Trenowanie modeli

Julia Janczyk

2024-05-20

CZESC I

Przygotowanie i filtracja danych

```
# 1
dane <- read.csv2("weather.csv", header = TRUE, sep=";") # NA
head(dane)
```

```
##   id DWD_ID         STATION.NAME        FEDERAL.STATE     LAT     LON ALTITUDE
## 1  0      1                Aach Baden-W\xfcrttemberg 47.8413  8.8493      478
## 2  1      3              Aachen  Nordrhein-Westfalen 50.7827  6.0941      202
## 3  2     44      Gro\xdfenkneten        Niedersachsen 52.9335  8.2370       44
## 4  6     71      Albstadt-Badkap Baden-W\xfcrttemberg 48.2156  8.9784      759
## 5  8     73 Aldersbach-Kriestorf               Bayern 48.6159 13.0506      340
## 6  9     78            Alfhausen        Niedersachsen 52.4853  7.9126       65
##      PERIOD RECORD.LENGTH MEAN.ANNUAL.AIR.TEMP MEAN.MONTHLY.MAX.TEMP
## 1 1931-1986            55                  8.2                  13.1
## 2 1851-2011           160                  9.8                  13.6
## 3 1971-2016            45                  9.2                  13.2
## 4 1986-2016            30                  7.4                  12.2
## 5 1952-2016            64                  8.4                  13.4
## 6 1961-2016            55                  9.3                  13.4
##   MEAN.MONTHLY.MIN.TEMP MEAN.ANNUAL.WIND.SPEED MEAN.CLOUD.COVER
## 1                   3.5                     2               67
## 2                   6.3                     3               67
## 3                   5.4                     2               67
## 4                   3.3                     2               66
## 5                   3.9                     1               65
## 6                   5.2                     2               67
##   MEAN.ANNUAL.SUNSHINE MEAN.ANNUAL.RAINFALL MAX.MONTHLY.WIND.SPEED MAX.AIR.TEMP
## 1                   NA                  755                      2         32.5
## 2                 1531                  820                      3         32.3
## 3                 1459                  759                      3         32.4
## 4                 1725                  919                      2         30.2
## 5                 1595                  790                      2         33.0
## 6                   NA                  794                      2         32.2
##   MAX.WIND.SPEED MAX.RAINFALL MIN.AIR.TEMP MEAN.RANGE.AIR.TEMP
## 1             NA           39        -16.3                 9.6
## 2           30.2           36        -10.9                 7.3
```

```
## 3           29.9           32        -12.6                 7.8
## 4             NA           43        -15.5                 8.9
## 5             NA           43        -19.2                 9.5
## 6             NA           33        -13.3                 8.2
```

```r
# 2
# typy zmiennych
str(dane)
```

