C5.0 [Release 2.07 GPL Edition]      Tue Mar 12 19:23:48 2019
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 105 cases (5 attributes) from undefined.data
##
## Decision tree:
##
## Petal.Length <= 1.9: setosa (37)
## Petal.Length > 1.9:
## :...Petal.Width > 1.7: virginica (30)
##     Petal.Width <= 1.7:
##     :...Petal.Length <= 4.9: versicolor (32/1)
##         Petal.Length > 4.9: virginica (6/2)
##
##
## Evaluation on training data (105 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       4    3( 2.9%)   <<
##
##
##     (a)   (b)   (c)    <-classified as
##    ----  ----  ----
```



Rysunek 4.7: Wykres drzewa decyzyjnego zbudowanego metodą C5.0

```
##      37                (a): class setosa
##            31     2    (b): class versicolor
##             1    34    (c): class virginica
##
##
##  Attribute usage:
##
##  100.00% Petal.Length
##   64.76% Petal.Width
##
##
## Time: 0.0 secs
```

Otrzymujemy identyczne drzewo jak w przypadku zastosowania algorytmu `ctree`.

```
plot(tree4)
```

Dla pewności przeprowadzimy sprawdzenie na zbiorze testowym.

```
pred4 <- predict(tree4, newdata = dt.test)
tab <- table(predykcja = pred4, obserwacja = dt.test$Species)
tab
```

```
##             obserwacja
## predykcja    setosa versicolor virginica
##   setosa         13          0         0
```

```
##   versicolor      0         16         0
##   virginica       0          1        15
```

Rozdział 5

Pochodne drzew decyzyjnych

Przykład zastosowania drzew decyzyjnych na zbiorze `iris` w poprzednich przykładach może skłaniać do przypuszczenia, że drzewa decyzyjne zawsze dobrze radzą sobie z predykcją wartości wynikowej. Niestety w przykładach nieco bardziej skomplikowanych, gdzie chociażby klasy zmiennej wynikowej nie są tak wyraźnie separowalne, drzewa decyzjne wypadają gorzej w prównaniu z innymi modelami nadzorowanego uczenia maszynowego.

I tak u podstaw metod bazujących na prostych drzewach decyzyjnych stał pomysł, że skoro jedno drzewo nie ma wystrczających własności predykcyjnych, to może zastosowanie wielu drzew połączonych w pewien sposób poprawi je. Tak powstały metody *bagging*, *random forest* i *boosting*[1].

5.1 Bagging

Technika ta zostala wprowadzona przez Breiman (1996) i ma na celu zmniejszenie wariancji modelu pojedynczego drzewa. Podobnie jak technika *bootstrap*, w której statystyki są wyliczane na wielu próbach pobranych z tego samego rozkładu (próby), w metodzie bagging losuje się wiele prób ze zbioru uczącego (najczęściej poprzez wielokrotne losowanie próby o rozmiarze zbioru uczącego ze zwracaniem), a następnie dla każdej próby bootstrapowej buduje się drzewo. W ten sposób otrzymujemy B drzew decyzyjnych $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$. Na koniec poprzez uśrednienie otrzymujemy model charakteryzujący się większą precyzją

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x). \tag{5.1}$$

Ponieważ podczas budowy drzew na podstawie prób bootstrapowych nie kontrolujemy złożoności, to w rezultacie każde z drzew może charakteryzować się dużą wariancją. Poprzez uśrednianie wyników pojedynczych drzew otrzymujemy mniejsze obciążenie ale również przy dostatecznie dużej liczbie prób (B często liczy się w setkach, czy tysiącach) zmniejszamy wariancję "średniej" predykcji z drzew. Oczywiście metodę tą trzeba dostosować do zadań klasyfikacyjnych, ponieważ nie istnieje średnia klasyfikacji z wielu drzew. W miejsce śreniej stosuje się modę, czyli wartość dominującą.

Przyjrzyjmy się jak maszyna losuje obserwacje ze zwracaniem

```
n <- NULL
m <- NULL
for(i in 1:1000){
    x <- sample(1:500, size = 500, replace = T)
    y <- setdiff(1:500, x)
```

[1] hyba tylko dla drógiej metody istniej dobre polskie tłumaczenie nazwy - las losowy

```
    z <- unique(x)
    n[i] <- length(z)
    m[i] <- length(y)
}
mean(n)/500*100
```

