รายงาน

Classification Model

With R

จัดทำโดย

นาย พัสกร โตวตระกูล 590610644

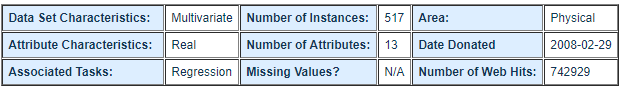
นาย รัชชานนท์ สมทา 590610679

นางสาว อมลวรรณ ลาภอินทรี 590610679

รายงานนี้เป็นส่วนหนึ่งของวิชา 261494 R for datascience

2/2561

Classification with R model

**Forest Fires Data Set**

**Source:**

Paulo Cortez, pcortez '@' dsi.uminho.pt, Department of Information Systems, University of Minho, Portugal.

Aníbal Morais, araimorais '@' gmail.com, Department of Information Systems, University of Minho, Portugal.

Attribute Information:

For more information, read [Cortez and Morais, 2007].

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9

2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9

3. month - month of the year: 'jan' to 'dec'

4. day - day of the week: 'mon' to 'sun'

5. FFMC - FFMC index from the FWI system: 18.7 to 96.20

6. DMC - DMC index from the FWI system: 1.1 to 291.3

7. DC - DC index from the FWI system: 7.9 to 860.6

8. ISI - ISI index from the FWI system: 0.0 to 56.10

9. temp - temperature in Celsius degrees: 2.2 to 33.30

10. RH - relative humidity in %: 15.0 to 100

11. wind - wind speed in km/h: 0.40 to 9.40

12. rain - outside rain in mm/m2 : 0.0 to 6.4

13. area - the burned area of the forest (in ha): 0.00 to 1090.84

(this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

Link : <https://archive.ics.uci.edu/ml/datasets/forest+fires>

Problem statement

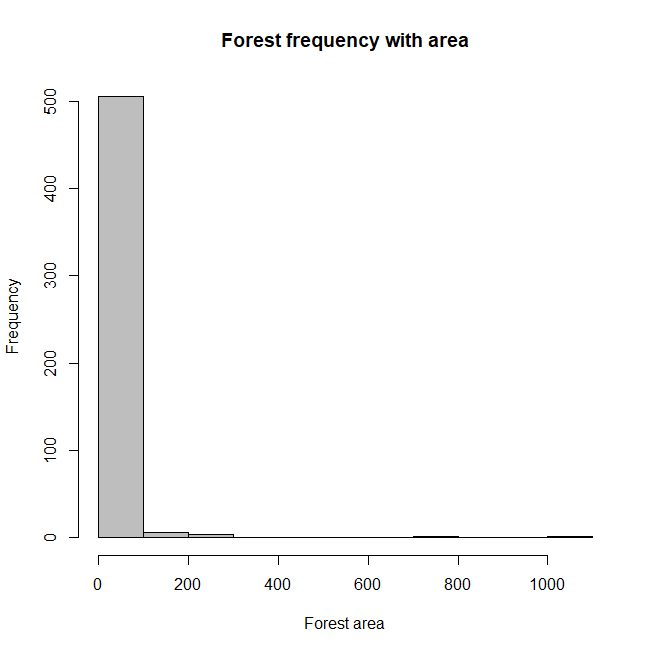
  จุดมุ่งหมายคือการทำนายพื้นที่ที่ถูกไฟไหม้ของไฟป่า ในภูมิภาคตะวันออกเฉียงเหนือของโปรตุเกส

