Forest Fires

Andrei Petrisor, Antonio Radu, Lorenzo Medici, Andrea Rusconi

2023-12-01

Contents

1 Dataset

Questo dataset è pubblico e a disposizione per la ricerca. I dettagli sul dataset possono essere trovati in Cortez e Morais (2007). Il dataset è composto dalle seguenti variabili:

1. Coordinata spaziale dell'asse X all'interno della mappa del parco Montesinho: da 1 a 9
2. Y coordinata spaziale dell'asse y all'interno della mappa del parco Montesinho: da 2 a 9
3. mese: mese dell'anno: da "gen" a "dic"
4. giorno della settimana: da "lunedì" a "domenica"
5. Indice FFMC dal sistema FWI: da 18,7 a 96,20
6. Indice DMC dal sistema FWI: da 1,1 a 291,3
7. Indice DC dal sistema FWI: da 7,9 a 860,6
8. Indice ISI del sistema FWI: da 0,0 a 56,10
9. temperatura temporanea in gradi Celsius: da 2,2 a 33,30
10. Umidità relativa RH in %: da 15,0 a 100
11. velocità del vento in km/h: da 0,40 a 9,40
12. pioggia in mm/m2: da 0,0 a 6,4
13. area della superficie bruciata della foresta (in ettari): da 0,00 a 1090,84.

1.1 Obiettivo

In questo dataset siamo interessati a modellare l'area bruciata della foresta come funzione delle altre variabili.

Cortez P. e Morais A. "Un approccio di data mining per prevedere gli incendi boschivi utilizzando dati meteorologici." In J. Neves, MF Santos e J. Machado Eds., "Nuove tendenze nell'intelligenza artificiale", Atti della 13a EPIA 2007 Conferenza portoghese sull'intelligenza artificiale, dicembre, Guimaraes, Por- tugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. 18-0-9. Disponibile a: http://www3.dsi.uminho.pt/pcortez/fires.pdf

Il dataset è composto dalle seguenti rilevazioni:

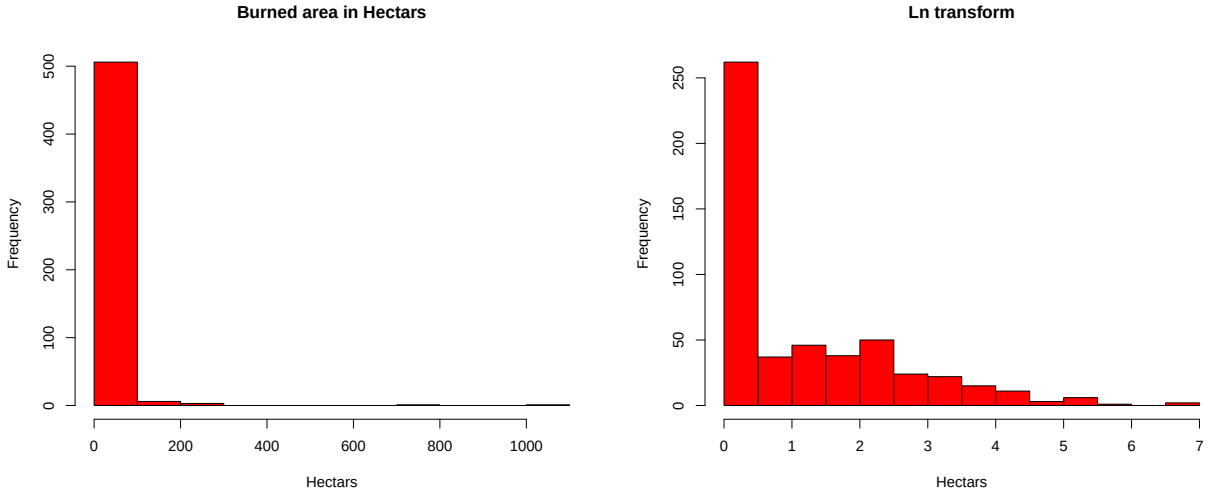
```
##        X               Y          month               day
##  Min.   :1.000   Min.   :2.0   Length:517         Length:517
##  1st Qu.:3.000   1st Qu.:4.0   Class :character   Class :character
##  Median :4.000   Median :4.0   Mode  :character   Mode  :character
##  Mean   :4.669   Mean   :4.3
##  3rd Qu.:7.000   3rd Qu.:5.0
##  Max.   :9.000   Max.   :9.0
##       FFMC            DMC              DC             ISI
##  Min.   :18.70   Min.   :  1.1   Min.   :  7.9   Min.   : 0.000
##  1st Qu.:90.20   1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500
##  Median :91.60   Median :108.3   Median :664.2   Median : 8.400
##  Mean   :90.64   Mean   :110.9   Mean   :547.9   Mean   : 9.022
##  3rd Qu.:92.90   3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800
##  Max.   :96.20   Max.   :291.3   Max.   :860.6   Max.   :56.100
##       temp             RH              wind            rain
##  Min.   : 2.20   Min.   : 15.00   Min.   :0.400   Min.   :0.00000
##  1st Qu.:15.50   1st Qu.: 33.00   1st Qu.:2.700   1st Qu.:0.00000
##  Median :19.30   Median : 42.00   Median :4.000   Median :0.00000
##  Mean   :18.89   Mean   : 44.29   Mean   :4.018   Mean   :0.02166
##  3rd Qu.:22.80   3rd Qu.: 53.00   3rd Qu.:4.900   3rd Qu.:0.00000
##  Max.   :33.30   Max.   :100.00   Max.   :9.400   Max.   :6.40000
##       area
```

```
##  Min.   :   0.00
##  1st Qu.:   0.00
##  Median :   0.52
##  Mean   :  12.85
##  3rd Qu.:   6.57
##  Max.   :1090.84
```

```