```
## [1] 63.2574
```

```
mean(m)/500*100
```

```
## [1] 36.7426
```

Faktycznie uczenie modelu metodą bagging odbywa się średnio na 2/3 obserwacji zbioru uczącego wylosowanych do prób bootstrapowych, a pozostała 1/3 (ang. *out-of-bag*) jest wykorzystana do oceny jakości predykcji.

Niewątpliwą zaletą drzew decyzyjnych była ich łatwa interpretacja. W przypadku metody bagging jest ona znacznie utrudniona, ponieważ jej wynik składa się z agregacji wielu drzew. Można natomiast ocenić ważność predyktorów (ang. *variable importance*). I tak, przez obserwację spadku RSS dla baggingu regresyjnego przy zastosowaniu danego predyktora w podziałach drzewa i uśrednieniu wyniku otrzymamy wskaźnik ważności predyktora dużo lepszy niż dla pojedynczego drzewa. W przypadku baggingu klasyfikacyjnego w miejsce RSS stosujemy indeks Gini'ego.

Implementacja R-owa metody bagging znajduje się w pakiecie **ipred**, a funkcja do budowy modelu nazywa się `bagging` (Peters and Hothorn, 2018). Można również stosować funkcję `randomForest` pakietu **randomForest** (Liaw and Wiener, 2002) - powody takiego działania wyjaśnią się w rodziale Lasy losowe.

5.1.1 Przykład

Tym razem cel zadania jest regresyjny i polega na ustaleniu miary tendencji centralnej ceny mieszkań w Bostonie na podstwie zmiennych umieszczonych w zbiorze `Boston` pakietu **MASS** (Venables and Ripley, 2002). Zmienną zależną będzie mediana cen mieszkań na przedmieściach Bostonu (`medv`).

```
library(MASS)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

```
set.seed(2019)
boston.train <- Boston %>%
    sample_frac(size = 2/3)
boston.test <- setdiff(Boston, boston.train)
```

Aby móc porównać wyniki predykcji z metody bagging, najpierw zostanie zbudowane jedno drzewo decyzyjne w oparciu o algorytm CART.

