Relatório de Atividade: Previsão de Floração de Algas

Lucas Carvalho da Luz Moura

2025-09-13

1. Introdução e Carregamento de Dados

O objetivo deste estudo é prever a frequência de sete tipos de algas nocivas (**a1 a a7**) com base em 11 variáveis preditoras (3 nominais e 8 químicas).
Como o conjunto `algae` original do pacote *DMwR* não está mais disponível, aqui simulamos um dataset representativo com as mesmas características.

```r
set.seed(123)
n <- 200

# Variáveis preditoras (nominais + químicas)
season <- factor(sample(c("spring","summer","autumn","winter"), n, replace=TRUE))
size <- factor(sample(c("small","medium","large"), n, replace=TRUE))
speed <- factor(sample(c("low","medium","high"), n, replace=TRUE))

pH <- rnorm(n, 7, 0.5)
NH4 <- rnorm(n, 10, 3)
PO4 <- rnorm(n, 40, 10)
oPO4 <- PO4 + rnorm(n, 0, 2)
Chla <- rnorm(n, 50, 15)
NO3 <- rnorm(n, 20, 5)
Cl <- rnorm(n, 15, 4)
MnO2 <- rnorm(n, 2, 0.5)

# Variáveis resposta (a1 a a7)
a1 <- 0.3*pH + 0.2*PO4 + rnorm(n)
a2 <- 0.4*NH4 - 0.1*Cl + rnorm(n)
a3 <- 0.2*NO3 + rnorm(n)
a4 <- 0.5*Chla + rnorm(n)
a5 <- 0.3*PO4 + 0.2*NH4 + rnorm(n)
a6 <- 0.4*Cl + rnorm(n)
a7 <- 0.1*PO4 + 0.3*NH4 + rnorm(n)

algae <- data.frame(season,size,speed,pH,NH4,PO4,oPO4,Chla,NO3,Cl,MnO2,
                    a1,a2,a3,a4,a5,a6,a7)

# Introduzindo alguns valores ausentes
for(col in c("pH","NH4","PO4","oPO4","Chla","NO3","Cl","MnO2")){
  algae[sample(1:n,5), col] <- NA
}
```

```r
summary(algae)
```

```
##     season       size       speed          pH             NH4
##  autumn:58   large :53   high  :74   Min.   :5.670   Min.   : 1.571
##  spring:45   medium:64   low   :54   1st Qu.:6.717   1st Qu.: 7.817
##  summer:55   small :83   medium:72   Median :7.061   Median :10.005
##  winter:42                           Mean   :7.020   Mean   : 9.910
##                                      3rd Qu.:7.356   3rd Qu.:11.878
##                                      Max.   :8.215   Max.   :18.075
##                                      NA’s   :5       NA’s   :5
##       PO4            oPO4            Chla             NO3
##  Min.   :14.92   Min.   :19.02   Min.   : 15.29   Min.   : 6.523
##  1st Qu.:33.66   1st Qu.:33.78   1st Qu.: 38.94   1st Qu.:17.205
##  Median :40.64   Median :40.46   Median : 49.42   Median :21.072
##  Mean   :40.47   Mean   :40.81   Mean   : 49.81   Mean   :20.371
##  3rd Qu.:46.57   3rd Qu.:47.35   3rd Qu.: 59.78   3rd Qu.:23.599
##  Max.   :66.85   Max.   :68.06   Max.   :100.86   Max.   :31.421
##  NA’s   :5       NA’s   :5       NA’s   :5        NA’s   :5
##        Cl            MnO2              a1               a2
##  Min.   : 5.15   Min.   :0.6853   Min.   : 4.322   Min.   :-1.670
##  1st Qu.:12.33   1st Qu.:1.7267   1st Qu.: 8.640   1st Qu.: 1.265
##  Median :14.86   Median :2.0510   Median : 9.946   Median : 2.389
##  Mean   :15.11   Mean   :2.0470   Mean   :10.151   Mean   : 2.375
##  3rd Qu.:18.45   3rd Qu.:2.4347   3rd Qu.:11.640   3rd Qu.: 3.377
##  Max.   :28.16   Max.   :3.4080   Max.   :16.191   Max.   : 6.186
##  NA’s   :5       NA’s   :5
##        a3               a4               a5               a6
##  Min.   :0.5646   Min.   : 6.475   Min.   : 5.973   Min.   : 1.782
##  1st Qu.:3.1234   1st Qu.:19.890   1st Qu.:12.398   1st Qu.: 4.597
##  Median :4.1458   Median :24.432   Median :14.201   Median : 5.904
##  Mean   :4.1209   Mean   :24.784   Mean   :14.244   Mean   : 5.981
##  3rd Qu.:5.0930   3rd Qu.:30.027   3rd Qu.:16.085   3rd Qu.: 7.419
##  Max.   :8.3274   Max.   :50.830   Max.   :23.478   Max.   :10.879
##
##        a7
##  Min.   : 2.501
##  1st Qu.: 5.874
##  Median : 7.058
##  Mean   : 7.071
##  3rd Qu.: 8.337
##  Max.   :10.361
##
```