```
## 'data.frame':    599 obs. of  22 variables:
##  $ id                     : int  0 1 2 6 8 9 10 12 14 18 ...
##  $ DWD_ID                 : int  1 3 44 71 73 78 91 98 116 132 ...
##  $ STATION.NAME           : chr  "Aach" "Aachen" "Gro\xdfenkneten" "Albstadt-Badkap" ...
##  $ FEDERAL.STATE          : chr  "Baden-W\xfcrttemberg" "Nordrhein-Westfalen" "Niedersachsen" "Baden-
##  $ LAT                    : num  47.8 50.8 52.9 48.2 48.6 ...
##  $ LON                    : num  8.85 6.09 8.24 8.98 13.05 ...
##  $ ALTITUDE               : num  478 202 44 759 340 65 300 780 213 750 ...
##  $ PERIOD                 : chr  "1931-1986" "1851-2011" "1971-2016" "1986-2016" ...
##  $ RECORD.LENGTH          : int  55 160 45 30 64 55 38 67 67 33 ...
##  $ MEAN.ANNUAL.AIR.TEMP   : num  8.2 9.8 9.2 7.4 8.4 9.3 8.2 5.1 8.4 5.7 ...
##  $ MEAN.MONTHLY.MAX.TEMP  : num  13.1 13.6 13.2 12.2 13.4 13.4 12.7 8.9 12.9 9.2 ...
##  $ MEAN.MONTHLY.MIN.TEMP  : num  3.5 6.3 5.4 3.3 3.9 5.2 4.1 2.2 4.2 2.7 ...
##  $ MEAN.ANNUAL.WIND.SPEED: num  2 3 2 2 1 2 3 3 2 3 ...
##  $ MEAN.CLOUD.COVER       : num  67 67 67 66 65 67 72 72 66 64 ...
##  $ MEAN.ANNUAL.SUNSHINE   : num  NA 1531 1459 1725 1595 ...
##  $ MEAN.ANNUAL.RAINFALL   : num  755 820 759 919 790 794 657 NA NA 915 ...
##  $ MAX.MONTHLY.WIND.SPEED: num  2 3 3 2 2 2 3 4 3 3 ...
##  $ MAX.AIR.TEMP           : num  32.5 32.3 32.4 30.2 33 32.2 31.6 27.6 33.2 29 ...
##  $ MAX.WIND.SPEED         : num  NA 30.2 29.9 NA NA NA NA NA NA NA ...
##  $ MAX.RAINFALL           : num  39 36 32 43 43 33 37 NA NA 40 ...
##  $ MIN.AIR.TEMP           : num  -16.3 -10.9 -12.6 -15.5 -19.2 -13.3 -15.2 -15.7 -17.5 -17.2 ...
##  $ MEAN.RANGE.AIR.TEMP    : num  9.6 7.3 7.8 8.9 9.5 8.2 8.6 6.7 8.6 6.5 ...
```

```r
# podsumowanie
summary(dane)
```

```
