Project Report

Analysis of Forest Fires

Group 5

	Contribution	E-mail
Wenzhuo Wu	linear regression and model diagnostic	wzhwu@ucdavis.edu
Tong Zhen	logistic regression and SVM	tzhen@ucdavis.edu
Hongyuan Wang	AIC and k-NN algorithm	hhywang@ucdavis.edu

Instructor: Emanuela Furfaro

STA 141A - Fundamentals of Statistical Data Science

University of California, Davis

December 8, 2021

A. Introduction and research question

The occurrence of forest fires (also known as wildfires) is a major environmental concern because it affects forest preservation, causes economic and ecological damage (Bowman et al., 2009). The wildfire will bring tragedy to the family who lives near the forest region, and they have to run away from their home and lose their valuables or sometimes even their life. Additionally, from the ecological perspective, wildfire is a natural phenomenon that creates disturbances to nature and provides benefits to the ecosystem. The best way to reduce the damage is fast detection of wildfire and accurately predict this occurrence. Therefore, evaluating the relationship between environmental index and occurrence of forest fires from the previous data may provide insights on how risk factors are associated, and which index can be used for future prediction of forest fires to improve safety.

The data collected from the northeast region of Portugal (Cortez and Morais, 2007) was used in this project. The objective of this project is to 1) build a model to predict the occurrence of forest fires and 2) build a model to predict the size of the burned area.

To achieve these goals, we will answer four specific questions:

1. Which environmental attributes have significant effects on the occurrence of forest fire?

2. Are there any environmental factors that can reduce the likelihood of a forest fire?

3. After the forest fire occurs, which environmental attributes have the greatest impact on the total burned area?

4. Are there any environmental factors that can reduce the size of the burned area?

B. Dataset Introduction

In this dataset, there are a total of 517 instances and 13 attributes. The attributes include the x and y location of the forest, the month and day, FFMC index, DMC index, DC index, ISI index, temperature, relative humidity, wind speed, rain (Table 1). The burned area of the forest is the output of the model which is measured in hectares.

Table 1. Attribute description

Attribute	Description
X	X-axis coordinate
Y	Y-axis coordinate
month	Month of the year
day	Day of the week
FFMC	The relative ease of ignition and flammability of the fuels
DMC	The depth that fire will burn in the moderate organic layer of the soil and medium-size woody materials

DC	The depth that fire will burn in the deep organic layer of the soil and the large size of woody materials
ISI	The rate that a fire will spread in the early stages
temp	Temperature (in °C)
RH	Relative humidity (in %)
wind	Wind speed (in km/h)
rain	Rain (in mm/m^2)
area	Total burned area (in ha)

In this project, we will focus on the index predictors (FFMC, DMC, DC and ISI) and the four environmental variables (temperature, relative humidity, wind and speed) to predict occurrence of the fire and total burned area.

C. Data visualization

The function chart.Correlation() was used to display a chart of a correlation matrix (Fig. 1). The correlation coefficient of DC and DMC (0.68), the correlation coefficient of ISI and FFMC (0.53) and the correlation coefficient of temp and RH (-0.53) are the top three highest in absolute value. However, the response variable area has low correlation coefficients with all explanatory variables.

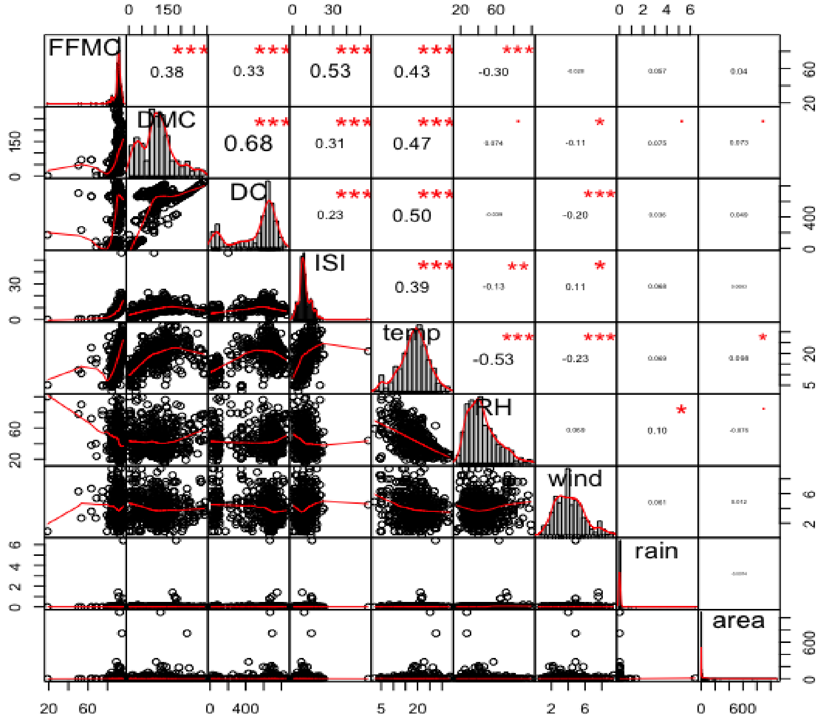

Figure 1. The distribution and correlation of each variable. (The bivariate scatter plot with fitted line showed on bottom of the diagonal and the correlation value on the top). p-values (0, 0.001, 0.01, 0.05, 0.1, 1) was represented by symbols(“*** ”, “**”, “*”, “.”, " “)

The dataset was separated into two categories to better compare the influence of environmental index on the absence and presence of forest fire. Group one has a burned area of 0 (Class = Absence), while group two has a burned area which is not equal to 0 (Class = Presence). Figure 2 shows the distribution of each variable for these two groups, and there isn't much of a difference.

