

# Regressão de Biomassa

Luan Roger Santos Santana

```
## Loading required package: ggplot2
## Loading required package: lattice
##
## Attaching package: 'caret'
## The following objects are masked from 'package:Metrics':
##
##   precision, recall
```

## Preparing data

### Loading Data

```
data_raw <- read.csv("../data_sets/Material 03 - 10 - Alunos - Dados.csv")
data_raw_new_cases <- read.csv("../data_sets/Material 03 - 10 - Alunos - Dados - Novos Casos.csv")
```

### Cleaning data

```
data <- data_raw
data_new_cases <- data_raw_new_cases
print(head(data))
```

```
##   school sex age address famsize Pstatus Medu Fedu   Mjob   Fjob   reason
## 1    GP   2  18      U    GT3      A    4    4  at_home teacher  course
## 2    GP   2  17      U    GT3      T    1    1  at_home   other  course
## 3    GP   2  15      U    LE3      T    1    1  at_home   other  other
## 4    GP   2  15      U    GT3      T    4    2  health services  home
## 5    GP   2  16      U    GT3      T    3    3   other   other  home
## 6    GP   1  16      U    LE3      T    4    3 services  other reputation
##   guardian travelttime studytime failures schoolsup famsup paid activities
## 1   mother           2         2         0        yes    no   no          no
## 2   father           1         2         0        no    yes  no          no
## 3   mother           1         2         3        yes    no  yes          no
## 4   mother           1         3         0        no    yes  yes          yes
## 5   father           1         2         0        no    yes  yes          no
## 6   mother           1         2         0        no    yes  yes          yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1    yes    yes      no      no      4         3     4     1     1     3
## 2    no    yes      yes      no      5         3     3     1     1     3
## 3    yes    yes      yes      no      4         3     2     2     3     3
## 4    yes    yes      yes     yes      3         2     2     1     1     5
## 5    yes    yes      no      no      4         3     2     1     2     5
## 6    yes    yes      yes      no      5         4     2     1     2     5
```

```
## absences G1 G2 G3
## 1      6  5  6  6
## 2      4  5  5  6
## 3     10  7  8 10
## 4      2 15 14 15
## 5      4  6 10 10
## 6     10 15 15 15
```

```
print(head(data_new_cases))
```

```
## school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason
## 1 GP 2 16 R GT3 A 4 4 at_home teacher course
## 2 GP 1 17 U GT3 T 1 1 at_home other course
## 3 GP 1 18 U LE3 T 1 1 at_home other other
## guardian traveltime studytime failures schoolsup famsup paid activities
## 1 mother 2 1 0 no no no no
## 2 father 1 2 0 no no no yes
## 3 mother 1 2 3 yes yes yes yes
## nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1 yes yes no no 4 3 1 1 1 3
## 2 no yes yes yes 5 3 3 1 1 3
## 3 yes yes yes no 4 3 2 2 3 3
## absences G1 G2 G3
## 1 6 2 3 ?
## 2 4 15 15 ?
## 3 10 10 10 ?
```

## Creating data partitioning

```
set.seed(1988)
# ran <- sample(1:nrow(data), 0.8 * nrow(data))
ind <- createDataPartition(data$G3, p=0.80, list = FALSE)
training_data <- data[ind,]
test_data <- data[-ind,]
```

## Training

### Using KNN

#### Creating the model

```
tuneGrid <- expand.grid(k = c(1,3,5,7,9))
set.seed(1988)
knn <- train(G3 ~ ., data = training_data, method = "knn", tuneGrid=tuneGrid)
print(knn)
```

```
## k-Nearest Neighbors
##
## 318 samples
## 32 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 318, 318, 318, 318, 318, 318, ...
```

```
## Resampling results across tuning parameters:
##
##  k  RMSE      Rsquared  MAE
##  1  2.377665  0.7314687  1.523125
##  3  2.080581  0.7871955  1.376019
##  5  1.931045  0.8162142  1.320403
##  7  1.885793  0.8270615  1.293422
##  9  1.870958  0.8319092  1.287775
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

```
prediction.knn <- predict(knn, test_data)
library(Metrics)
rmse(test_data$G3, prediction.knn)
```

### Checking the model with training data

```
## [1] 2.442221
```

```
r2 <- function(predito, observado) {
  return(1 - (sum((predito-observado)^2) / sum((predito-mean(observado))^2)))
}
r2(prediction.knn,test_data$G3)
```

### R<sup>2</sup> function

```
## [1] 0.4002866
```

### Checking for new cases

```
prediction.knn_new_data <- predict(knn, data_new_cases)
data_new_cases$G3 <- NULL
result <- cbind(data_new_cases, G3=prediction.knn_new_data)
print(result)
```

```
##  school sex age address famsize Pstatus Medu Fedu  Mjob  Fjob reason
## 1    GP   2  16      R    GT3      A    4    4 at_home teacher course
## 2    GP   1  17      U    GT3      T    1    1 at_home  other course
## 3    GP   1  18      U    LE3      T    1    1 at_home  other  other
##  guardian traveltime studytime failures schoolsup famsup paid activities
## 1  mother          2          1          0          no    no    no          no
## 2  father          1          2          0          no    no    no          yes
## 3  mother          1          2          3          yes   yes   yes          yes
##  nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1    yes    yes      no      no      4          3    1    1    1    3
## 2     no    yes      yes     yes     5          3    3    1    1    3
## 3    yes    yes      yes     no     4          3    2    2    3    3
##  absences G1 G2      G3
## 1         6  2  3  5.666667
## 2         4 15 15 14.545455
## 3        10 10 10 10.600000
```