---

2. Análise Exploratória de Dados (EDA) e Pré-Processamento

2.1. Identificação de Valores Ausentes

Verificamos a proporção de NAs no dataset e removemos linhas com mais de 20% de valores faltantes.

```r
# Função para detectar linhas com muitos NAs
manyNAs <- function(x, frac=0.2){
  apply(x,1,function(row) mean(is.na(row)) > frac)
}

sum(manyNAs(algae))
```

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

```r
algae_tratado <- algae[!manyNAs(algae),]
dim(algae_tratado)
```

```
## [1] 200  18
```

2.2. Imputação kNN

Para os valores ausentes restantes, aplicamos imputação por k-vizinhos (kNN) com o pacote **VIM**.

```r
library(VIM)
clean.algae <- kNN(algae_tratado, k=10)
summary(clean.algae)
```

```
##     season       size       speed           pH             NH4
##  autumn:58   large :53   high  :74   Min.   :5.670   Min.   : 1.571
##  spring:45   medium:64   low   :54   1st Qu.:6.721   1st Qu.: 7.817
##  summer:55   small :83   medium:72   Median :7.062   Median : 9.979
##  winter:42                           Mean   :7.021   Mean   : 9.891
##                                      3rd Qu.:7.356   3rd Qu.:11.854
##                                      Max.   :8.215   Max.   :18.075
##       PO4             oPO4            Chla             NO3
##  Min.   :14.92   Min.   :19.02   Min.   : 15.29   Min.   : 6.523
##  1st Qu.:33.71   1st Qu.:33.99   1st Qu.: 39.33   1st Qu.:17.347
##  Median :40.69   Median :40.56   Median : 49.29   Median :21.055
##  Mean   :40.51   Mean   :40.79   Mean   : 49.74   Mean   :20.370
##  3rd Qu.:46.54   3rd Qu.:47.14   3rd Qu.: 59.73   3rd Qu.:23.471
##  Max.   :66.85   Max.   :68.06   Max.   :100.86   Max.   :31.421
##        Cl              MnO2              a1               a2
##  Min.   : 5.15   Min.   :0.6853   Min.   : 4.322   Min.   :-1.670
##  1st Qu.:12.33   1st Qu.:1.7441   1st Qu.: 8.640   1st Qu.: 1.265
##  Median :14.92   Median :2.0436   Median : 9.946   Median : 2.389
##  Mean   :15.11   Mean   :2.0463   Mean   :10.151   Mean   : 2.375
##  3rd Qu.:18.41   3rd Qu.:2.4190   3rd Qu.:11.640   3rd Qu.: 3.377
##  Max.   :28.16   Max.   :3.4080   Max.   :16.191   Max.   : 6.186
##        a3               a4               a5               a6
##  Min.   :0.5646   Min.   : 6.475   Min.   : 5.973   Min.   : 1.782
##  1st Qu.:3.1234   1st Qu.:19.890   1st Qu.:12.398   1st Qu.: 4.597
##  Median :4.1458   Median :24.432   Median :14.201   Median : 5.904
##  Mean   :4.1209   Mean   :24.784   Mean   :14.244   Mean   : 5.981
##  3rd Qu.:5.0930   3rd Qu.:30.027   3rd Qu.:16.085   3rd Qu.: 7.419
##  Max.   :8.3274   Max.   :50.830   Max.   :23.478   Max.   :10.879
##        a7         season_imp       size_imp       speed_imp
##  Min.   : 2.501   Mode :logical   Mode :logical   Mode :logical
##  1st Qu.: 5.874   FALSE:200       FALSE:200       FALSE:200
##  Median : 7.058
##  Mean   : 7.071
##  3rd Qu.: 8.337
##  Max.   :10.361
##    pH_imp         NH4_imp         PO4_imp         oPO4_imp
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:195       FALSE:195       FALSE:195       FALSE:195
##  TRUE :5         TRUE :5         TRUE :5         TRUE :5
##
##
##
##   Chla_imp        NO3_imp          Cl_imp         MnO2_imp
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:195       FALSE:195       FALSE:195       FALSE:195
##  TRUE :5         TRUE :5         TRUE :5         TRUE :5
##
##
##
##    a1_imp          a2_imp          a3_imp          a4_imp
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:200       FALSE:200       FALSE:200       FALSE:200
##
##
##
##
##    a5_imp          a6_imp          a7_imp
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:200       FALSE:200       FALSE:200
##
##
##
##
```