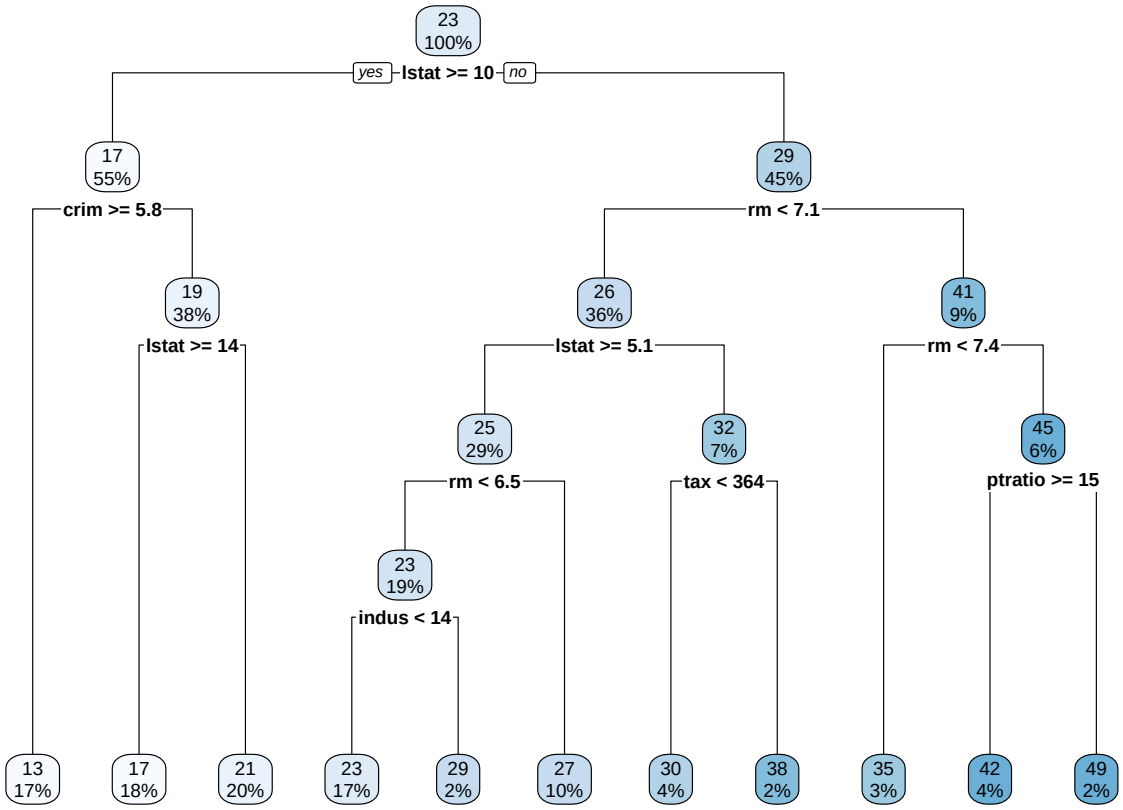
```
library(rpart)
library(rpart.plot)
boston.rpart <- rpart(medv~., data = boston.train)
x <- summary(boston.rpart)
```

```
## Call:
## rpart(formula = medv ~ ., data = boston.train)
##   n= 337
##
##            CP nsplit rel error    xerror       xstd
## 1  0.43506104      0 1.0000000 1.0037495 0.10496568
## 2  0.21114710      1 0.5649390 0.6856438 0.07732133
## 3  0.05641774      2 0.3537919 0.4393220 0.05974589
## 4  0.04154842      3 0.2973741 0.3726563 0.05716622
## 5  0.02707678      4 0.2558257 0.3520312 0.05569786
## 6  0.01489117      5 0.2287489 0.3238915 0.05681943
## 7  0.01202564      6 0.2138578 0.2922610 0.05311293
## 8  0.01057622      7 0.2018321 0.2889364 0.05318206
## 9  0.01031677      8 0.1912559 0.2838433 0.05152251
## 10 0.01006729      9 0.1809391 0.2838187 0.05152098
## 11 0.01000000     10 0.1708718 0.2815210 0.05152993
##
## Variable importance
##   lstat     nox   indus    crim     tax      rm     age     dis ptratio
##      24      13      13      13      11      10      10       2       2
##     rad   black
##       1       1
##
## Node number 1: 337 observations,    complexity param=0.435061
##   mean=22.61157, MSE=79.33004
##   left son=2 (186 obs) right son=3 (151 obs)
##   Primary splits:
##       lstat   < 10.02    to the right, improve=0.4350610, (0 missing)
##       rm      < 6.8375   to the left,  improve=0.4305766, (0 missing)
##       indus   < 6.66     to the right, improve=0.2914821, (0 missing)
##       ptratio < 19.15    to the right, improve=0.2608119, (0 missing)
##       nox     < 0.5125   to the right, improve=0.2169607, (0 missing)
##   Surrogate splits:
##       indus < 7.625    to the right, agree=0.846, adj=0.656, (0 split)
##       nox   < 0.519    to the right, agree=0.828, adj=0.616, (0 split)
##       crim  < 0.12995  to the right, agree=0.786, adj=0.523, (0 split)
##       age   < 63.9     to the right, agree=0.777, adj=0.503, (0 split)
##       tax   < 377      to the right, agree=0.769, adj=0.483, (0 split)
##
## Node number 2: 186 observations,    complexity param=0.05641774
##   mean=17.31828, MSE=19.86042
##   left son=4 (58 obs) right son=5 (128 obs)
##   Primary splits:
##       crim  < 5.84803  to the right, improve=0.4083024, (0 missing)
##       dis   < 2.0754   to the left,  improve=0.3684093, (0 missing)
##       lstat < 14.405   to the right, improve=0.3516672, (0 missing)
##       nox   < 0.657    to the right, improve=0.3255969, (0 missing)
##       age   < 84.9     to the right, improve=0.2247741, (0 missing)
##   Surrogate splits:
```

```
##       rad   < 16       to the right, agree=0.855, adj=0.534, (0 split)
##       tax   < 551.5    to the right, agree=0.839, adj=0.483, (0 split)
##       nox   < 0.657    to the right, agree=0.828, adj=0.448, (0 split)
##       dis   < 2.0754   to the left,  agree=0.801, adj=0.362, (0 split)
##       lstat < 19.055   to the right, agree=0.796, adj=0.345, (0 split)
##
## Node number 3: 151 observations,    complexity param=0.2111471
##   mean=29.13179, MSE=75.5574
##   left son=6 (120 obs) right son=7 (31 obs)
##   Primary splits:
##       rm      < 7.127    to the left,  improve=0.4947648, (0 missing)
##       lstat   < 4.495    to the right, improve=0.4054324, (0 missing)
##       nox     < 0.574    to the left,  improve=0.1389706, (0 missing)
##       ptratio < 14.75    to the right, improve=0.1349232, (0 missing)
##       age     < 89.45    to the left,  improve=0.1133301, (0 missing)
##   Surrogate splits:
##       lstat   < 3.21     to the right, agree=0.841, adj=0.226, (0 split)
##       ptratio < 14.15    to the right, agree=0.828, adj=0.161, (0 split)
##       tax     < 207      to the right, agree=0.808, adj=0.065, (0 split)
##       nox     < 0.639    to the left,  agree=0.801, adj=0.032, (0 split)
##
## Node number 4: 58 observations
##   mean=13.08793, MSE=14.14485
##
## Node number 5: 128 observations,    complexity param=0.01489117
##   mean=19.23516, MSE=10.66681
##   left son=10 (61 obs) right son=11 (67 obs)
##   Primary splits:
##       lstat   < 14.405   to the right, improve=0.2915760, (0 missing)
##       dis     < 1.99235  to the left,  improve=0.2280873, (0 missing)
##       age     < 84.15    to the right, improve=0.1950219, (0 missing)
##       ptratio < 20.95    to the right, improve=0.1349341, (0 missing)
##       rm      < 5.706    to the left,  improve=0.1194638, (0 missing)
##   Surrogate splits:
##       age   < 91.15    to the right, agree=0.758, adj=0.492, (0 split)
##       dis   < 2.0418   to the left,  agree=0.664, adj=0.295, (0 split)
##       nox   < 0.607    to the right, agree=0.633, adj=0.230, (0 split)
##       indus < 18.84    to the right, agree=0.625, adj=0.213, (0 split)
##       rm    < 5.703    to the left,  agree=0.617, adj=0.197, (0 split)
##
## Node number 6: 120 observations,    complexity param=0.04154842
##   mean=26.02417, MSE=34.39883
##   left son=12 (98 obs) right son=13 (22 obs)
##   Primary splits:
##       lstat < 5.145    to the right, improve=0.2690898, (0 missing)
##       dis   < 2.0891   to the right, improve=0.2163813, (0 missing)
##       rm    < 6.543    to the left,  improve=0.2036454, (0 missing)
##       age   < 89.45    to the left,  improve=0.1796977, (0 missing)
##       tax   < 548      to the left,  improve=0.1751322, (0 missing)
##   Surrogate splits:
##       zn    < 92.5     to the left,  agree=0.833, adj=0.091, (0 split)
##       nox   < 0.4035   to the right, agree=0.833, adj=0.091, (0 split)
##       indus < 1.495    to the right, agree=0.825, adj=0.045, (0 split)
##       dis   < 1.48495  to the right, agree=0.825, adj=0.045, (0 split)
```