Used Model

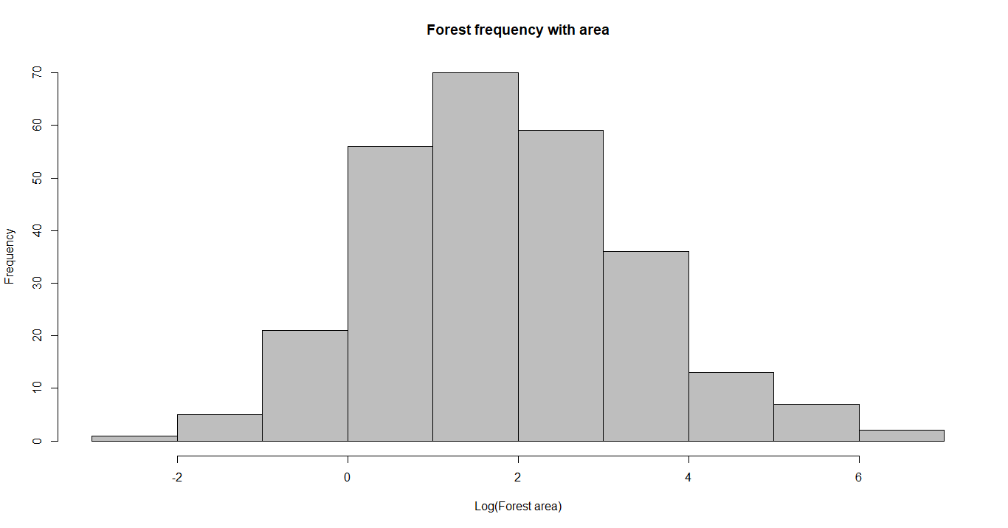
Support vector machine

Summary

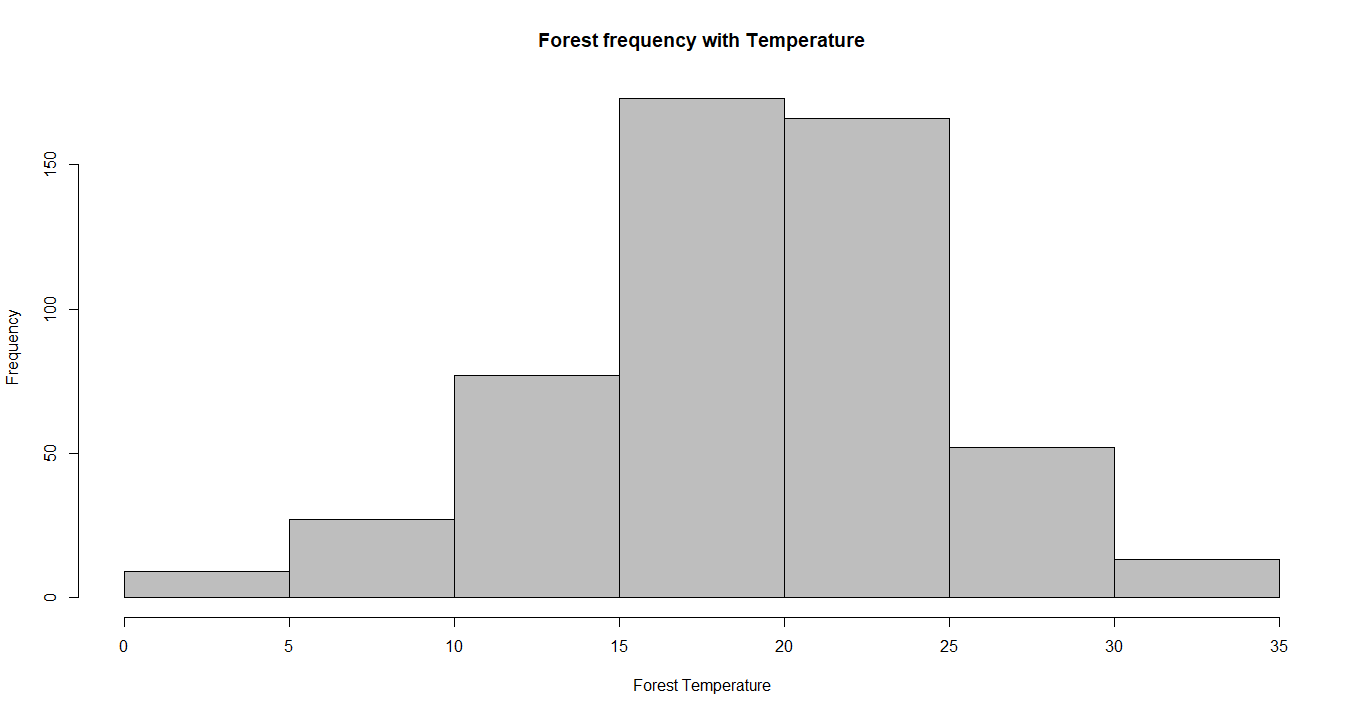
Forest fire frequency with area [no log]



Forest fire frequency with area [log]



