# Applied Statistics and Data Analysis

*Edoardo Marangone—Gabriele Venturato*

**Abstract**

This is the abstract.

The aim of this paper is to explore and give an insight into the *rminer* package of R. The purposes are to introduce linear regression, present regression methods and then use them on a dataset to show their applications.

The rminer package can be installed using R package installation or by typing the command:

```r
install.packages("rminer")
```

The command to load the package is:

```r
library(rminer)
library(kknn)
library(ggplot2)
```

This package allows to do *data preparation*, *modeling* and *evaluation*.

## Data preparation

### 1. Loading Data

The rminer package assumes that a dataset is available as a dataframe. As example in our case we load a csv of the dataset found on Kaglle,

```r
lifeexp.df = read.csv("Life Expectancy Data.csv")
```

To see details about the dataset just loaded:

```r
str(lifeexp.df)
```

```
## 'data.frame':    2938 obs. of  22 variables:
##  $ Country                        : Factor w/ 193 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2 2 2 2 2
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

```
summary(lifeexp.df)
```

```
##                 Country          Year              Status
##  Afghanistan        :  16   Min.   :2000   Developed : 512
##  Albania            :  16   1st Qu.:2004   Developing:2426
##  Algeria            :  16   Median :2008
##  Angola             :  16   Mean   :2008
##  Antigua and Barbuda:  16   3rd Qu.:2012
##  Argentina          :  16   Max.   :2015
##  (Other)            :2842
##  Life.expectancy Adult.Mortality infant.deaths       Alcohol
##  Min.   :36.30   Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100
##  1st Qu.:63.10   1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775
##  Median :72.10   Median :144.0   Median :   3.0   Median : 3.7550
##  Mean   :69.22   Mean   :164.8   Mean   :  30.3   Mean   : 4.6029
##  3rd Qu.:75.70   3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025
##  Max.   :89.00   Max.   :723.0   Max.   :1800.0   Max.   :17.8700
##  NA's   :10      NA's   :10                       NA's   :194
##  percentage.expenditure  Hepatitis.B    Measles                 BMI
##  Min.   :    0.000      Min.   : 1.00   Min.   :     0.0   Min.   : 1.00
##  1st Qu.:    4.685      1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30
##  Median :   64.913      Median :92.00   Median :    17.0   Median :43.50
##  Mean   :  738.251      Mean   :80.94   Mean   :  2419.6   Mean   :38.32
##  3rd Qu.:  441.534      3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20
##  Max.   :19479.912      Max.   :99.00   Max.   :212183.0   Max.   :87.30
##                         NA's   :553                        NA's   :34
##  under.five.deaths     Polio       Total.expenditure   Diphtheria
##  Min.   :   0.00   Min.   : 3.00   Min.   : 0.370    Min.   : 2.00
##  1st Qu.:   0.00   1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00
##  Median :   4.00   Median :93.00   Median : 5.755    Median :93.00
##  Mean   :  42.04   Mean   :82.55   Mean   : 5.938    Mean   :82.32
##  3rd Qu.:  28.00   3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00
##  Max.   :2500.00   Max.   :99.00   Max.   :17.600    Max.   :99.00
##                    NA's   :19      NA's   :226       NA's   :19
##     HIV.AIDS           GDP              Population
##  Min.   : 0.100   Min.   :     1.68   Min.   :3.400e+01
##  1st Qu.: 0.100   1st Qu.:   463.94   1st Qu.:1.958e+05
##  Median : 0.100   Median :  1766.95   Median :1.387e+06
##  Mean   : 1.742   Mean   :  7483.16   Mean   :1.275e+07
##  3rd Qu.: 0.800   3rd Qu.:  5910.81   3rd Qu.:7.420e+06
##  Max.   :50.600   Max.   :119172.74   Max.   :1.294e+09
##                   NA's   :448         NA's   :652
##  thinness..1.19.years thinness.5.9.years Income.composition.of.resources
##  Min.   : 0.10        Min.   : 0.10      Min.   :0.0000
##  1st Qu.: 1.60        1st Qu.: 1.50      1st Qu.:0.4930
##  Median : 3.30        Median : 3.30      Median :0.6770
##  Mean   : 4.84        Mean   : 4.87      Mean   :0.6276
##  3rd Qu.: 7.20        3rd Qu.: 7.20      3rd Qu.:0.7790
##  Max.   :27.70        Max.   :28.60      Max.   :0.9480
##  NA's   :34           NA's   :34         NA's   :167
##    Schooling
##  Min.   : 0.00
##  1st Qu.:10.10
##  Median :12.30
##  Mean   :11.99
##  3rd Qu.:14.30
##  Max.   :20.70
##  NA's   :163
```

In summary we see that for some variables there are missing values, so we need to take care of them in next section.

### 2. Data Selection and transformation

This is a crucial step because it inclues operations, such as outlier detection and removal, attribute and instance selection, assuring data quality etc. In rminer package there are some useful functions, such as

- delevels
- imputation
- CasesSeries.

*Delevels:* reduces or replace factors levels;

*Imputation:* replaces missing data with values according to the other values of the dataframe

*CasesSeries:* Creates a dataframe from a time series using a sliding window. The sliding window contains a set of time lags used to pull out variable inputs from a series.

In this first part we used only the *imputation* function to deal with missing values. We tried with three different methods: 1. Deleting all the entries with missing values 2. Using imputation to substitute missing values with the mode 3. Using imputation to substitute missing values using the hotdeck method