```
##
## Node number 7: 31 observations,    complexity param=0.02707678
##   mean=41.16129, MSE=52.78882
##   left son=14 (11 obs) right son=15 (20 obs)
##   Primary splits:
##       rm      < 7.437    to the left,  improve=0.4423448, (0 missing)
##       lstat   < 5.185    to the right, improve=0.3125696, (0 missing)
##       ptratio < 15.05    to the right, improve=0.1896089, (0 missing)
##       black   < 392.715  to the right, improve=0.1133472, (0 missing)
##       age     < 37.6     to the right, improve=0.0737298, (0 missing)
##   Surrogate splits:
##       lstat < 4.635    to the right, agree=0.774, adj=0.364, (0 split)
##       indus < 2.32     to the left,  agree=0.742, adj=0.273, (0 split)
##       dis   < 5.9736   to the right, agree=0.710, adj=0.182, (0 split)
##       black < 390.095  to the right, agree=0.710, adj=0.182, (0 split)
##       crim  < 0.10593  to the left,  agree=0.677, adj=0.091, (0 split)
##
## Node number 10: 61 observations
##   mean=17.38689, MSE=8.122779
##
## Node number 11: 67 observations
##   mean=20.91791, MSE=7.041172
##
## Node number 12: 98 observations,    complexity param=0.01202564
##   mean=24.58265, MSE=20.9745
##   left son=24 (64 obs) right son=25 (34 obs)
##   Primary splits:
##       rm    < 6.543    to the left,  improve=0.1564077, (0 missing)
##       black < 364.385  to the right, improve=0.1331323, (0 missing)
##       age   < 89.45    to the left,  improve=0.1241124, (0 missing)
##       tax   < 223.5    to the right, improve=0.1204819, (0 missing)
##       dis   < 4.46815  to the right, improve=0.1048755, (0 missing)
##   Surrogate splits:
##       dis   < 3.6589   to the right, agree=0.704, adj=0.147, (0 split)
##       rad   < 6.5      to the left,  agree=0.704, adj=0.147, (0 split)
##       age   < 68.9     to the left,  agree=0.694, adj=0.118, (0 split)
##       indus < 1.605    to the right, agree=0.673, adj=0.059, (0 split)
##       nox   < 0.4045   to the right, agree=0.673, adj=0.059, (0 split)
##
## Node number 13: 22 observations,    complexity param=0.01031677
##   mean=32.44545, MSE=43.70884
##   left son=26 (15 obs) right son=27 (7 obs)
##   Primary splits:
##       tax   < 364      to the left,  improve=0.2868266, (0 missing)
##       lstat < 3.855    to the right, improve=0.2413545, (0 missing)
##       age   < 31.85    to the left,  improve=0.1598075, (0 missing)
##       dis   < 5.4085   to the right, improve=0.1258591, (0 missing)
##       black < 381.59   to the right, improve=0.1052855, (0 missing)
##   Surrogate splits:
##       crim  < 2.6956   to the left,  agree=0.773, adj=0.286, (0 split)
##       indus < 14       to the left,  agree=0.773, adj=0.286, (0 split)
##       nox   < 0.5875   to the left,  agree=0.773, adj=0.286, (0 split)
##       age   < 89.65    to the left,  agree=0.773, adj=0.286, (0 split)
##       dis   < 2.3371   to the right, agree=0.773, adj=0.286, (0 split)
```

```
##
## Node number 14: 11 observations
##   mean=34.64545, MSE=3.304298
##
## Node number 15: 20 observations,    complexity param=0.01057622
##   mean=44.745, MSE=43.81147
##   left son=30 (12 obs) right son=31 (8 obs)
##   Primary splits:
##       ptratio < 15.4     to the right, improve=0.3226860, (0 missing)
##       rad     < 6        to the right, improve=0.2170243, (0 missing)
##       tax     < 270      to the right, improve=0.1545997, (0 missing)
##       age     < 71.85    to the right, improve=0.1331209, (0 missing)
##       zn      < 10       to the left,  improve=0.1328727, (0 missing)
##   Surrogate splits:
##       zn   < 10       to the left,  agree=0.80, adj=0.500, (0 split)
##       nox  < 0.541    to the left,  agree=0.80, adj=0.500, (0 split)
##       age  < 86.7     to the left,  agree=0.80, adj=0.500, (0 split)
##       dis  < 2.5813   to the right, agree=0.80, adj=0.500, (0 split)
##       crim < 0.45114  to the left,  agree=0.75, adj=0.375, (0 split)
##
## Node number 24: 64 observations,    complexity param=0.01006729
##   mean=23.2625, MSE=21.96891
##   left son=48 (57 obs) right son=49 (7 obs)
##   Primary splits:
##       indus < 14.48    to the left,  improve=0.19142190, (0 missing)
##       crim  < 0.841845 to the left,  improve=0.17407590, (0 missing)
##       black < 374.635  to the right, improve=0.14590640, (0 missing)
##       dis   < 2.6499   to the right, improve=0.13374910, (0 missing)
##       age   < 79.85    to the left,  improve=0.08856433, (0 missing)
##   Surrogate splits:
##       crim  < 1.163695 to the left,  agree=0.984, adj=0.857, (0 split)
##       nox   < 0.589    to the left,  agree=0.984, adj=0.857, (0 split)
##       age   < 84.35    to the left,  agree=0.984, adj=0.857, (0 split)
##       dis   < 2.28545  to the right, agree=0.969, adj=0.714, (0 split)
##       black < 361.635  to the right, agree=0.969, adj=0.714, (0 split)
##
## Node number 25: 34 observations
##   mean=27.06765, MSE=9.646894
##
## Node number 26: 15 observations
##   mean=30.02667, MSE=14.56062
##
## Node number 27: 7 observations
##   mean=37.62857, MSE=66.76776
##
## Node number 30: 12 observations
##   mean=41.675, MSE=48.28521
##
## Node number 31: 8 observations
##   mean=49.35, MSE=1.7575
##
## Node number 48: 57 observations
##   mean=22.54386, MSE=10.87053
##
```