Forest fire with temperature



#True => observed&&forest fires are triggered

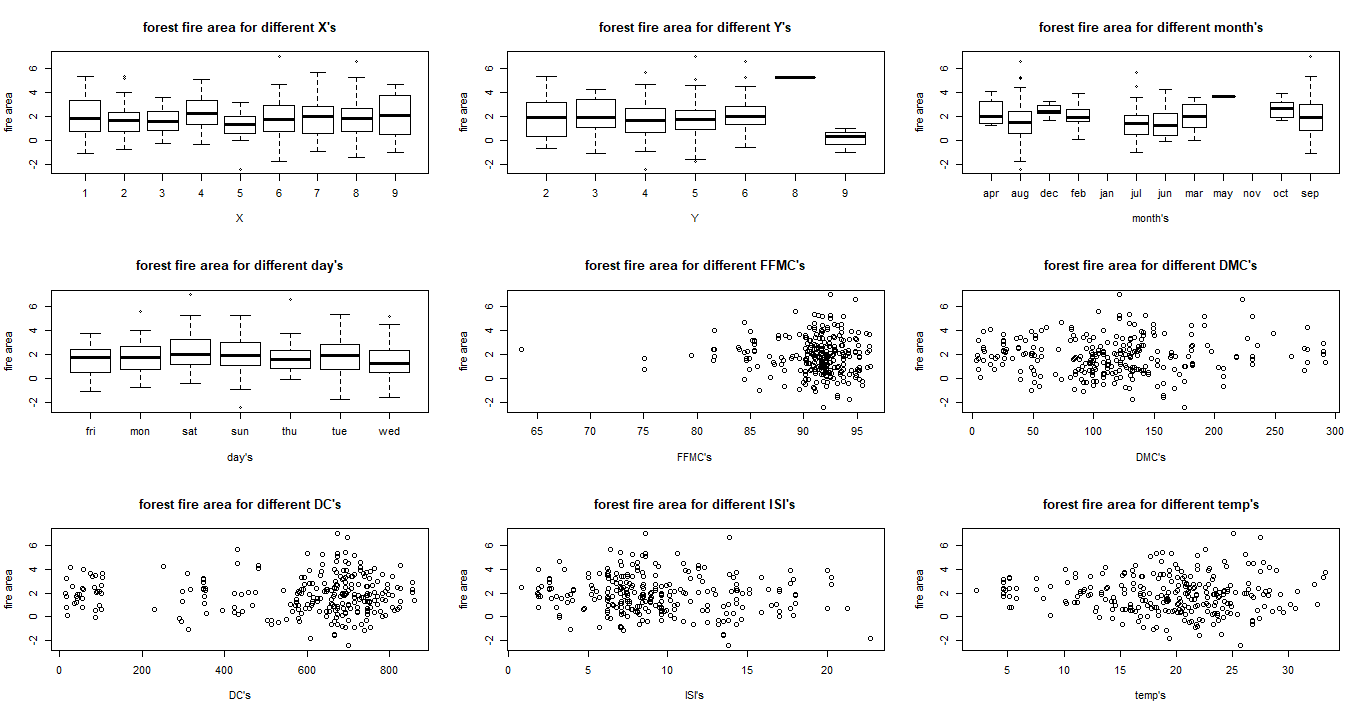
#False => observed&&forest fires aren't triggered

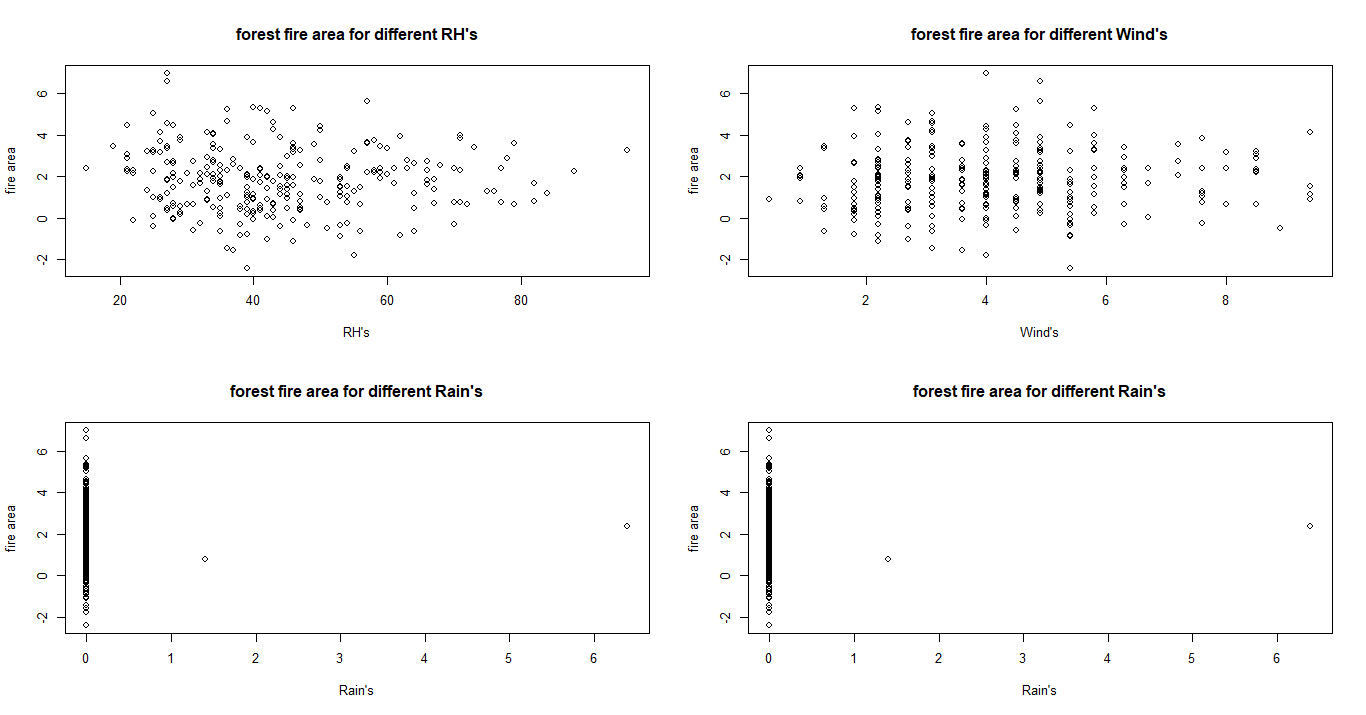
round(table(forest\_area == 0) / (n\_row), 3)

output:

FALSE TRUE

0.522 0.478





เตรียมข้อมูลเพื่อวิเคราะห์แยกเป็นฤดูๆไป

#Seasons

forest\_fire$season <- rep("spring", n\_row)

for (i in 1:n\_row) {

if (forest\_fire$month[i] %in% c("feb", "jan", "dec"))

forest\_fire$season[i] <- "winter"

if (forest\_fire$month[i] %in% c("oct", "nov", "sep"))

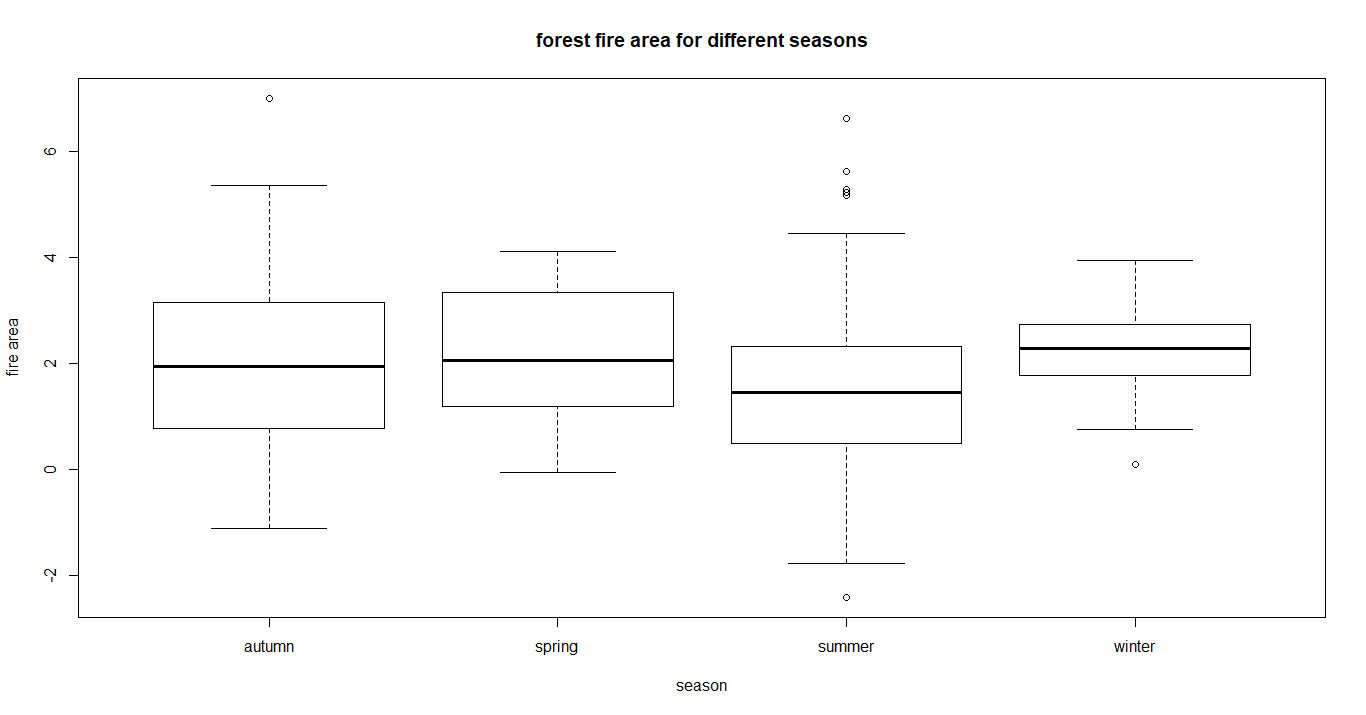
forest\_fire$season[i] <- "autumn"

if (forest\_fire$month[i] %in% c("aug", "jul", "jun"))

forest\_fire$season[i] <- "summer"

}

forest\_fire$season <- as.factor(forest\_fire$season)