```r
# 1st method: case deletion
lifeexp.na.del = na.omit(lifeexp.df)
summary(lifeexp.na.del)
```

```
##         Country          Year            Status     Life.expectancy
##  Afghanistan:  16   Min.   :2000   Developed : 242   Min.   :44.0  
##  Albania    :  16   1st Qu.:2005   Developing:1407   1st Qu.:64.4  
##  Armenia    :  15   Median :2008                     Median :71.7  
##  Austria    :  15   Mean   :2008                     Mean   :69.3  
##  Belarus    :  15   3rd Qu.:2011                     3rd Qu.:75.0  
##  Belgium    :  15   Max.   :2015                     Max.   :89.0  
##  (Other)    :1557                                                  
##  Adult.Mortality infant.deaths        Alcohol       percentage.expenditure
##  Min.   :  1.0   Min.   :   0.00   Min.   : 0.010   Min.   :    0.00  
##  1st Qu.: 77.0   1st Qu.:   1.00   1st Qu.: 0.810   1st Qu.:   37.44  
##  Median :148.0   Median :   3.00   Median : 3.790   Median :  145.10  
##  Mean   :168.2   Mean   :  32.55   Mean   : 4.533   Mean   :  698.97  
##  3rd Qu.:227.0   3rd Qu.:  22.00   3rd Qu.: 7.340   3rd Qu.:  509.39  
##  Max.   :723.0   Max.   :1600.00   Max.   :17.870   Max.   :18961.35  
##                                                                       
##   Hepatitis.B       Measles            BMI         under.five.deaths
##  Min.   : 2.00   Min.   :     0   Min.   : 2.00   Min.   :   0.00  
##  1st Qu.:74.00   1st Qu.:     0   1st Qu.:19.50   1st Qu.:   1.00  
##  Median :89.00   Median :    15   Median :43.70   Median :   4.00  
##  Mean   :79.22   Mean   :  2224   Mean   :38.13   Mean   :  44.22  
##  3rd Qu.:96.00   3rd Qu.:   373   3rd Qu.:55.80   3rd Qu.:  29.00  
##  Max.   :99.00   Max.   :131441   Max.   :77.10   Max.   :2100.00  
```

```
##
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.740    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:81.00   1st Qu.: 4.410    1st Qu.:82.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.840    Median :92.00   Median : 0.100
##  Mean   :83.56   Mean   : 5.956    Mean   :84.16   Mean   : 1.984
##  3rd Qu.:97.00   3rd Qu.: 7.470    3rd Qu.:97.00   3rd Qu.: 0.700
##  Max.   :99.00   Max.   :14.390    Max.   :99.00   Max.   :50.600
##
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.100
##  1st Qu.:   462.15   1st Qu.:1.919e+05   1st Qu.: 1.600
##  Median :  1592.57   Median :1.420e+06   Median : 3.000
##  Mean   :  5566.03   Mean   :1.465e+07   Mean   : 4.851
##  3rd Qu.:  4718.51   3rd Qu.:7.659e+06   3rd Qu.: 7.100
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.200
##
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.100     Min.   :0.0000                  Min.   : 4.20
##  1st Qu.: 1.700     1st Qu.:0.5090                  1st Qu.:10.30
##  Median : 3.200     Median :0.6730                  Median :12.30
##  Mean   : 4.908     Mean   :0.6316                  Mean   :12.12
##  3rd Qu.: 7.100     3rd Qu.:0.7510                  3rd Qu.:14.00
##  Max.   :28.200     Max.   :0.9360                  Max.   :20.70
##
```