Rysunek 5.1: Drzewo regresyjne pełne

```
## Node number 49: 7 observations
##   mean=29.11429, MSE=73.89265
```

```
rpart.plot(boston.rpart)
```

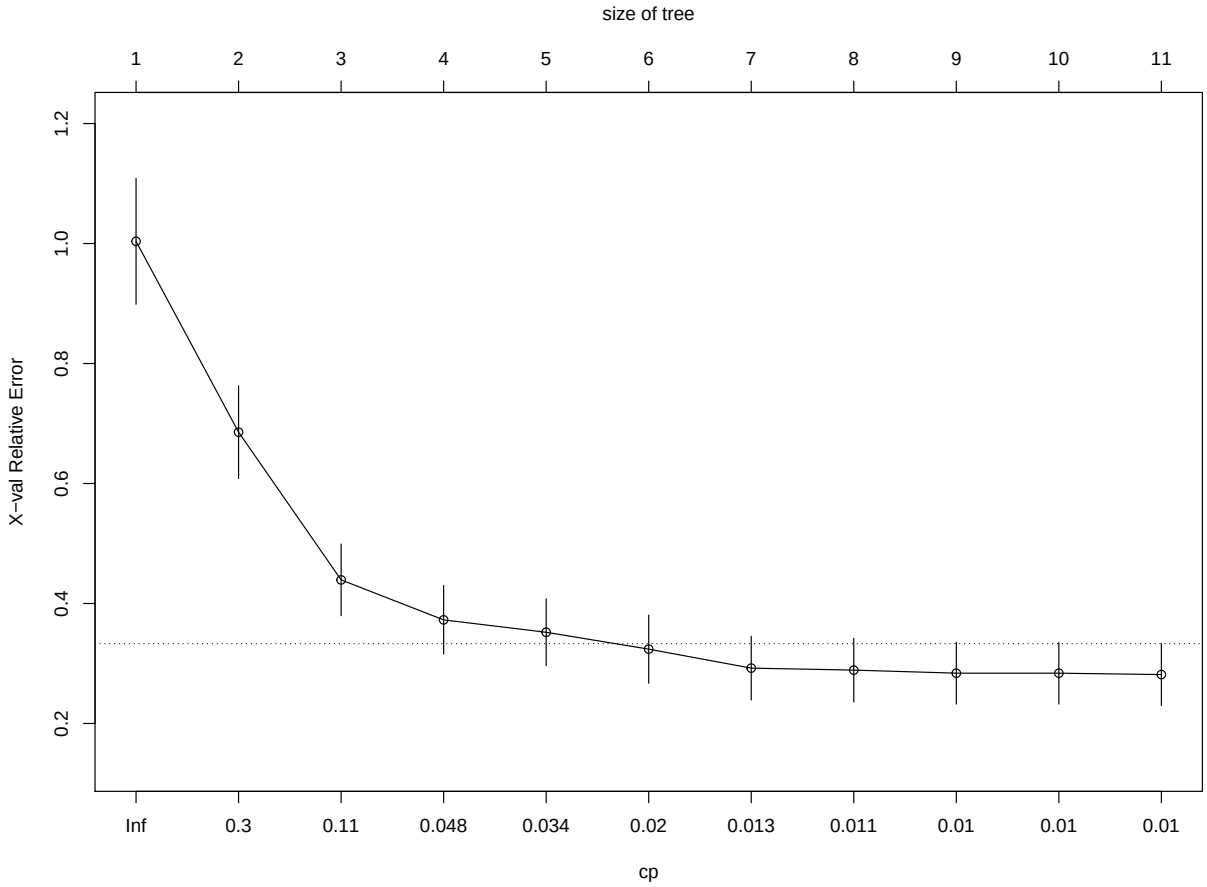
Przycinamy drzewo...

```
printcp(boston.rpart)
```

```
##
## Regression tree:
## rpart(formula = medv ~ ., data = boston.train)
##
## Variables actually used in tree construction:
## [1] crim    indus   lstat   ptratio rm      tax
##
## Root node error: 26734/337 = 79.33
##
## n= 337
##
##          CP nsplit rel error  xerror     xstd
## 1  0.435061      0   1.00000 1.00375 0.104966
## 2  0.211147      1   0.56494 0.68564 0.077321
```

```
## 3  0.056418      2   0.35379 0.43932 0.059746
## 4  0.041548      3   0.29737 0.37266 0.057166
## 5  0.027077      4   0.25583 0.35203 0.055698
## 6  0.014891      5   0.22875 0.32389 0.056819
## 7  0.012026      6   0.21386 0.29226 0.053113
## 8  0.010576      7   0.20183 0.28894 0.053182
## 9  0.010317      8   0.19126 0.28384 0.051523
## 10 0.010067      9   0.18094 0.28382 0.051521
## 11 0.010000     10   0.17087 0.28152 0.051530
```

```
plotcp(boston.rpart)
```




```
boston.rpart2 <- prune(boston.rpart, cp = 0.012026)
```

```
rpart.plot(boston.rpart2)
```