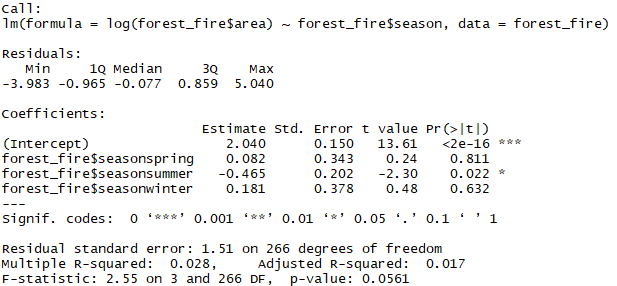


ผลสรุปจาก boxplot แสดงให้เห็นว่า ในฤดูร้อนมีค่า mean ของ forest fires ที่น้อยกว่าในอีกสามฤดู

**ใช้ Fitting Linear Models**

*reg\_season* = lm(log(forest\_fire$area) ~ forest\_fire$season, *data* = forest\_fire)

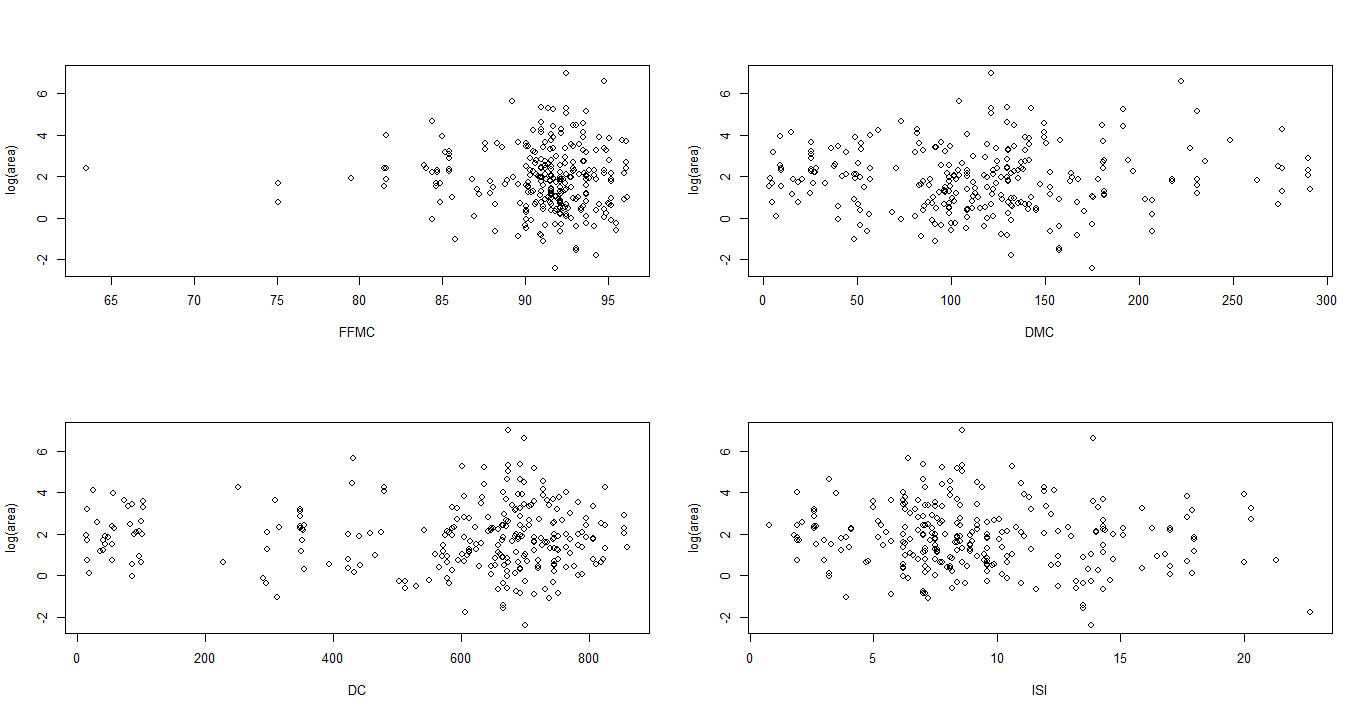
summary(reg\_season)



