2A. Porówanie metod klasyfikacji - pakiet R

Konrad Cendrowski

Jacek Giedronowicz

# Wstęp

Przedstawimy wyniki porównania czterech metod klasyfikacji:

* Decision Trees
* K-Nearest Neighbours
* Support Vector Machine
* Random Forest

Modele zostały zbudowane na danych z ogólnodostępnych baz tj:

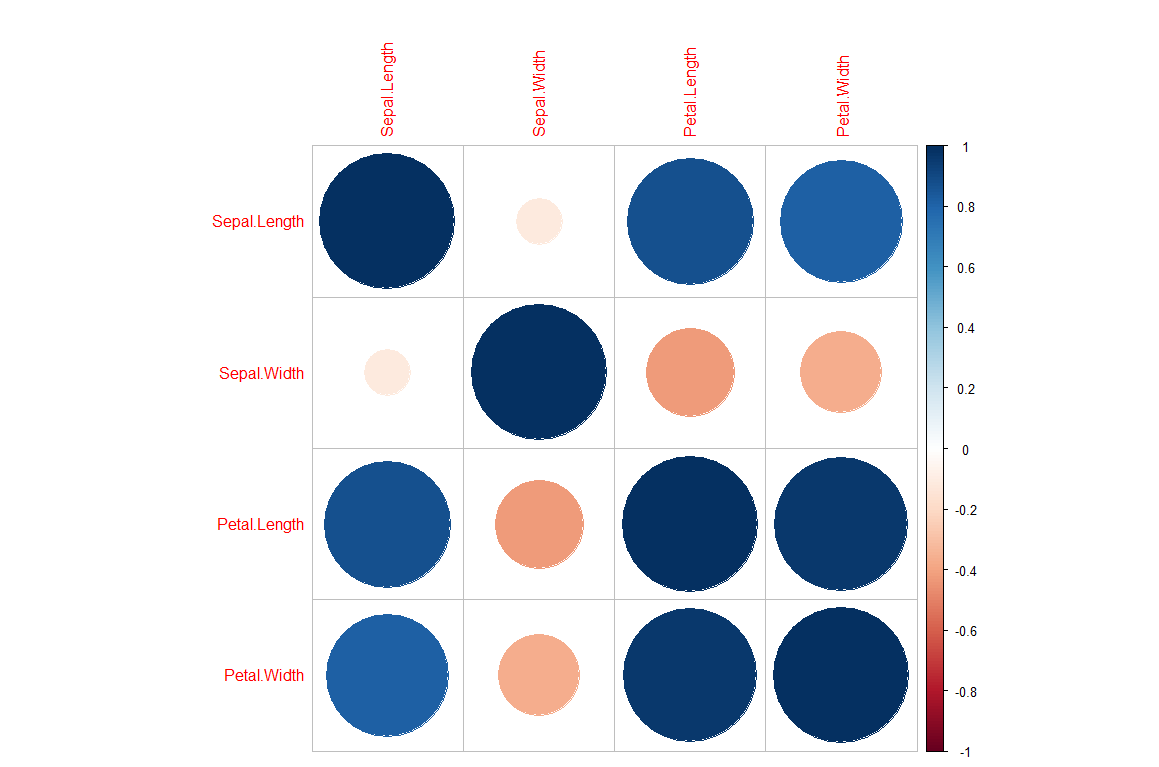
* iris
* imports-85
* adults

# Krótki opis danych

* iris – dane opisują pewne cechy kwiatów irysów tj:
  + Sepal.Length – długość działki kielicha przy kwiecie liczone w centymetrach,
  + Sepal.Width – szerokość działki kielicha przy kwiecie liczone w centymetrach,
  + Petal.Length – długość płatków kwiatu liczone w centymetrach,
  + Petal.Width – szerokość płatków kwiatu liczone w centymetrach,
  + Species – nazwa odmiany.
* imports-85 – dane opisują pewne cechy samochodów
  + symboling – wartości całkowite od -3 do 3 opisujące ryzykowność samochodu. Wartość +3 oznacza, że auto jest ryzykowne, -3 że prawdopodobnie jest całkiem bezpieczne.
  + normalized-losses – znormalizowana wartość straty w użytkowaniu w porównaniu z innymi samochodami.
  + make – nazwa producenta, który wyprodukował samochód
  + fuel-type – rodzaj paliwa
  + aspiration – możliwe dwie wartość tej zmiennej opisują czy samochód ma aspiracje na samochód sportowy (turbo). W przeciwnym wypadku przyjmuje wartość ‘std’.
  + price – cena samochodu
  + Inne fizyczne parametry samochodu takie jak :
    - num-of-doors, body-style , drive-wheels, engine-location, wheel-base, length, width, height, curb-weight, engine-type, num-of-cylinders, engine-size, fuel-system, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg
* adult – dane opisują pewne cechy dorosłych
  + age - wiek
  + workclass – klasa robocza
    - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
  + fnlwgt: continuous.
  + education – rodzaj wykształcenia
  + education-num: continuous.
  + marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
  + occupation - zawód
  + relationship - posiadane związki (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
  + race - rasa
  + sex - płeć
  + capital-gain / capital-loss – zysk / strata kapitałowa
  + hours-per-week – przepracowane godziny tygodniowo
  + native-country – pochodzenie
  + income – zarobki (progowo)

# Iris

Wykres korelacji:



**Decision Trees:**

Czas wykonania: 0.009008884 secs

Dokładność predykcji: 96%

**k-Nearest Neighbour:**

Czas wykonania: 0.00800705 secs

Dokładność predykcji: 96%

**Support Vector Machine:**

Czas wykonania: 0.0100081 secs

Dokładność predykcji: 98%

**Random Forest:**

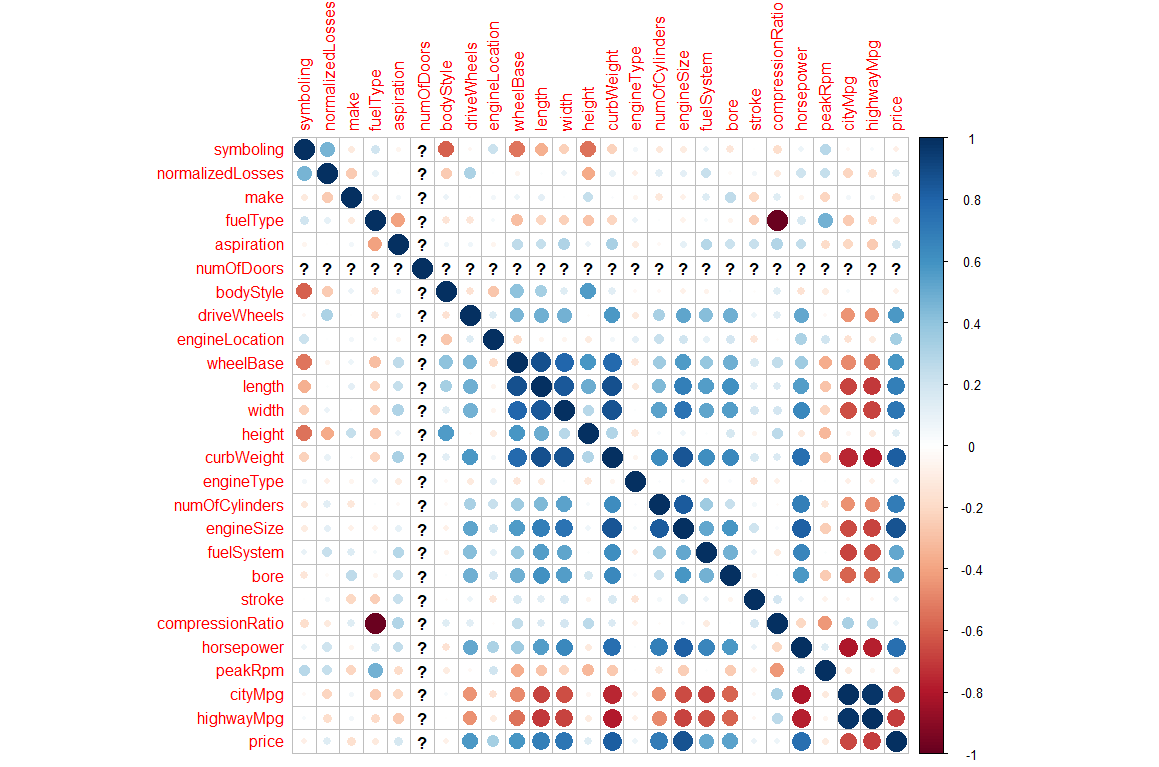
Czas wykonania: 0.03403115 secs

Dokładność predykcji: 96%

Pod względem dokładności predykcji, najlepiej wypadł algorytm Support Vector Machine, natomiast najszybszy okazał się algorytm k-Nearest Neighbours.

# Imports-85

Wykres korelacji:



**Decision Trees:**

Czas wykonania: 0.02101922 secs

Dokładność predykcji: 67%

**k-Nearest Neighbour:**

Czas wykonania:

Dokładność predykcji:

**Support Vector Machine:**

Czas wykonania:

Dokładność predykcji:

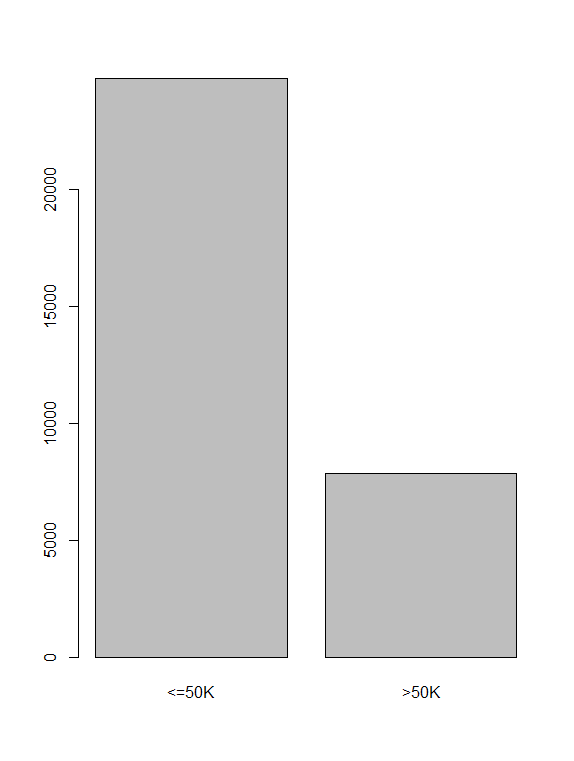
**Random Forest:**

Czas wykonania:

Dokładność predykcji:

# Adult

Zmienna income:



**Decision Trees:**

Czas wykonania: 0.4904451 secs

Dokładność predykcji: 84%

**k-Nearest Neighbour:**

Czas wykonania:

Dokładność predykcji:

**Support Vector Machine:**

Czas wykonania:

Dokładność predykcji:

**Random Forest:**

Czas wykonania: 11.35732 secs

Dokładność predykcji: 86%