##     month              day                FFMC            DMC
##  Length:517         Length:517         Min.   :18.70   Min.   :  1.1
##  Class :character   Class :character   1st Qu.:90.20   1st Qu.: 68.6
##  Mode  :character   Mode  :character   Median :91.60   Median :108.3
##                                        Mean   :90.64   Mean   :110.9
##                                        3rd Qu.:92.90   3rd Qu.:142.4
##                                        Max.   :96.20   Max.   :291.3
##        DC              ISI              temp             RH
##  Min.   :  7.9   Min.   : 0.000   Min.   : 2.20   Min.   : 15.00
##  1st Qu.:437.7   1st Qu.: 6.500   1st Qu.:15.50   1st Qu.: 33.00
##  Median :664.2   Median : 8.400   Median :19.30   Median : 42.00
##  Mean   :547.9   Mean   : 9.022   Mean   :18.89   Mean   : 44.29
##  3rd Qu.:713.9   3rd Qu.:10.800   3rd Qu.:22.80   3rd Qu.: 53.00
##  Max.   :860.6   Max.   :56.100   Max.   :33.30   Max.   :100.00
##       wind            rain              area
##  Min.   :0.400   Min.   :0.00000   Min.   :   0.00
##  1st Qu.:2.700   1st Qu.:0.00000   1st Qu.:   0.00
##  Median :4.000   Median :0.00000   Median :   0.52
##  Mean   :4.018   Mean   :0.02166   Mean   :  12.85
##  3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.:   6.57
##  Max.   :9.400   Max.   :6.40000   Max.   :1090.84
```

```
##       Days
## Months mon tue wed thu fri sat sun
##    jan   0   0   0   0   0   1   1
##    feb   3   2   1   1   5   4   4
##    mar  12   5   4   5  11  10   7
##    apr   1   0   1   2   1   1   3
##    may   0   0   0   0   1   1   0
##    jun   3   0   3   2   3   2   4
##    jul   4   6   3   3   3   8   5
##    aug  15  28  25  26  21  29  40
##    sep  28  19  14  21  38  25  27
##    oct   4   2   2   0   1   3   3
##    nov   0   1   0   0   0   0   0
##    dic   0   0   0   0   0   0   0
```