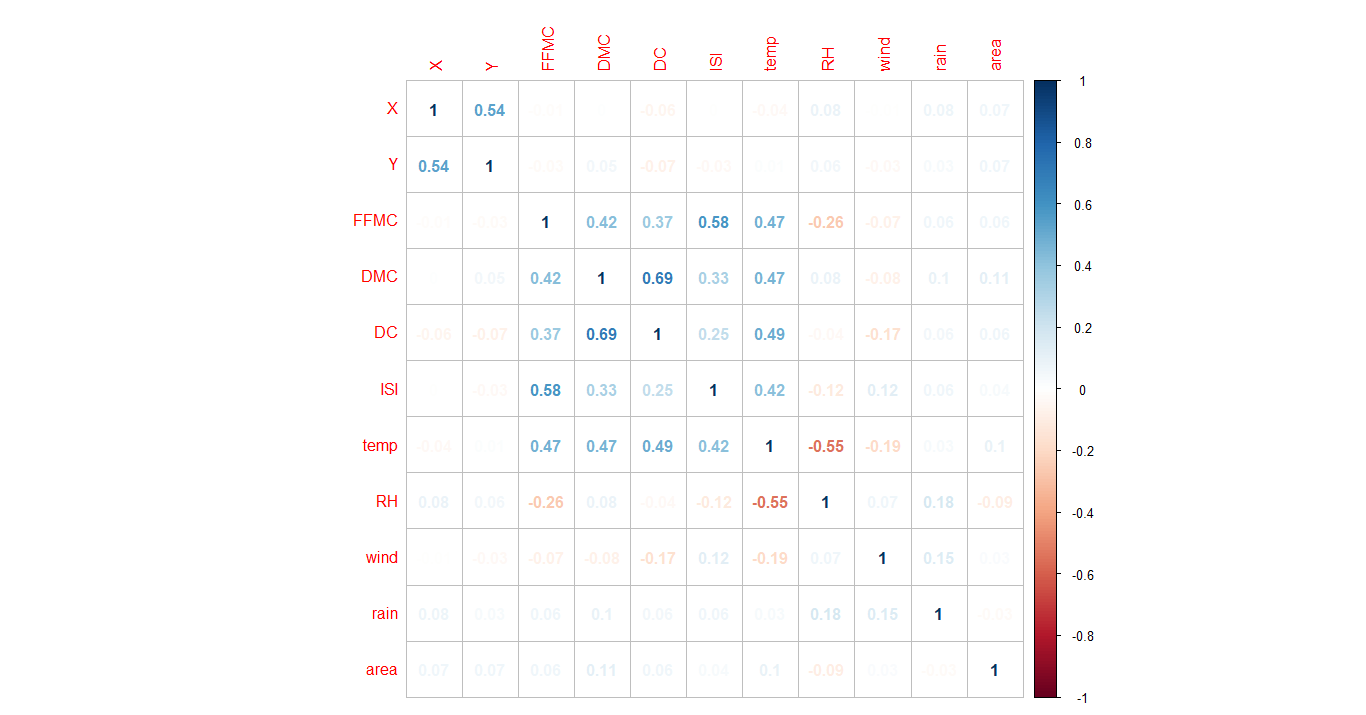
par(*mfrow* = c(2, 2))

plot(log(area) ~ FFMC + DMC + DC + ISI + temp + RH + wind + rain + season,

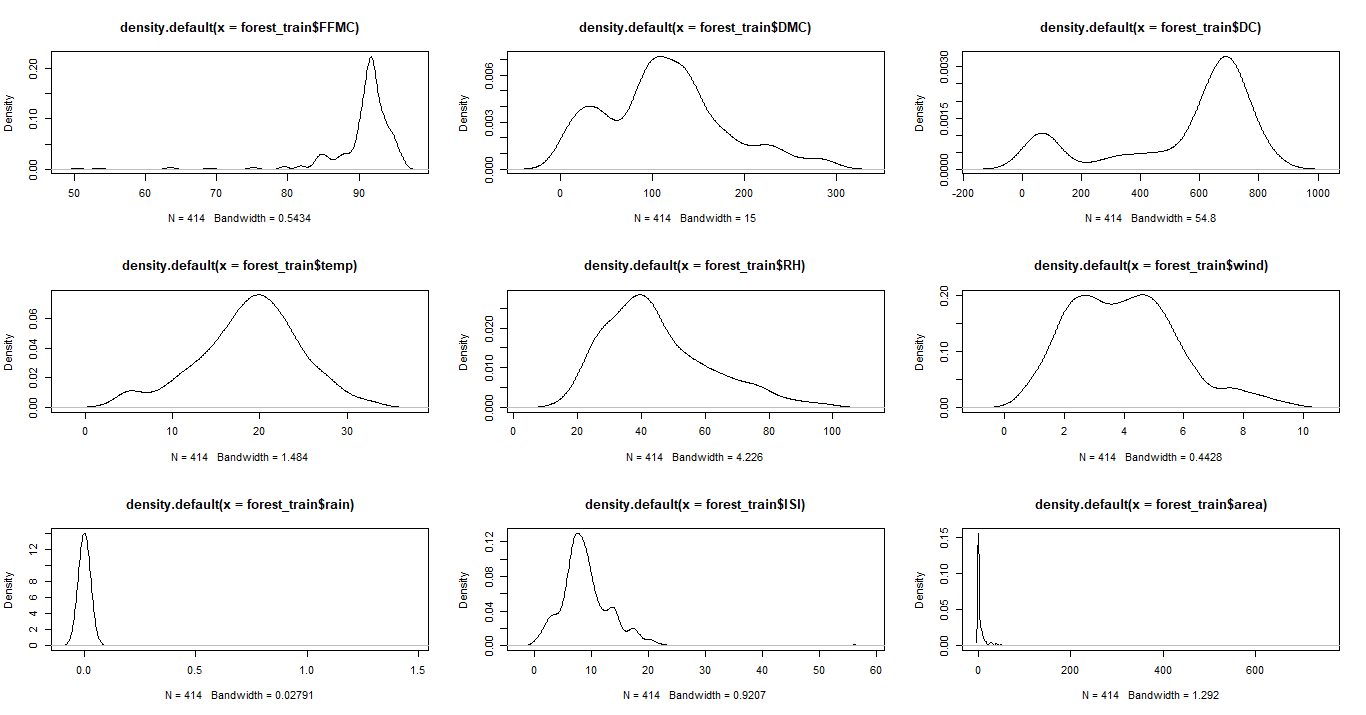
*data* = forest\_fire)