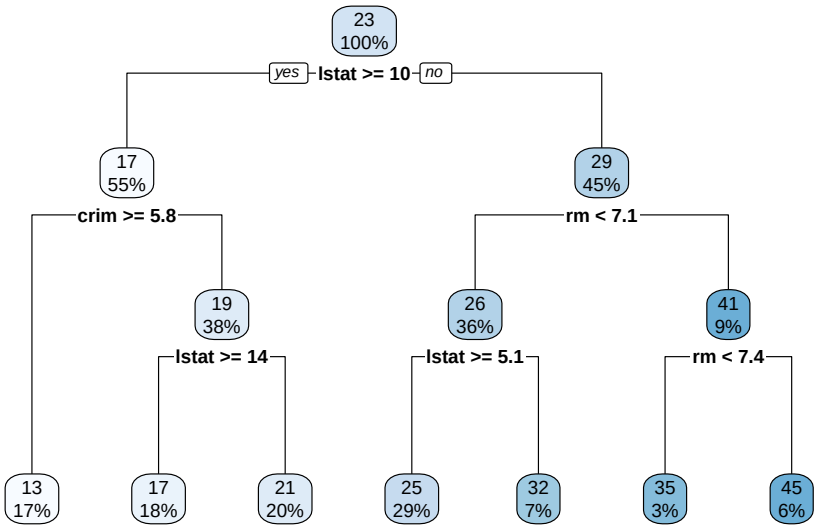
Predykcja na podstawie drzewa na zbiorze testowym.

```
boston.pred <- predict(boston.rpart2, newdata = boston.test)
rmse <- function(pred, obs) sqrt(1/length(pred)*sum((pred-obs)^2))
rmse(boston.pred, boston.test$medv)
```