```
# 2nd method: imputation by mode
lifeexp.imp.mode = lifeexp.df
for (i in 1:ncol(lifeexp.df)) {
  if ( any(is.na(lifeexp.df[,i])) ) {
    lifeexp.imp.mode = imputation("value", lifeexp.imp.mode, i, Value=which.max(table(na.omit(lifeexp.d
  }
}
summary(lifeexp.imp.mode)
```

```
##                    Country           Year            Status
##  Afghanistan        :  16   Min.   :2000   Developed : 512
##  Albania            :  16   1st Qu.:2004   Developing:2426
##  Algeria            :  16   Median :2008
##  Angola             :  16   Mean   :2008
##  Antigua and Barbuda:  16   3rd Qu.:2012
##  Argentina          :  16   Max.   :2015
##  (Other)            :2842
##  Life.expectancy  Adult.Mortality infant.deaths       Alcohol
##  Min.   : 36.30   Min.   :  1.0   Min.   :   0.0   Min.   : 0.010
##  1st Qu.: 63.20   1st Qu.: 73.0   1st Qu.:   0.0   1st Qu.: 1.000
##  Median : 72.10   Median :144.0   Median :   3.0   Median : 3.130
##  Mean   : 69.87   Mean   :164.3   Mean   :  30.3   Mean   : 4.365
##  3rd Qu.: 75.70   3rd Qu.:227.0   3rd Qu.:  22.0   3rd Qu.: 7.390
##  Max.   :259.00   Max.   :723.0   Max.   :1800.0   Max.   :17.870
##
##  percentage.expenditure  Hepatitis.B       Measles
##  Min.   :    0.000      Min.   : 1.00   Min.   :     0.0
##  1st Qu.:    4.685      1st Qu.:82.00   1st Qu.:     0.0
##  Median :   64.913      Median :87.00   Median :    17.0
##  Mean   :  738.251      Mean   :82.08   Mean   :  2419.6
##  3rd Qu.:  441.534      3rd Qu.:96.00   3rd Qu.:   360.2
##  Max.   :19479.912      Max.   :99.00   Max.   :212183.0
##
##       BMI          under.five.deaths     Polio       Total.expenditure
##  Min.   :  1.00   Min.   :   0.00   Min.   : 3.00   Min.   :  0.37
##  1st Qu.: 19.40   1st Qu.:   0.00   1st Qu.:77.00   1st Qu.:  4.37
##  Median : 43.90   Median :   4.00   Median :93.00   Median :  5.95
##  Mean   : 43.46   Mean   :  42.04   Mean   :82.49   Mean   : 27.10
##  3rd Qu.: 56.48   3rd Qu.:  28.00   3rd Qu.:97.00   3rd Qu.:  8.19
##  Max.   :482.00   Max.   :2500.00   Max.   :99.00   Max.   :281.00
##
##    Diphtheria       HIV.AIDS           GDP            Population
##  Min.   : 2.00   Min.   : 0.100   Min.   :     1.0   Min.   :2.200e+01
##  1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   190.2   1st Qu.:5.874e+03
##  Median :93.00   Median : 0.100   Median :  1172.0   Median :5.394e+05
##  Mean   :82.32   Mean   : 1.742   Mean   :  6342.2   Mean   :9.923e+06
##  3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  4779.4   3rd Qu.:4.584e+06
##  Max.   :99.00   Max.   :50.600   Max.   :119172.7   Max.   :1.294e+09
##
##  thinness..1.19.years thinness.5.9.years Income.composition.of.resources
##  Min.   : 0.100       Min.   : 0.100     Min.   :0.0000
##  1st Qu.: 1.600       1st Qu.: 1.600     1st Qu.:0.5042
##  Median : 3.400       Median : 3.400     Median :0.6895
##  Mean   : 4.899       Mean   : 4.918     Mean   :0.6487
##  3rd Qu.: 7.300       3rd Qu.: 7.300     3rd Qu.:0.7970
##  Max.   :27.700       Max.   :28.600     Max.   :1.0000
##
##    Schooling
##  Min.   :  0.00
##  1st Qu.: 10.30
##  Median : 12.50
##  Mean   : 16.93
##  3rd Qu.: 14.70
##  Max.   :101.00
##
```

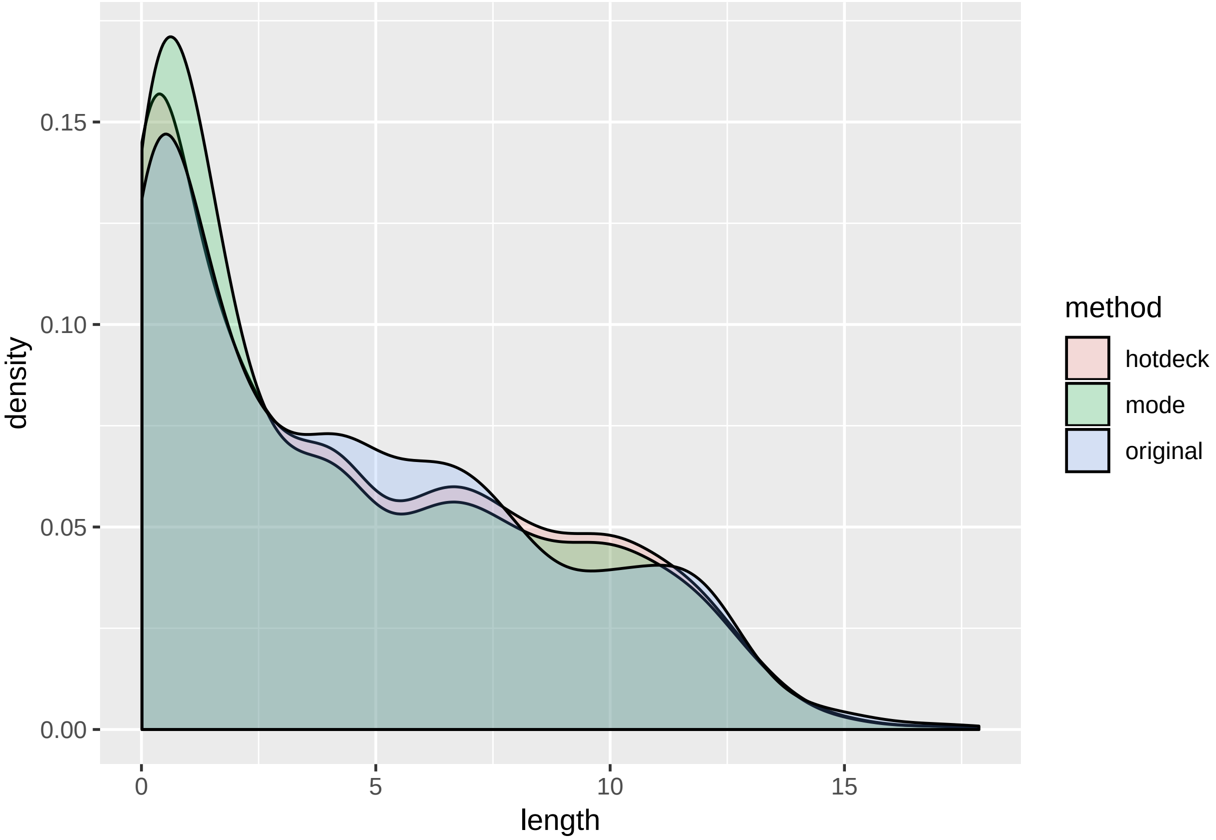
```
# 3rd mode: imputation by hotdeck
lifeexp.imp.hotdeck = lifeexp.df
for (i in 1:ncol(lifeexp.df)) {
  if ( any(is.na(lifeexp.df[,i])) ) {
    lifeexp.imp.hotdeck = imputation("hotdeck", lifeexp.imp.hotdeck, i)
  }
}
summary(lifeexp.imp.mode)
```