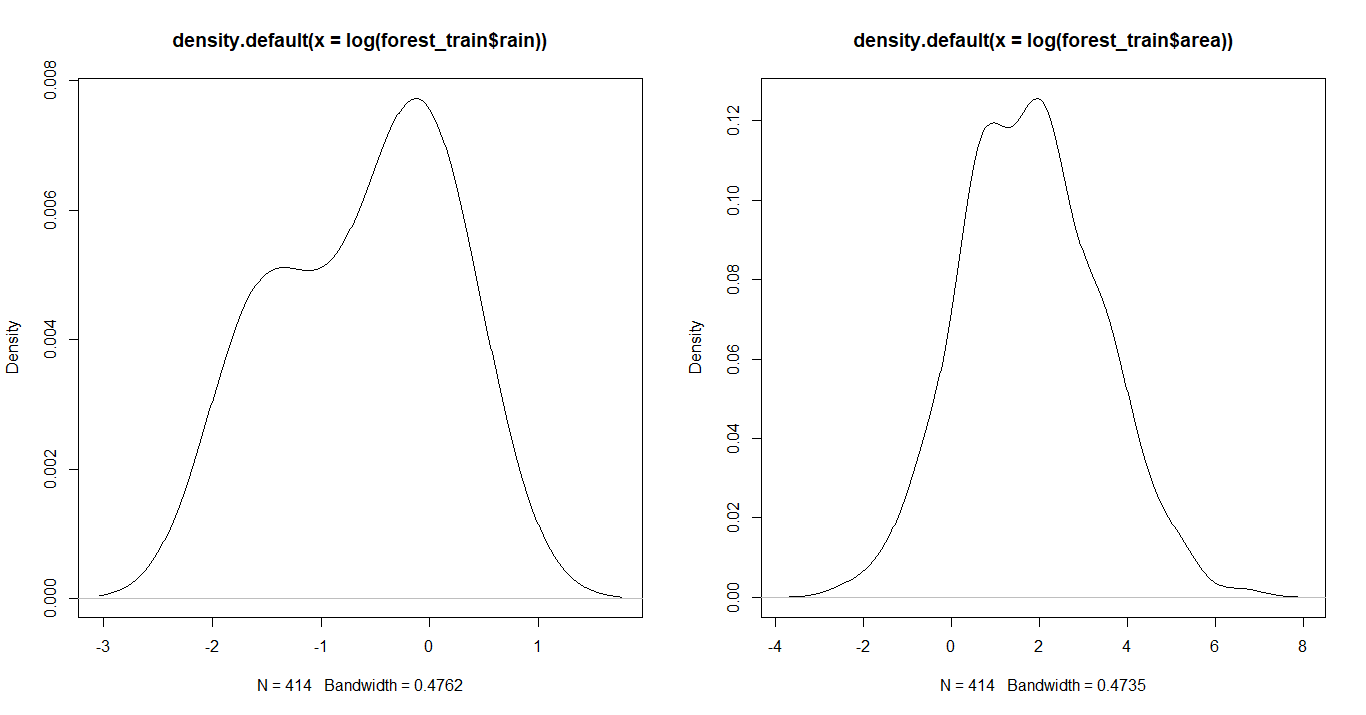
correlation plot correlation between variables



density curve for other variables



density curve for other variables [log]



SVM prediction model

Prediction SVM

Normalize Data

#nornalize data

normalise <- function(*x*) {

return((x - min(x)) / (max(x) - min(x)))