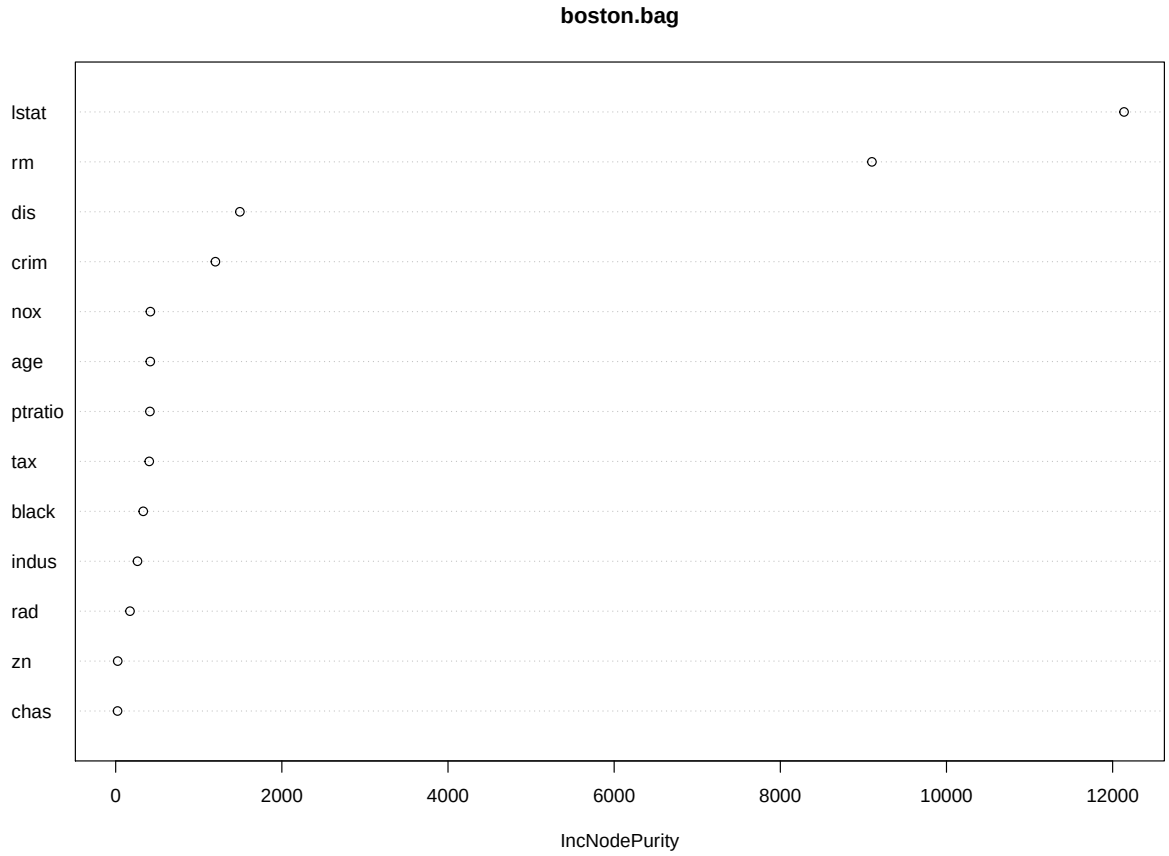
```
## [1] 4.825862
```

Teraz zbudujemy model metodą bagging.

```
library(randomForest)
boston.bag <- randomForest(medv~., data = boston.train,
```



Rysunek 5.2: Drzewo regresyjne przycięte



Rysunek 5.3: Wykres ważności predyktorów

```
                            mtry = ncol(boston.train)-1)
boston.bag
```

```
##
## Call:
##  randomForest(formula = medv ~ ., data = boston.train, mtry = ncol(boston.train) -      1)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 13
##
##           Mean of squared residuals: 13.06701
##                     % Var explained: 83.53
```

Predykcja na podstawie modelu

```
boston.pred2 <- predict(boston.bag, newdata = boston.test)
rmse(boston.pred2, boston.test$medv)
```

```
## [1] 3.039308
```

Zatem predykcja na podstwie modelu bagging jest nico lepsza niż z pojedynczego drzewa. Dodatkowo możemy ocenić ważność zmiennych użytych w budowie drzew.

```
varImpPlot(boston.bag)
```

```
importance(boston.bag)
```

```
##         IncNodePurity
## crim       1200.11828
## zn           24.17836
## indus       262.33396
## chas         22.27133
## nox         417.32236
## rm         9102.58339
## age         416.48170
## dis        1494.79734
## rad         171.92103
## tax         403.66309
## ptratio     411.88528
## black       331.58495
## lstat     12137.38999
```

```
x$variable.importance
```

```
##      lstat        nox      indus       crim        tax         rm
## 15197.8587  8683.8225  8325.2431  8074.7200  6991.0756  6768.5423
##        age        dis    ptratio        rad      black         zn
##  6538.5039  1305.3786  1193.2073   853.4309   323.8576   242.3521
```

W porównaniu do ważności zmiennych dla pojedynczego drzewa widać pewne rożnice.

5.2 Lasy losowe

Lasy losowe są uogólnieniem metody bagging, polegającą na losowaniu dla każdego drzewa wchodzącego w skład lasu m predyktorów spośród p dostępnych, a następnie budowaniu drzew z wykorzystaniem tylko tych predyktorów. Dzięki temu za każdy razem drzewo jest budowane w oparciu o nowy zestaw cech (najczęściej przyjmujemy $m = \sqrt{p}$). W przypadku modeli bagging za każdym razem najsilniejszy predyktor wchodził w skład zbioru uczacego, a co za tym idzie również uczestniczył w tworzeniu reguł podziału. Wówczas wiele drzew zawierało reguły stosujące dany atrubut, a wtedy predykcje otrzymywane za pomocą drzew były skorelowane. Dlatego nawet duża liczba prób bootstrapowych nie zapewniała poprawy precyzji.