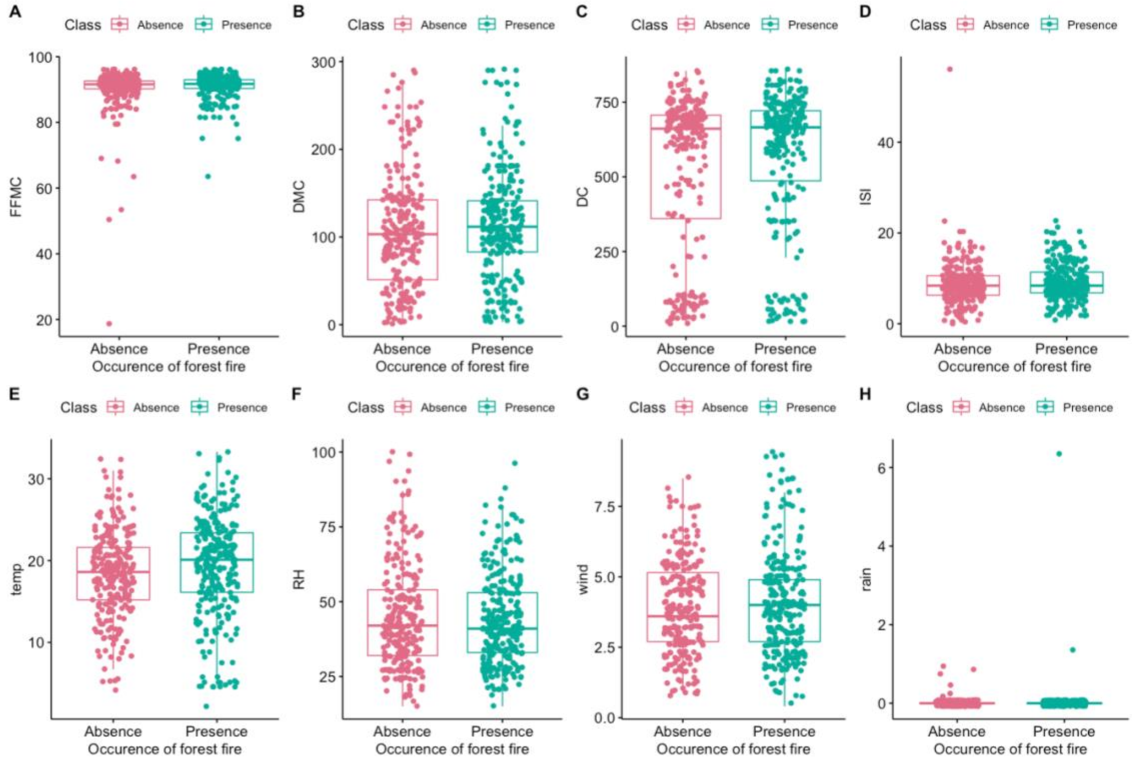

Figure 2. The distribution of each explanatory variable by occurrence of forest fire. (A-H: FFMC, DMC, DC, ISI, temp, RH, wind and rain)

D. Method

The project has two stages. The first stage is to predict whether the burn area is 0 or not.   In this case, we applied several binary classification methods. We compared the results from logistic regression, support vector machines (SVM), random forests and k-NN algorithms to classify the location that will have potential wildfire. According to Cortez and Morais (2007), the four environmental data, including temperature, rain, relative humidity, and wind speed can predict small fires with better performance. All the models will focus on these four environmental predictors. We created both training and testing dataset from the original data. Testing data is used to create a confusion matrix and calculate error rate for each model. The framework of the confusion matrix is shown in Table 2 below.

Table 2. Framework of the confusion matrix to be used in model-wide comparison.

	Predicted Wildfire Absent	Predicted Wildfire Present
Actual Wildfire Absent	True Negatives	False Positives
Actual Wildfire Present	False Negatives	True Positives

The second stage is to predict the size of the burned area if forest fire happens. The data with burned area greater than 0 was used in this part. Model selection was performed to find the best model to run linear regression. Assumptions of linear regression were checked.

E. Result and discussion

Stage one: occurrence of forest fire (Classification)

Several classification methods were used in this part. For the logistic regression analysis, a full model was developed first. The performance of the model is not satisfactory based on the confusion matrix result which is shown in table 3. The value of false positives and false negatives are high. In addition, no predictor is significant in the model. However, wind, DC and DMC are more important than other predictors due to low p-values (0.146, 0.141 and 0.137).