1.2 Histogrammi area bruciata



Come possiamo vedere dall'istogramma(sin) i dati sono assimetrici verso lo zero, possiamo percio valutare i dati facendo la trasformata grazie al logaritmo, cosifacendo otteniamo un grafico più preciso.

```
summary(forest$area)
```

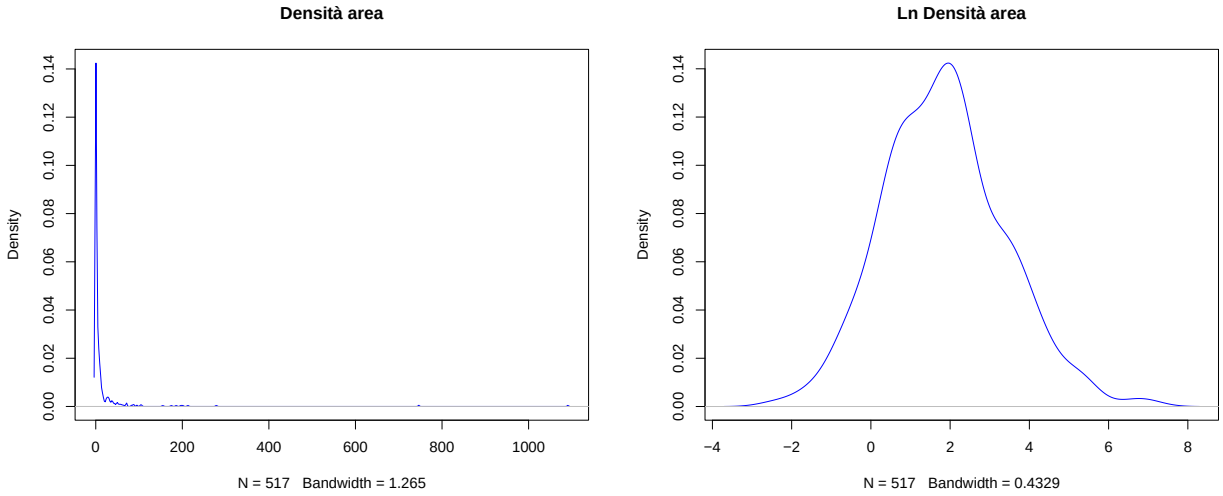
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.52   12.85    6.57 1090.84
```

Facendo il summary otteniamo i valori di: min, max, media, mediana, possiamo considerare i nostri dati molto vicini allo zero.