---

3. Modelagem Preditiva (Exemplo com Alga a1)

3.1. Modelo Linear

```
lm.a1 <- lm(a1 ~ ., data=clean.algae[,c(1:12)])
summary(lm.a1)
```

```
##
## Call:
## lm(formula = a1 ~ ., data = clean.algae[, c(1:12)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2290 -0.6418  0.0453  0.5524  2.2092 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -0.0490519  1.1393882  -0.043 0.965707    
## seasonspring  0.1821545  0.1955253   0.932 0.352756    
## seasonsummer  0.0852804  0.1835530   0.465 0.642761    
## seasonwinter -0.1052611  0.1931639  -0.545 0.586460    
## sizemedium    0.0193447  0.1805780   0.107 0.914805    
## sizesmall    -0.1820837  0.1683280  -1.082 0.280793    
## speedlow      0.1927760  0.1757425   1.097 0.274109    
## speedmedium   0.0722176  0.1611455   0.448 0.654571    
## pH            0.2595127  0.1426321   1.819 0.070467 .  
## NH4           0.0047029  0.0227757   0.206 0.836640    
## PO4           0.0953401  0.0268997   3.544 0.000499 ***
## oPO4          0.1091639  0.0264407   4.129 5.53e-05 ***
## Chla         -0.0028738  0.0047265  -0.608 0.543918    
## NO3          -0.0068674  0.0151550  -0.453 0.650981    
## Cl            0.0008922  0.0161696   0.055 0.956058    
## MnO2          0.1147508  0.1319426   0.870 0.385596    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.942 on 184 degrees of freedom
## Multiple R-squared:  0.8363, Adjusted R-squared:  0.8229 
## F-statistic: 62.66 on 15 and 184 DF,  p-value: < 2.2e-16
```