##        id             DWD_ID      STATION.NAME       FEDERAL.STATE
##  Min.   :   0.0   Min.   :    1   Length:599         Length:599
##  1st Qu.: 259.5   1st Qu.: 1368   Class :character   Class :character
##  Median : 479.0   Median : 2812   Mode  :character   Mode  :character
##  Mean   : 489.2   Mean   : 2902
##  3rd Qu.: 731.5   3rd Qu.: 4338
##  Max.   :1058.0   Max.   :15526
##
##       LAT             LON            ALTITUDE         PERIOD
##  Min.   :47.40   Min.   : 6.094   Min.   :   1.0   Length:599
##  1st Qu.:49.27   1st Qu.: 8.477   1st Qu.:  75.0   Class :character
##  Median :50.64   Median : 9.966   Median : 224.0   Mode  :character
##  Mean   :50.75   Mean   :10.120   Mean   : 285.3
##  3rd Qu.:51.96   3rd Qu.:11.703   3rd Qu.: 418.0
##  Max.   :55.01   Max.   :14.951   Max.   :2964.0
##
##  RECORD.LENGTH    MEAN.ANNUAL.AIR.TEMP MEAN.MONTHLY.MAX.TEMP
```

```
##  Min.   : 30.00   Min.   : 2.500       Min.   : 3.30
##  1st Qu.: 54.00   1st Qu.: 8.000       1st Qu.:12.10
##  Median : 70.00   Median : 8.500       Median :12.90
##  Mean   : 80.07   Mean   : 8.401       Mean   :12.66
##  3rd Qu.:103.00   3rd Qu.: 9.100       3rd Qu.:13.50
##  Max.   :297.00   Max.   :11.000       Max.   :15.60
##                   NA's   :1            NA's   :2
##  MEAN.MONTHLY.MIN.TEMP MEAN.ANNUAL.WIND.SPEED MEAN.CLOUD.COVER
##  Min.   :0.300         Min.   :1.000          Min.   :56.0
##  1st Qu.:3.800         1st Qu.:2.000          1st Qu.:65.0
##  Median :4.600         Median :2.000          Median :67.0
##  Mean   :4.488         Mean   :2.124          Mean   :66.8
##  3rd Qu.:5.300         3rd Qu.:2.000          3rd Qu.:69.0
##  Max.   :7.300         Max.   :6.000          Max.   :79.0
##  NA's   :4             NA's   :11             NA's   :11
##  MEAN.ANNUAL.SUNSHINE MEAN.ANNUAL.RAINFALL MAX.MONTHLY.WIND.SPEED
##  Min.   :   0         Min.   : 446.0       Min.   :1.000
##  1st Qu.:1441         1st Qu.: 640.2       1st Qu.:2.000
##  Median :1543         Median : 737.5       Median :3.000
##  Mean   :1517         Mean   : 787.2       Mean   :2.721
##  3rd Qu.:1635         3rd Qu.: 857.0       3rd Qu.:3.000
##  Max.   :1846         Max.   :1995.0       Max.   :7.000
##  NA's   :193          NA's   :13           NA's   :11
##   MAX.AIR.TEMP   MAX.WIND.SPEED   MAX.RAINFALL    MIN.AIR.TEMP
##  Min.   :13.90   Min.   : 3.80   Min.   :25.00   Min.   :-25.40
##  1st Qu.:31.10   1st Qu.:25.45   1st Qu.:34.00   1st Qu.:-16.70
##  Median :32.20   Median :27.50   Median :36.00   Median :-14.90
##  Mean   :31.84   Mean   :27.56   Mean   :38.55   Mean   :-14.93
##  3rd Qu.:33.10   3rd Qu.:29.50   3rd Qu.:41.00   3rd Qu.:-13.30
##  Max.   :35.40   Max.   :54.30   Max.   :76.00   Max.   : -5.30
##  NA's   :2       NA's   :380     NA's   :14      NA's   :2
##  MEAN.RANGE.AIR.TEMP
##  Min.   : 0.000
##  1st Qu.: 7.600
##  Median : 8.400
##  Mean   : 8.168
##  3rd Qu.: 8.900
##  Max.   :11.100
##
```

```
# 3
istotne = subset(dane, select = -c(LAT, LON, PERIOD, RECORD.LENGTH, FEDERAL.STATE))
```

```
# 4 obrobka danych, usuwanie pustych

liczba <- nrow(istotne)

nowe_dane <- na.omit(istotne)
liczba_bp <- nrow(nowe_dane) # liczba wierszy z pominietymi brakujacymi danymi

liczba_usunietych = liczba - liczba_bp
liczba_usunietych
```

```
## [1] 395
```

```r
# 5
wymiary <- dim(nowe_dane)
wymiary # wiersze, kolumny
```

```
## [1] 204  17
```

```r
# 6 podział zestawu
# install.packages("caTools")
library(caTools)
```

```
## Warning: pakiet 'caTools' został zbudowany w wersji R 4.3.3
```

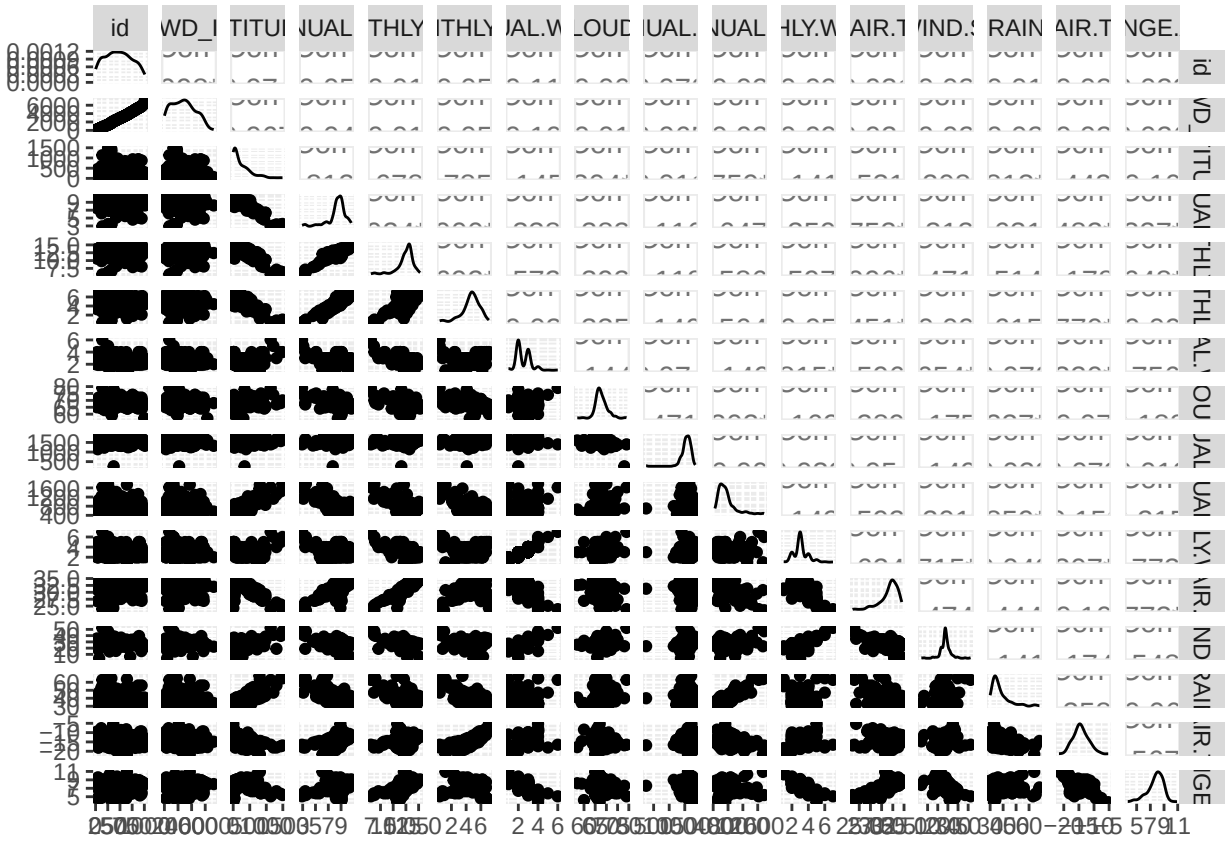
```r
split <- sample.split(nowe_dane, 0.7)

trening <- subset(nowe_dane, split == TRUE)