```
##                  Country          Year             Status
##  Afghanistan        :  16   Min.   :2000   Developed : 512
##  Albania            :  16   1st Qu.:2004   Developing:2426
##  Algeria            :  16   Median :2008
##  Angola             :  16   Mean   :2008
##  Antigua and Barbuda:  16   3rd Qu.:2012
##  Argentina          :  16   Max.   :2015
##  (Other)            :2842
##  Life.expectancy  Adult.Mortality infant.deaths       Alcohol
##  Min.   : 36.30   Min.   :  1.0   Min.   :   0.0   Min.   : 0.010
##  1st Qu.: 63.20   1st Qu.: 73.0   1st Qu.:   0.0   1st Qu.: 1.000
##  Median : 72.10   Median :144.0   Median :   3.0   Median : 3.130
##  Mean   : 69.87   Mean   :164.3   Mean   :  30.3   Mean   : 4.365
##  3rd Qu.: 75.70   3rd Qu.:227.0   3rd Qu.:  22.0   3rd Qu.: 7.390
##  Max.   :259.00   Max.   :723.0   Max.   :1800.0   Max.   :17.870
##
##  percentage.expenditure  Hepatitis.B       Measles
##  Min.   :    0.000      Min.   : 1.00   Min.   :     0.0
##  1st Qu.:    4.685      1st Qu.:82.00   1st Qu.:     0.0
##  Median :   64.913      Median :87.00   Median :    17.0
##  Mean   :  738.251      Mean   :82.08   Mean   :  2419.6
##  3rd Qu.:  441.534      3rd Qu.:96.00   3rd Qu.:   360.2
##  Max.   :19479.912      Max.   :99.00   Max.   :212183.0
##
##       BMI         under.five.deaths     Polio       Total.expenditure
##  Min.   :  1.00   Min.   :   0.00   Min.   : 3.00   Min.   :  0.37
##  1st Qu.: 19.40   1st Qu.:   0.00   1st Qu.:77.00   1st Qu.:  4.37
##  Median : 43.90   Median :   4.00   Median :93.00   Median :  5.95
##  Mean   : 43.46   Mean   :  42.04   Mean   :82.49   Mean   : 27.10
##  3rd Qu.: 56.48   3rd Qu.:  28.00   3rd Qu.:97.00   3rd Qu.:  8.19
##  Max.   :482.00   Max.   :2500.00   Max.   :99.00   Max.   :281.00
##
##    Diphtheria       HIV.AIDS           GDP             Population
##  Min.   : 2.00   Min.   : 0.100   Min.   :     1.0   Min.   :2.200e+01
##  1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   190.2   1st Qu.:5.874e+03
##  Median :93.00   Median : 0.100   Median :  1172.0   Median :5.394e+05
##  Mean   :82.32   Mean   : 1.742   Mean   :  6342.2   Mean   :9.923e+06
##  3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  4779.4   3rd Qu.:4.584e+06
##  Max.   :99.00   Max.   :50.600   Max.   :119172.7   Max.   :1.294e+09
##
##  thinness..1.19.years thinness.5.9.years Income.composition.of.resources
##  Min.   : 0.100       Min.   : 0.100     Min.   :0.0000
##  1st Qu.: 1.600       1st Qu.: 1.600     1st Qu.:0.5042
##  Median : 3.400       Median : 3.400     Median :0.6895
##  Mean   : 4.899       Mean   : 4.918     Mean   :0.6487
##  3rd Qu.: 7.300       3rd Qu.: 7.300     3rd Qu.:0.7970
##  Max.   :27.700       Max.   :28.600     Max.   :1.0000
##
##    Schooling
##  Min.   :  0.00
##  1st Qu.: 10.30
##  Median : 12.50
##  Mean   : 16.93
##  3rd Qu.: 14.70
##  Max.   :101.00
##
```

We performed also a comparison among the different techniques. Here we report the result for the *Alcohol* variable:

```
meth1=data.frame(length=lifeexp.na.del$Alcohol)
meth2=data.frame(length=lifeexp.imp.mode$Alcohol)
meth3=data.frame(length=lifeexp.imp.hotdeck$Alcohol)
meth1$method="original"
```