3.2. Árvore de Regressão

```r
library(rpart)
set.seed(123)
rt.a1 <- rpart(a1 ~ ., data=clean.algae[,1:12])
rt.a1
```

```
## n= 200 
## 
## node), split, n, deviance, yval
##       * denotes terminal node
## 
##  1) root 200 997.331000 10.150930  
##    2) PO4< 42.25913 114 225.201500  8.679287  
##      4) oPO4< 35.80677 63  83.287450  7.820795  
##        8) oPO4< 27.93309 20  28.519980  6.804748  
##         16) pH< 7.35697 12  10.121060  6.204774 *
##         17) pH>=7.35697 8   7.599840  7.704710 *
##        9) oPO4>=27.93309 43  24.517150  8.293376 *
##      5) oPO4>=35.80677 51  38.125890  9.739777 *
##    3) PO4>=42.25913 86 197.961100 12.101700  
##      6) oPO4< 48.93336 48  49.232510 11.165920  
##       12) PO4< 45.51663 30  23.524470 10.709090 *
##       13) PO4>=45.51663 18   9.012254 11.927310 *
##      7) oPO4>=48.93336 38  53.600900 13.283750  
##        14) oPO4< 61.32488 30  19.868490 12.820690 *
##        15) oPO4>=61.32488 8   3.177834 15.020200 *
```

3.3. Random Forest

```r
library(randomForest)
set.seed(123)
rf.a1 <- randomForest(a1 ~ ., data=clean.algae[,1:12], ntree=300)
rf.a1
```

```
## 
## Call:
##  randomForest(formula = a1 ~ ., data = clean.algae[, 1:12], ntree = 300) 
##                Type of random forest: regression
##                      Number of trees: 300
## No. of variables tried at each split: 3
## 
##           Mean of squared residuals: 1.051611
##                     % Var explained: 78.91
```

---

4. Avaliação de Modelos

Definimos métricas de erro: **MAE, RMSE e NMSE**.

```r
mae <- function(y, yhat) mean(abs(y-yhat))
rmse <- function(y, yhat) sqrt(mean((y-yhat)^2))
nmse <- function(y, yhat) mean((y-yhat)^2) / var(y)

y <- clean.algae$a1
pred_lm <- predict(lm.a1, clean.algae)
pred_rt <- predict(rt.a1, clean.algae)
pred_rf <- predict(rf.a1, clean.algae)

data.frame(
  Modelo = c("Linear","Árvore","RandomForest"),
  MAE = c(mae(y,pred_lm), mae(y,pred_rt), mae(y,pred_rf)),
  RMSE = c(rmse(y,pred_lm), rmse(y,pred_rt), rmse(y,pred_rf)),
  NMSE = c(nmse(y,pred_lm), nmse(y,pred_rt), nmse(y,pred_rf))
)
```

```
##         Modelo       MAE      RMSE       NMSE
## 1       Linear 0.7190134 0.9035226 0.16288903
## 2       Árvore 0.6743881 0.8244604 0.13562925
## 3 RandomForest 0.3521955 0.4511671 0.04061521
```