test <- subset(nowe_dane, split == FALSE)
```

```r
# 7   Wyodrębnij tą zmienną w postaci wektora
treningV <- trening$MEAN.ANNUAL.RAINFALL
testV <- test$MEAN.ANNUAL.RAINFALL


trening_matrix = as.matrix(subset(trening, select = -c(MEAN.ANNUAL.RAINFALL)))

test_matrix = as.matrix(subset(test, select = -c(MEAN.ANNUAL.RAINFALL)))
```



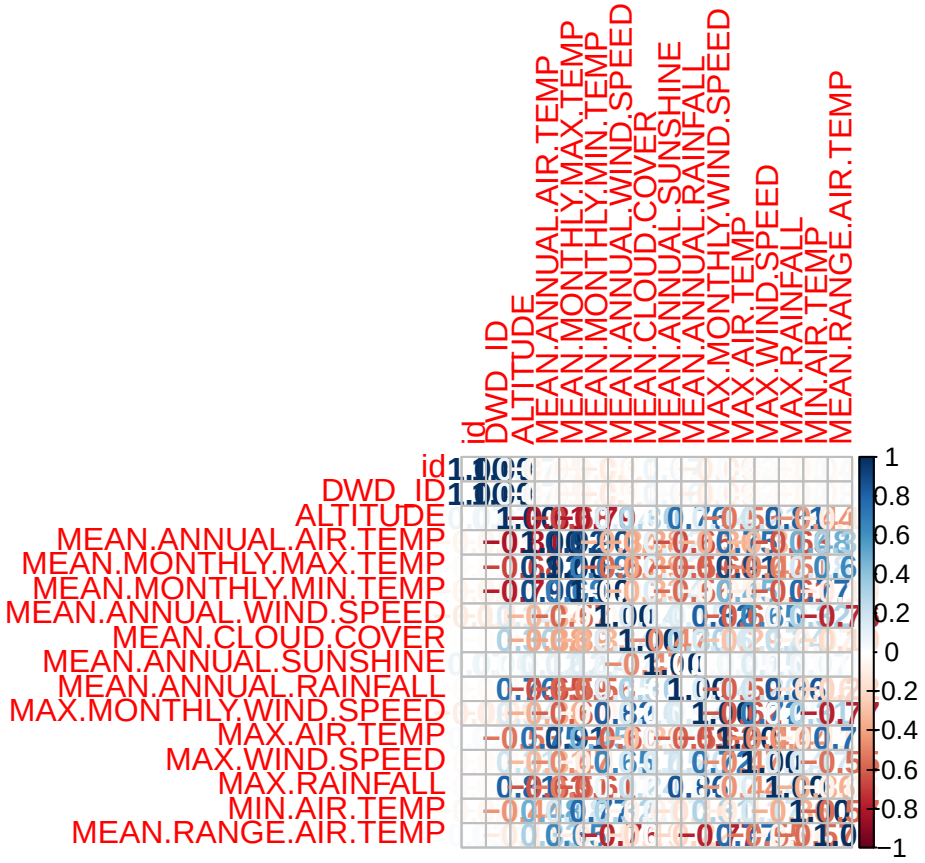
```r
# 9
corr = cor(dane8)
```

```r
# 10
library(corrplot)
```

```
## Warning: pakiet 'corrplot' został zbudowany w wersji R 4.3.2
```

```
## corrplot 0.92 loaded
```

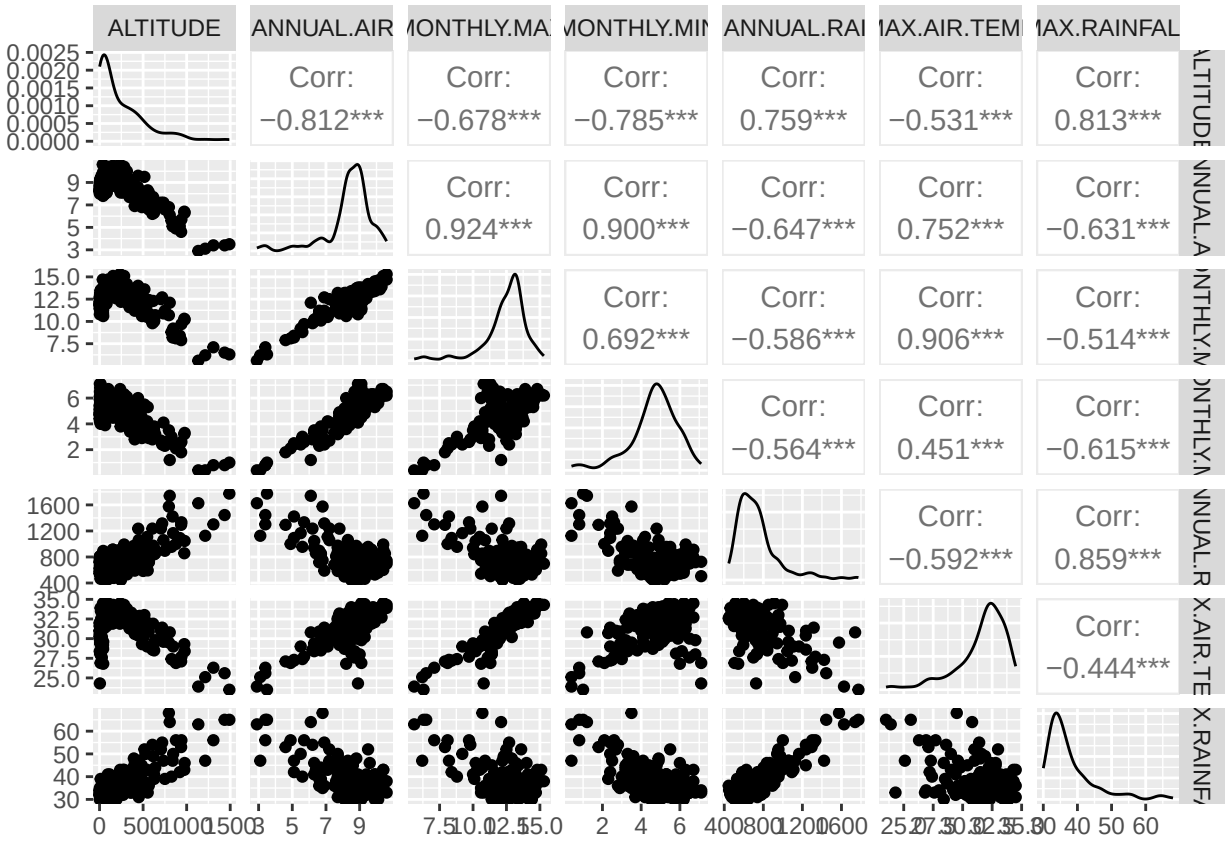
```r
corrplot(corr, method = "number") # wartość korelacji w sposób liczbowy
```




```r
# 11
# filtracja par zmiennych  z wartością współczynnika korelacji |r|>=0.5

zmienne = abs(corr['MEAN.ANNUAL.RAINFALL',]) >= 0.5
wybrane = names(zmienne[zmienne])
z11 <- nowe_dane[wybrane]
```

```r
# 12
ggpairs(z11)
```



CZESC II

Tworzenie modeli

```
# 1

# srednia opadow na wszystkich stacjach
srednia_opadow <- mean(treningV)
srednia_opadow
```

```
## [1] 738.3258
```

```
# porownaj ze srednia dla kazdej stacji
srednia_stacje <- aggregate(MEAN.ANNUAL.RAINFALL ~ STATION.NAME, data = trening, FUN = mean)
head(srednia_stacje)
```

```
##           STATION.NAME MEAN.ANNUAL.RAINFALL
## 1           G\xf6rlitz                  681
## 2 Gie\xdfen/Wettenberg                  624
## 3            L\xfcchow                  539
## 4  Gro\xdf L\xfcsewitz                  612
## 5   Bremerv\xf6rde (A)                  758
## 6         G\xf6ttingen                  608
```

```r
model_bazowy <- lm(MEAN.ANNUAL.RAINFALL ~ 1, data = trening)
```

```r
# 2

RMSE2 <- sqrt(sum((srednia_opadow - treningV)^2)/length(treningV) )
# RMSE <- sqrt(mean((srednia_opadow - treningV)^2))
RMSE2
```

```
## [1] 222.4818
```

```r