```
## [1] 0.4777563
```

Si evidenzia che il dataset ha il 48% dei valori che valgono 0.

1.3 Grafici densità area bruciata



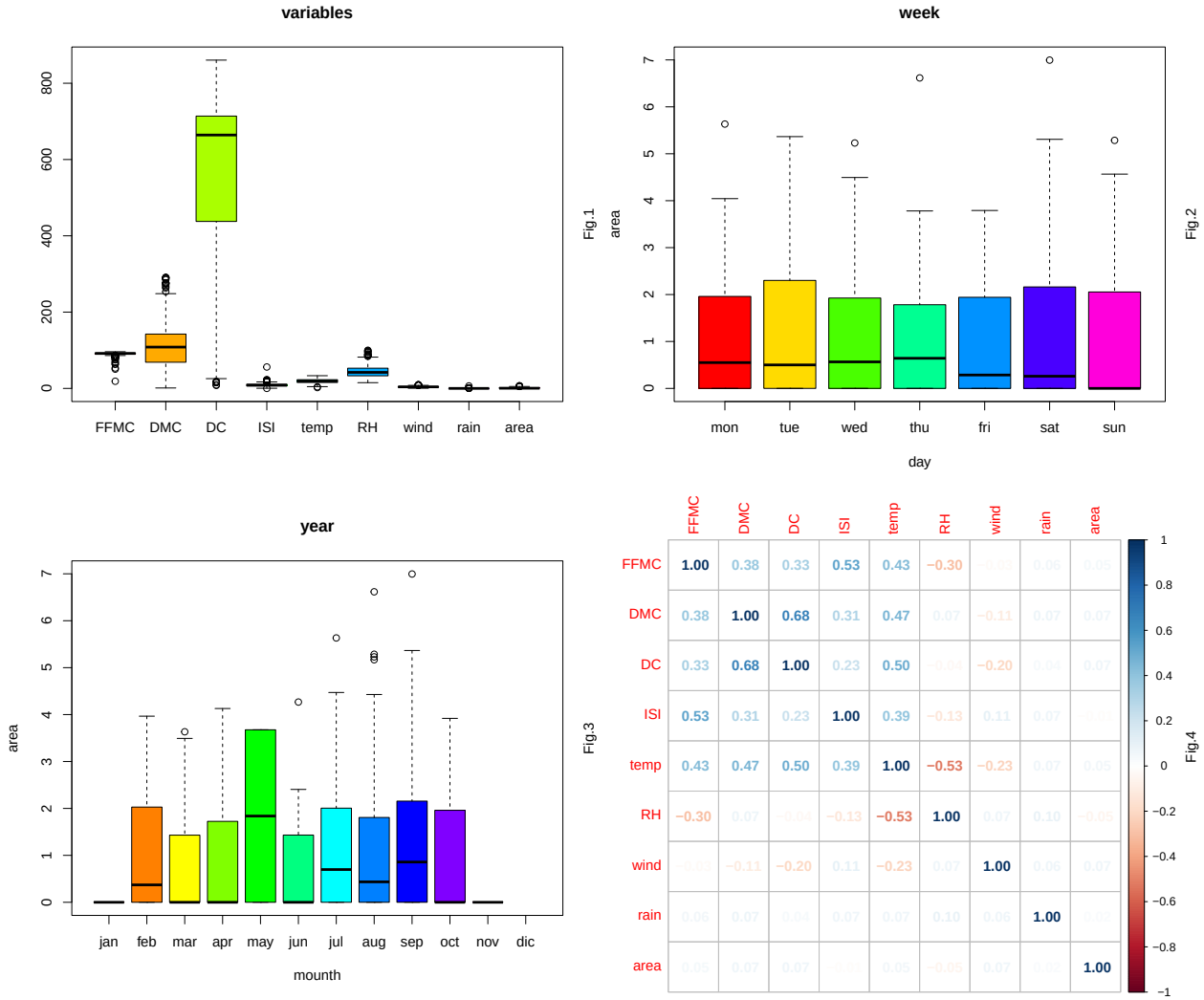
Dobbiamo tenere in considerazione del numero di zeri presenti per considerare un andamento normale, visto

che log(0)= -inf

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.4187  1.1110  2.0242  6.9956
```

come possiamo vedere la mediana e la media sono diminuite.

```
## Warning: il pacchetto 'corrplot' è stato creato con R versione 4.3.2
```



-Come possiamo vedere da questo correlation plot(fig.4) le variabili che sono correlate positivamente di più sono: DC con DMC, ISI e FFMC, temp e DC. Le variabili correlate negativamente di più sono, invece: RH e temp.

-Dal box plot delle variabili(fig.1) possiamo determinare che le mediane di quasi tutte le variabili sono più o meno simili quindi ci sono poche differenze,inoltre possiamo notare come in tutte le variabili ci siano degli outliers sia al minimo che al massimo.La variabile DMC,invece,presenta dei baffi più lunghi il che implica che tale variabile ha valori più incoerenti rispetto alle altre, la mediane lontana da tutte le altre e spostata molto verso il terzo quartile.

-Dal box plot dei giorni della settimana(fig.2) possiamo vedere come le mediane sono simili tra di loro come anche i baffi, questo dimostra come i giorni della settimana hanno tutti valori coerenti tra di loro.

-Dal box plot dei mesi dell'anno(fig.3) notiamo come il mese di dicembre e quello di maggio siano quelli più incoerenti rispetto agli altri mesi. Il mese di agosto presenta molti più outliers rispetto agli altri mesi.