```
meth2$method="mode"
meth3$method="hotdeck"
all=rbind(meth1,meth2,meth3)
ggplot(all,aes(length,fill=method))+geom_density(alpha = 0.2)
```



we can see that the hotdeck method is the average solution, compared with the mode that is too much extreme, so we decided to keep the dataset with missing values substituted with the hotdeck technique.

### Modeling

The rminer package contains 15 regression methods. These methods can be used by *fit*, *predict* and *mining* functions. We focused our attention on *RandomForest* model.

1. *fit:* adjusts a selected model to a dataset
2. *predict:* given a fitted model, it computes the predictions for a new dataset
3. *mining:* trains and tests a particular fit model under several runs and a given validation method

### Holdout

First of all we trained a model using the *holdout* technique to divide the dataset in training and test sets.

Model training:

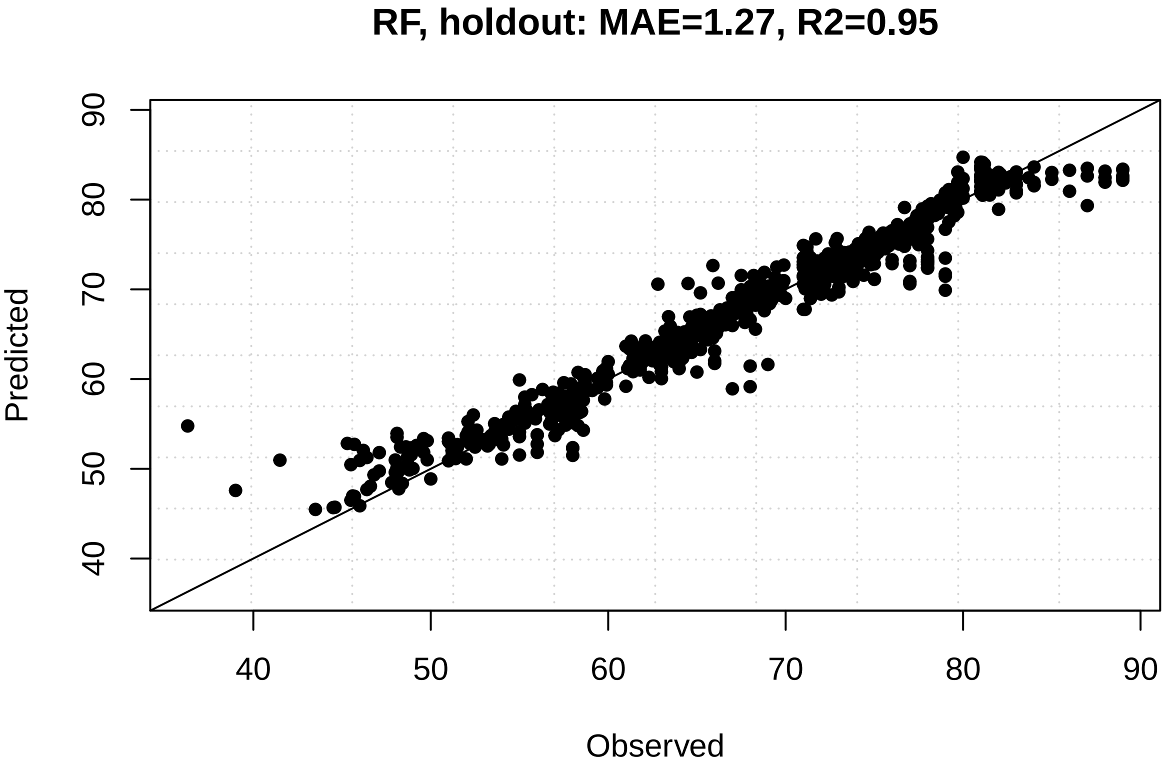
```
H = holdout(lifeexp$Life.expectancy, ratio=2/3, seed=12345)
summary(H)
```

```
##     Length Class  Mode
## tr  1958   -none- numeric
## itr    0   -none- NULL
## val    0   -none- NULL
## ts   980   -none- numeric
```

```
model1 = fit( Life.expectancy~., lifeexp[H$tr,c(inputs,dvar)], model="randomForest")
```

Model testing:

```
# get predictions on test set (new data)
pred1 = predict(model1, lifeexp[H$ts,c(inputs,dvar)])
# show scatter plot with quality of the predictions:
target1 = lifeexp[H$ts,]$Life.expectancy
e1 = mmetric(target1, pred1, metric=c("MAE","R2"))
error = paste("RF, holdout: MAE=", round(e1[1],2), ", R2=", round(e1[2],2), sep="")
mgraph(target1, pred1, graph="RSC", Grid=10, main=error)
```



**Evaluation**

The ‘rminer’ package contains evaluation metrics and graphs that can be used to assess the quality of the fitted models and to get informations from the models. In order to do that, the *mmetric* and *mgraph()* functions are needed.