# 3
predykcje_bazowe <- predict(model_bazowy, newdata = test)
pred = srednia_opadow * length(treningV)
RMSE3 <- sqrt(mean(( predykcje_bazowe - testV)^2))
RMSE3
```

```
## [1] 247.9997
```

```r
# 4
corr_srednia <- cor(dane8)[, "MEAN.ANNUAL.RAINFALL"]
# zmienna ALTITUDE ma korelacje 0.758532239 ze srednimi opadami
model_liniowy4 <- lm(MEAN.ANNUAL.RAINFALL ~ ALTITUDE, data = trening)
RMSE4 <- sqrt(mean((predict(model_liniowy4) - srednia_opadow)^2))
```

```r
# 5
predykcje_liniowe5 <- predict(model_liniowy4, newdata = test)
RMSE5 <- sqrt(mean((predykcje_liniowe5 - srednia_opadow)^2))
RMSE5
```

```
## [1] 185.2929
```

```r
# 6
model_liniowy6 <- lm(MEAN.ANNUAL.RAINFALL ~ MAX.RAINFALL, data = trening)
RMSE6 <- sqrt(mean((predict(model_liniowy6) - srednia_opadow)^2))
RMSE6
```

```
## [1] 184.2358
```

```r
# 7
predykcje_liniowe6<- predict(model_liniowy6, newdata = test)
RMSE7 <- sqrt(mean((predykcje_liniowe5 - testV)^2))
RMSE7
```

```
## [1] 150.6471
```

```r
# 8
model_liniowy_wielokrotny8 <- lm(MEAN.ANNUAL.RAINFALL ~ ALTITUDE + MEAN.ANNUAL.AIR.TEMP + MAX.RAINFALL,
R2_8 <- summary(model_liniowy_wielokrotny8)$r.squared
RMSE8 <- sqrt(mean((predict(model_liniowy_wielokrotny8) - srednia_opadow)^2))
RMSE8
```

```
## [1] 186.9019
```

```r
# 9
predykcje_9 <- predict(model_liniowy_wielokrotny8, newdata = test)
RMSE9 <- sqrt(mean((predykcje_9 - testV)^2))
RMSE9
```

```
## [1] 103.1275
```

```r
# 10
RMSE_trening <- c(RMSE2, RMSE4, RMSE6, RMSE8)
RMSE_test <- c(RMSE3, RMSE5, RMSE7, RMSE9)

modele <- c("średnią opadów", "lin altitude", "lin max opady", "lin wielokrotny")

barplot(
  rbind(RMSE_trening, RMSE_test),
  beside = TRUE,
  names.arg = modele,
  col = c("blue", "red"),
  legend.text = c("Trening", "Test"),
  main = "Porównanie RMSE dla różnych modeli",
  xlab = "Modele",
  ylab = "Wartość RMSE"
)
```