2 Testing

Per la fase di testing prendiamo in considerazione solamente il mese di agosto, per poter spiegare come varia l'area bruciata in relazione alle altre variabili. Dividiamo la fase di Testing in: train e test, con il train al 70% e il test al 30%. Questo è il nostro nuovo dataset:

```
forest_subset <- forest[which(forest$month=="aug"),]
forest_subset <- forest_subset[,-c(1)]#elimino colonna month dal dataset
summary(forest_subset)
```

```
##   day          FFMC            DMC              DC             ISI
##  fri:21   Min.   :63.50   Min.   : 47.9   Min.   :100.7   Min.   : 0.800
##  mon:15   1st Qu.:91.40   1st Qu.:111.2   1st Qu.:604.3   1st Qu.: 7.775
##  sat:29   Median :92.10   Median :145.4   Median :647.1   Median :10.600
##  sun:40   Mean   :92.34   Mean   :153.7   Mean   :641.1   Mean   :11.072
##  thu:26   3rd Qu.:94.53   3rd Qu.:181.3   3rd Qu.:685.6   3rd Qu.:14.100
##  tue:28   Max.   :96.20   Max.   :273.8   Max.   :819.1   Max.   :22.700
##  wed:25
##       temp             RH             wind            rain
##  Min.   : 5.10   Min.   :19.00   Min.   :0.400   Min.   :0.0000
##  1st Qu.:18.90   1st Qu.:33.75   1st Qu.:2.700   1st Qu.:0.0000
##  Median :21.25   Median :42.00   Median :4.000   Median :0.0000
##  Mean   :21.63   Mean   :45.49   Mean   :4.086   Mean   :0.0587
##  3rd Qu.:24.23   3rd Qu.:56.00   3rd Qu.:4.900   3rd Qu.:0.0000
##  Max.   :33.30   Max.   :96.00   Max.   :8.900   Max.   :6.4000
##
##       area
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.435
##  Mean   :1.045
##  3rd Qu.:1.796
##  Max.   :6.616
##
```

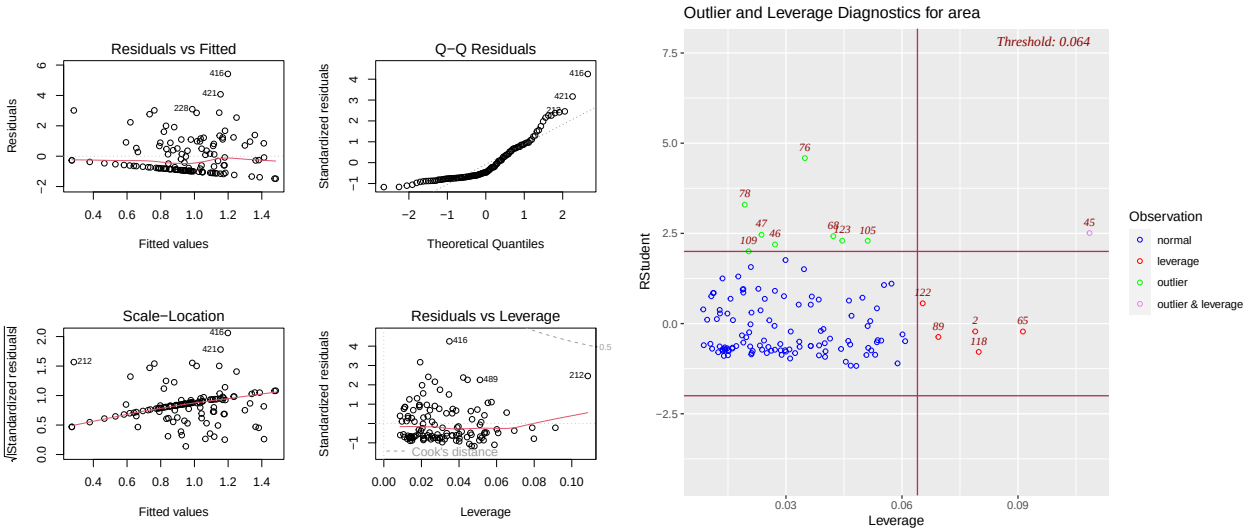
```
##
## Call:
## lm(formula = area ~ wind + temp + DMC, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4829 -0.9296 -0.5967  0.7071  5.4180
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.031317   0.669679  -0.047   0.9628
## wind         0.038797   0.074404   0.521   0.6030
## temp         0.046240   0.023812   1.942   0.0545 .
## DMC         -0.001043   0.002325  -0.449   0.6546
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.298 on 121 degrees of freedom
## Multiple R-squared:  0.03301,    Adjusted R-squared:  0.009034
## F-statistic: 1.377 on 3 and 121 DF,  p-value: 0.2532
```

```
## Warning: il pacchetto 'olsrr' è stato creato con R versione 4.3.2
```

```
##
## Call:
## lm(formula = area ~ wind + temp + DMC, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4829 -0.9296 -0.5967  0.7071  5.4180
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.031317   0.669679  -0.047   0.9628
## wind         0.038797   0.074404   0.521   0.6030
## temp         0.046240   0.023812   1.942   0.0545 .
## DMC         -0.001043   0.002325  -0.449   0.6546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.298 on 121 degrees of freedom
## Multiple R-squared:  0.03301,    Adjusted R-squared:  0.009034
## F-statistic: 1.377 on 3 and 121 DF,  p-value: 0.2532
```

```
## Analysis of Variance Table
##
## Response: area
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## wind        1   0.570  0.5703  0.3384 0.56185
## temp        1   6.052  6.0523  3.5908 0.06049 .
## DMC         1   0.339  0.3391  0.2012 0.65457
## Residuals 121 203.944  1.6855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



2.1 Confronto tra modelli:

```
# Create vector with values
a = c(AIC(model3), BIC(model3), AIC(model2), BIC(model2), AIC(model1), BIC(model1))

# Akaike Information Criterion (AIC) estimates the in-sample prediction error and indicates the relativ

# Create vector with nforest
b = c("AIC lm 3", "BIC lm 3", "AIC lm 2", "BIC lm 2", "AIC lm 1", "BIC lm 1")

# Link the values with nforest
names(a) = b
print(a)
```

```
## AIC lm 3 BIC lm 3 AIC lm 2 BIC lm 2 AIC lm 1 BIC lm 1
## 441.0094 486.2624 425.9263 440.0679 425.9263 440.0679
```

2.2 Previsioni con Model2:

```
#use lasso regression model to predict response value
new = test
previsioni_mod3 = predict(model3, newdata = new)#previsione puntuale

#find SST and SSE
sst <- sum((test$area - mean(test$area))^2)
sse <- sum((previsioni_mod3 - test$area)^2)

# Root Mean Squared Error: è una misura dell'errore che compiamo
sqrt(mean((test$area - previsioni_mod3)^2))#--
```

```
## [1] 1.514435
```

```r
# intervallo di previsione
predict(model3,new,interval="predict", level=0.95)#intervallo di previsione
```

```
##           fit        lwr      upr
## 6   0.9487703 -1.8230481 3.720589
## 24  1.0600006 -1.7178802 3.837881
## 25  1.4849119 -1.3663048 4.336129
## 54  0.7354892 -2.0066080 3.477586
## 74  0.5349060 -2.2664078 3.336220
## 80  0.4222109 -2.5344634 3.378885
## 82  0.6702891 -2.0526964 3.393275
## 93  0.8367600 -1.9511148 3.624635
## 100 0.6461023 -2.0736790 3.365884
## 102 0.6277356 -2.1559404 3.411412
## 108 0.7882901 -1.9296009 3.506181
## 136 1.2436126 -1.5945723 4.081798
## 143 0.8323316 -2.0268367 3.691500
## 146 1.3812468 -1.3402921 4.102786
## 159 0.6513526 -2.0854284 3.388134
## 176 1.1900731 -1.5513044 3.931451
## 180 0.6277356 -2.1559404 3.411412
## 196 0.7957120 -1.9880612 3.579485
## 207 1.3450427 -1.5383433 4.228429
## 230 0.9973048 -1.7827211 3.777331
## 236 0.6489473 -2.0716798 3.369574
## 243 0.7759577 -1.9402971 3.492213
## 245 1.0692478 -1.6845882 3.823084
## 246 1.2573000 -1.4807247 3.995325
## 248 1.4104337 -1.4773823 4.298250
## 250 0.9317782 -1.8227223 3.686279
## 251 1.0773405 -1.6355602 3.790241
## 256 1.4851202 -1.3256793 4.295920
## 259 1.5890806 -1.2425934 4.420755
## 260 1.5281589 -1.2552005 4.311518
## 262 0.3283275 -2.5387750 3.195430
## 263 0.7258043 -2.0445016 3.496110
## 264 0.3388053 -2.4419450 3.119556
## 271 1.2781794 -1.4542669 4.010626
## 272 1.1035701 -1.6078366 3.814977
## 374 0.9451373 -1.7985944 3.688869
## 377 0.3305879 -2.4352983 3.096474
## 378 1.0304474 -1.7154354 3.776330
## 385 0.8487732 -1.9465224 3.644069
## 389 0.4627402 -2.3312492 3.256729
## 403 0.6338729 -2.1386541 3.406400
## 414 1.2644800 -1.4851860 4.014146
## 420 0.9392173 -1.7683516 3.646786
## 426 1.0245692 -1.7162768 3.765415
## 428 0.6222619 -2.1557669 3.400291
## 439 1.3304474 -1.4118361 4.072731
## 442 0.3322193 -2.4478130 3.112252
## 452 0.5487137 -2.3311134 3.428541
## 455 0.6466401 -2.0909886 3.384269
```

```
## 458 0.6218600 -2.1479476 3.391668
## 460 0.6647884 -2.2126596 3.542236
## 485 1.7685465 -0.9957238 4.532817
## 491 1.4813376 -1.2566969 4.219372
## 499 1.7018071 -1.0897619 4.493376
## 503 1.1812365 -1.5758594 3.938332
## 505 1.4713636 -1.3113538 4.254081
## 506 1.6113128 -1.1795002 4.402126
## 509 0.9336057 -1.8338074 3.701019
## 512 1.6687804 -1.1940250 4.531586
```

2.3 LASSO regression:

```
#penalizza le covariate,avra una parte classica piu un errore
# install.packages("glmnet") # se non è già stato installato
library(glmnet)
```

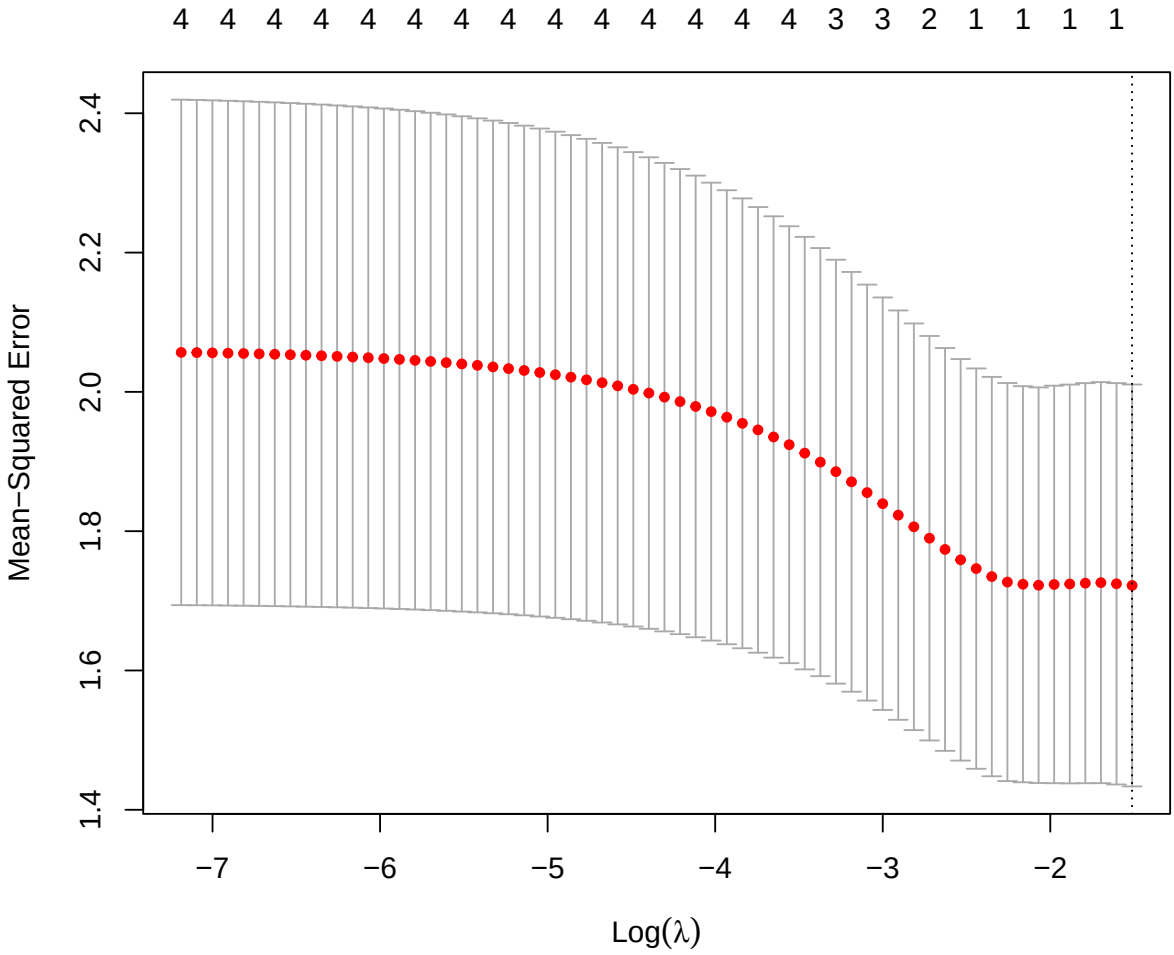
```
## Warning: il pacchetto 'glmnet' è stato creato con R versione 4.3.2
```

```
## Warning: il pacchetto 'Matrix' è stato creato con R versione 4.3.2
```

```
#define response variable
y <- train$area

#define matrix of predictor variables (uso solo poche variabili ma potete farlo con tutte da togliere p
x <- data.matrix(train[, c("wind","rain","temp","FFMC")])

#perform k-fold cross-validation to find optimal lambda value, la cross-validation è un ottimo modo per
cv_model <- cv.glmnet(x, y, alpha = 1)

#find optimal lambda value that minimizes test MSE
best_lambda <- cv_model$lambda.min
best_lambda#miglior lambda per penalizzare
```

```
## [1] 0.2205342
```

```
#produce plot of test MSE by lambda value
plot(cv_model)
```




```
# Fittiamo il modello con il best lambda (penalizzazione)

best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)
coef(best_model)
```

```
## 5 x 1 sparse Matrix of class "dgCMatrix"
##                    s0
## (Intercept) 0.9622985
## wind        0.0000000
## rain        .
## temp        .
## FFMC        .
```

2.4 Previsioni con Lasso:

```
#use lasso regression model to predict response value
new = data.matrix(test[,c("wind","rain","temp","FFMC")])
```

```r
previsioni = predict(best_model, s = best_lambda, newx = new)

# Root Mean Squared Error (RMSE): è una misura dell'errore che compiamo
sqrt(mean((test$area - previsioni)^2))#errore minimo
```

```
## [1] 1.479789
```

Ora potete confrontare modelli diversi con lasso e lm classici, vedete cosa vi dice BIC/AIC e RMSE per decidere quale è il modello ottimale.