5.2.1 Przykład

Kontynuując poprzedni przykład możemy zbudować las losowy aby przekonać się czy nastąpi poprawa predykcji zmiennej wynikowej.

```
boston.rf <- randomForest(medv~., data = boston.train)
boston.rf
```

```
##
## Call:
##  randomForest(formula = medv ~ ., data = boston.train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 13.09902
##                     % Var explained: 83.49
```

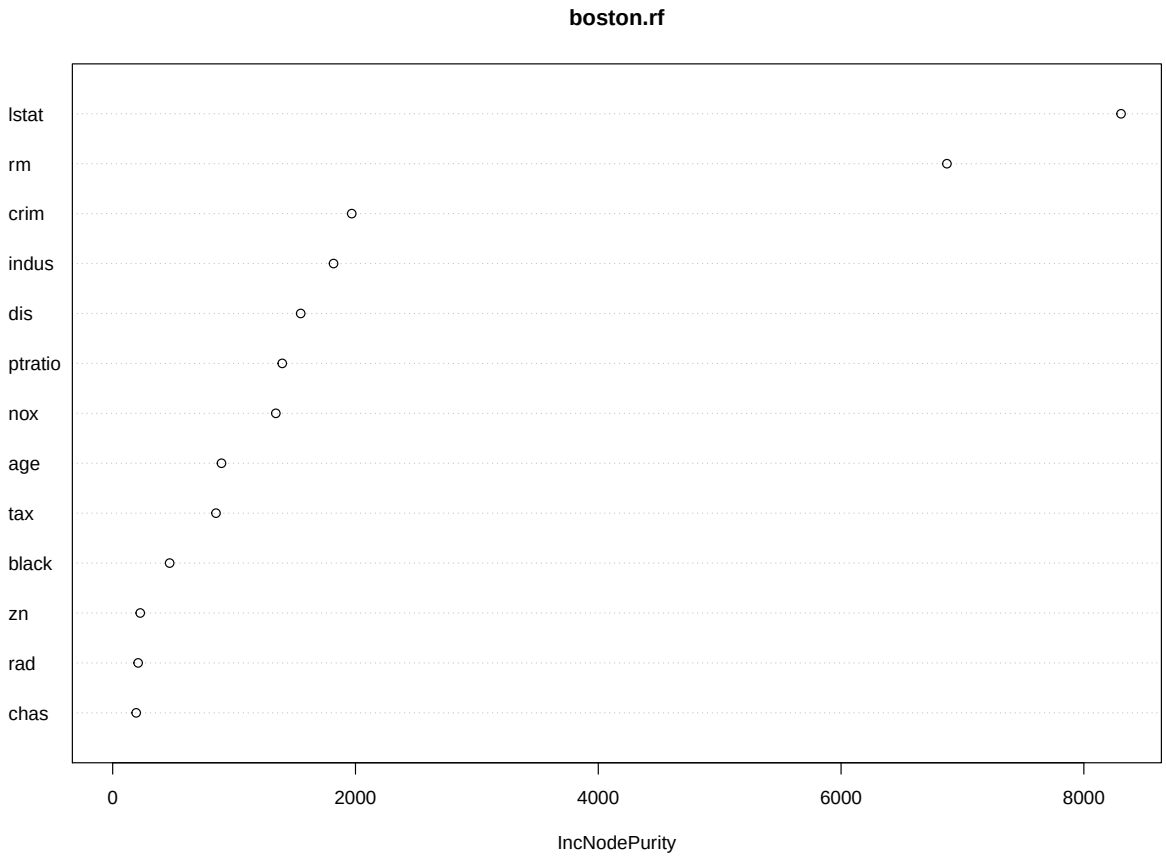
Porownanie MSE na próbach uczących pomiędzy lasem losowym i modelem bagging wypada nieco na korzyść bagging.

```
boston.pred3 <- predict(boston.rf, newdata = boston.test)
rmse(boston.pred3, boston.test$medv)
```

```
## [1] 3.418302
```

Ważność zmiennych również się nieco różni.

```
varImpPlot(boston.rf)
```



5.3 Boosting

Bibliografia

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2018). *rmarkdown: Dynamic Documents for R*. R package version 1.11.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). *Classification and Regression Trees*. Routledge.

Firke, S. (2018). *janitor: Simple Tools for Examining and Cleaning Dirty Data*. R package version 1.1.1.

hong Chan, C. and Leeper, T. J. (2018). *rio: A Swiss-Army Knife for Data I/O*. R package version 0.5.16.

Hornik, K., Buchta, C., and Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics*, 24(2):225–232.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hothorn, T. and Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in r. *Journal of Machine Learning Research*, 16:3905–3909.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15:104 – 116.

Kuhn, M. and Quinlan, R. (2018). *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.2.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Peters, A. and Hothorn, T. (2018). *ipred: Improved Predictors*. R package version 0.9-8.

Quinlan, J. R. (1993). *C4. 5: Programs for Machine Learning*. Morgan Kaufmann.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Team, T. F. S. P. (2015). *CHAID: CHi-squared Automated Interaction Detection*. R package version 0.1-2.

Templ, M., Kowarik, A., Alfons, A., and Prantner, B. (2019). *VIM: Visualization and Imputation of Missing Values*. R package version 4.8.0.

Therneau, T. and Atkinson, B. (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.

Torgo, L. (2013). *DMwR: Functions and data for "Data Mining with R"*. R package version 0.4.1.

van Buuren, S. and Groothuis-Oudshoorn, K. (2018). *mice: Multivariate Imputation by Chained Equations*. R package version 3.3.0.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Xie, Y. (2018a). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.9.

Xie, Y. (2018b). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.21.