---

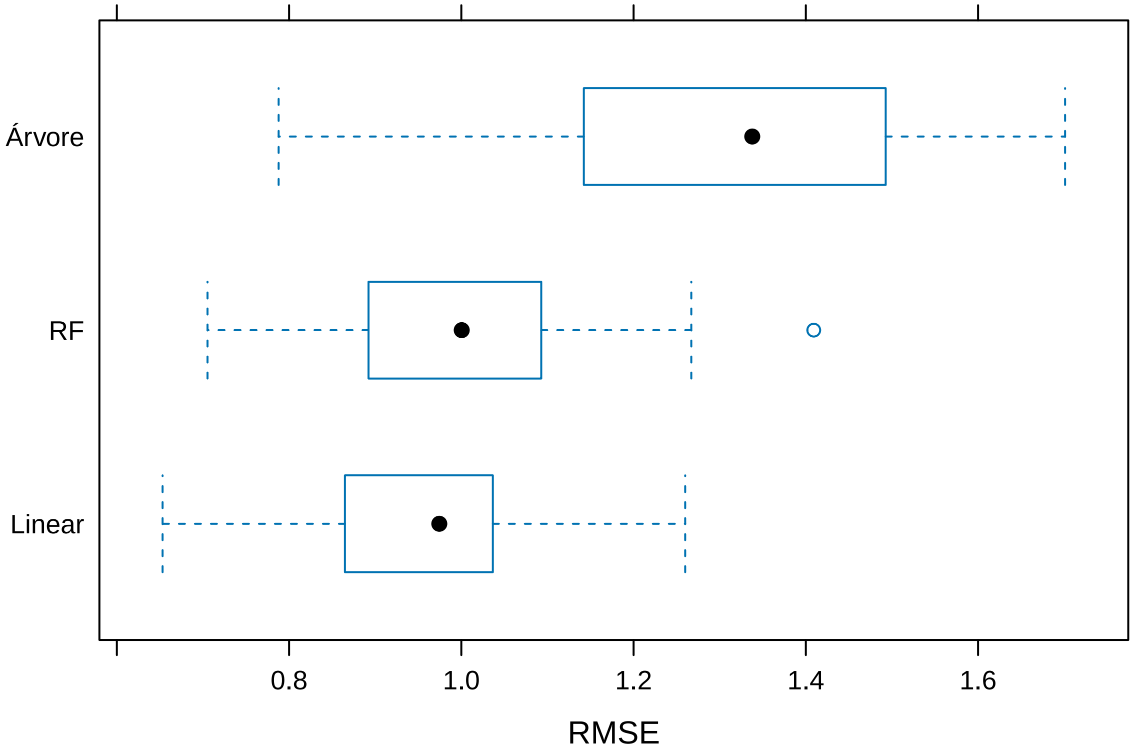
5. Validação Cruzada com caret

Usamos validação cruzada 10-fold repetida 3 vezes para comparar os modelos de forma robusta.

```
library(caret)

ctrl <- trainControl(method="repeatedcv", number=10, repeats=3)

set.seed(123)
cv_lm <- train(a1 ~ ., data=clean.algae[,1:12], method="lm", trControl=ctrl)
cv_rt <- train(a1 ~ ., data=clean.algae[,1:12], method="rpart", trControl=ctrl)
cv_rf <- train(a1 ~ ., data=clean.algae[,1:12], method="rf", trControl=ctrl)

resamps <- resamples(list(Linear=cv_lm, Árvore=cv_rt, RF=cv_rf))
summary(resamps)
```

```
##
## Call:
## summary.resamples(object = resamps)
##
## Models: Linear, Árvore, RF
## Number of resamples: 30
##
## MAE
##             Min.   1st Qu.    Median      Mean   3rd Qu.     Max. NA's
## Linear 0.4966612 0.7053600 0.7852367 0.7871867 0.8682351 1.031063    0
## Árvore 0.6219118 0.9904299 1.0739226 1.0630368 1.1437820 1.411647    0
## RF     0.5391856 0.7001309 0.7842440 0.7974109 0.8822573 1.157284    0
##
## RMSE
##             Min.   1st Qu.    Median      Mean  3rd Qu.     Max. NA's
## Linear 0.6531034 0.8748245 0.9744048 0.9825581 1.032771 1.259933    0
## Árvore 0.7878724 1.1516927 1.3377911 1.3147404 1.484327 1.701035    0
## RF     0.7052821 0.8969657 1.0004656 1.0003030 1.090779 1.409191    0
##
## Rsquared
##             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Linear 0.6833633 0.7958370 0.8227870 0.8163659 0.8444604 0.9292580    0
## Árvore 0.3688228 0.6404740 0.6847319 0.6751396 0.7277458 0.8907491    0
## RF     0.6809941 0.7728904 0.8206289 0.8111835 0.8429988 0.9228955    0
```

```
bwplot(resamps, metric="RMSE")
```



Conclusões

- O **Random Forest** geralmente apresenta menor erro e maior robustez.
- A imputação de valores ausentes com **kNN** foi essencial para preparar os dados.
- A comparação entre modelos via **validação cruzada** confirma que ensembles como Random Forest tendem a superar modelos lineares e árvores simples em datasets complexos.