}

mydata$temp <- normalise(mydata$temp)

mydata$rain <- normalise(mydata$rain)

mydata$RH <- normalise(mydata$RH)

mydata$wind <- normalise(mydata$wind)

mydata$size <- NULL

mydata$size <- factor(ifelse(mydata$area < 5, 1, 0),

*labels* = c("small", "large"))

train <-

sample(*x* = nrow(mydata),

*size* = 0.2 \* nrow(mydata),

*replace* = FALSE)

include library:

library(kernlab)

m.poly <- ksvm(

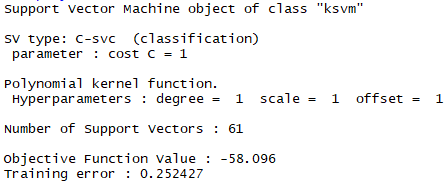
*size* ~ *temp* + *RH* + *wind* + rain,

*data* = mydata[train,],

*kernel* = "polydot",

*C* = 1

)



Training error = 25.2427%

m.rad <- ksvm(

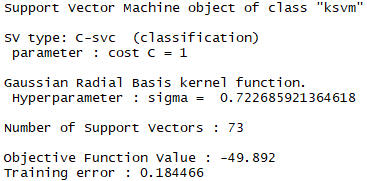
*size* ~ *temp* + *RH* + *wind* + rain,

*data* = mydata[train,],

*kernel* = "rbfdot",

*C* = 1

)



Training error = 18.4466%

m.tan <- ksvm(

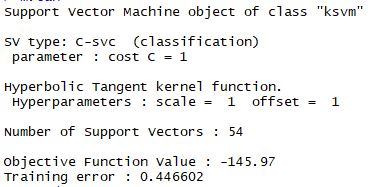
*size* ~ *temp* + *RH* + *wind* + rain,

*data* = mydata[train,],

*kernel* = "tanhdot",

*C* = 1

)



Training error = 44.6602%

pred <-predict(m.rad, *newdata* = mydata[-train,], *type* = "response")

library(e1071)

library(caret) # include library to use confusionMatrix

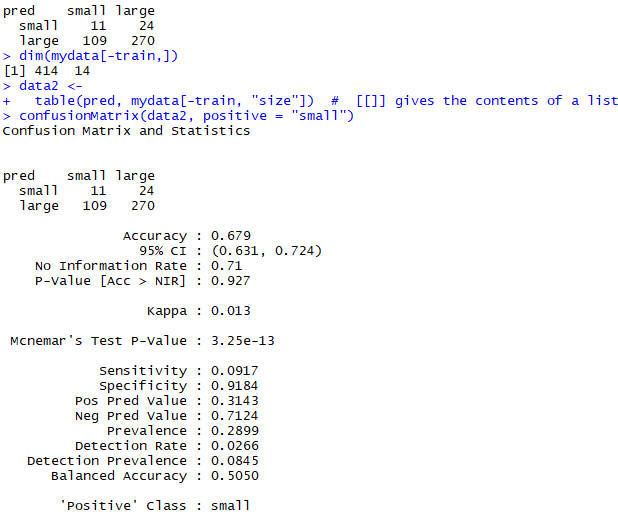
table(pred, mydata[-train, "size"])

dim(mydata[-train, ])

data2 <-

table(pred, mydata[-train, "size"]) # [[]] gives the contents of a list

confusionMatrix(data2, *positive* = "small")



Feature selection

include library

library(randomForest)

area.rf <-

randomForest(

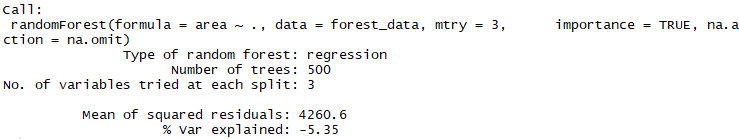
*area* ~ .,

*data* = forest\_data,

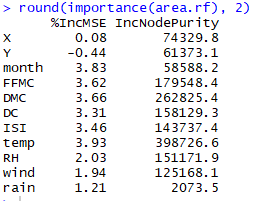
*mtry* = 3,

*importance* = TRUE,

*na.action* = na.omit

 )

round(importance(area.rf), 2)

first 5 importance attributes: temp, DMC, DC, FFMC, RH

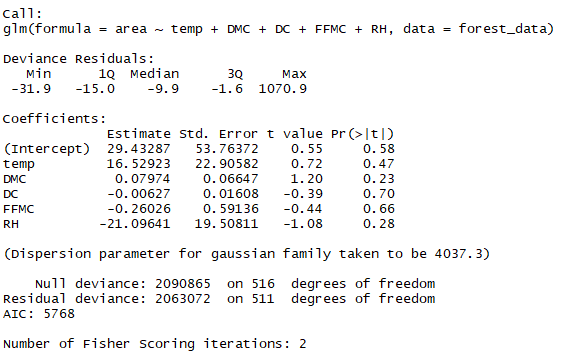
Logistic Regression

*lr* = glm(area ~ ., *data* = forest\_data)

summary(lr)

*lr.importance* = glm(area ~ temp + DMC + DC + FFMC + RH, *data* = forest\_data)

summary(lr.importance)



R- Code : <https://pastebin.com/xng65Vhi>