Table 3. Confusion matrix of the full logistic model, including predictors: FFMC, DMC, DC, ISI, temp, RH, wind and rain)

Logistic Mmodel (Full)	Predicted Wildfire Absent	Predicted Wildfire Present
Actual Wildfire Absent	30	19
Actual Wildfire Present	55	13

Table 4. Summary of full logistic model

	Estimate	**Standard error**	**t value**	**P**
(intercept)	-3.506e+00	2.838e+00	-1.235	0.217
FFMC	3.495e-02	3.032e-02	1.152	0.249
DMC	4.669e-03	3.143e-03	1.485	0.137
DC	0.0009924	5.964e-04	1.471	0.141
ISI	-0.0230860	3.052e-02	-0.780	0.435
temp	-0.0356216	3.263e-02	-0.869	0.385
RH	-0.0089571	9.334e-03	-0.498	0.619
wind	0.0734126	6.364e-02	1.453	0.146
rain	-7.343e+01	3.106e+03	-0.024	0.981

Then the reduced model was tested based on the four environmental predictors we are interested in. This model is also not satisfactory. According to Table 5, even though we have a small number of false positives, we have extremely large amounts of false negatives. However, this model is good at predicting the case of no actual wildfires. In the reduced model no predictor is significant. However, wind and temp are more important than other predictors because of low p-values (0.187 and 0.160).

Table 5. Confusion matrix of the reduced logistic model, including predictors: temp, rain, RH and wind.

Logistic Model (temp + rain + RH + wind)	Predicted Wildfire Absent	Predicted Wildfire Present
Actual Wildfire Absent	47	2
Actual Wildfire Present	67	1

Table 6. Summary of reduced logistic model

	Estimate	**Standard error**	**t value**	**P**
(intercept)	-1.04906	0.80226	-1.308	0.191
wind	0.07915	0.06004	1.318	0.187
RH	0.00373	0.00792	0.471	0.638
temp	0.03391	0.02412	1.406	0.160
rain	-73.40205	3091.2310	-0.024	0.981

The next classification model we used is the SVM. Table 7 shows that this model has a large number (36) of false positives, while it still has 10 false negatives. This model is good at predicting the scenario that wildfire presents. In terms of random forest classification, this model has a similar number of false positives (33) and false negatives (25) (table 8).

Table 7. Confusion matrix of the support vector machines model (SVM), including predictors: temp, rain, RH and wind.

SVM (temp + rain + RH + wind)	Predicted Wildfire Absent	Predicted Wildfire Present
Actual Wildfire Absent	13	36
Actual Wildfire Present	10	58

Table 8. Confusion matrix of the random forests model, including predictors: temp, rain, RH and wind.

Random Forests (temp + rain + RH + wind)	Predicted Wildfire Absent	Predicted Wildfire Present
Actual Wildfire Absent	16	33
Actual Wildfire Present	25	43

For the k-NN model, the best k-value was determined based on Figure 3. Then the confusion matrix was computed (Table 9). In this model, there are a large amount of false positives and a relatively small amount of false negatives. Therefore, this model is also good at predicting wildfires when wildfires occur.

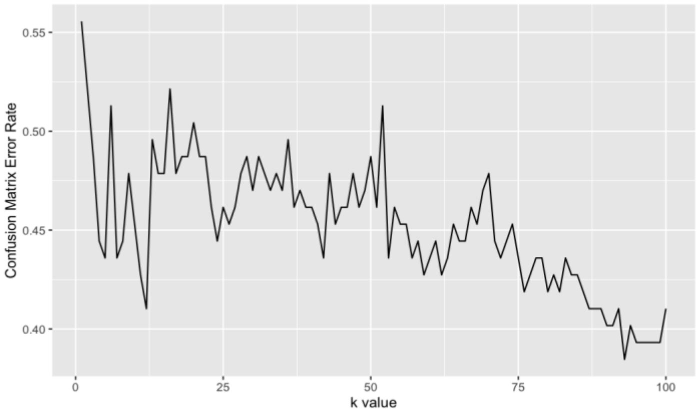

Figure 3. K values in k-NN model against the confusion matrix error rate. The k-NN model includes predictors: temp, rain, RH and wind.

Table 9. Confusion matrix of the k-NN (k=93) model, including predictors: temp, rain, RH and wind.

k-NN (k =93)	Predicted Wildfire Absent	Predicted Wildfire Present
Actual Wildfire Absent	10	39
Actual Wildfire Present	6	62

Based on the evaluation of all the models and their confusion matrix, the overall performance of all the models is not satisfactory. According to Table 10, the logistic model performs the worst with an error rate 0.59. SVM and k-NN have similar performance with the error rate of 0.39 and 0.38 respectively. Additionally, both SVM and k-NN have a relatively high false positive rate. In real-life situations and wildfire prediction, the result of having a false positive is much better than having a false negative, because it is better to monitor all the false positives and prevent the occurrence of wildfires for fire prevention. However, the cost of a wildfire is much more expensive than the cost of monitoring and prevention. Therefore, having a lower false negative error rate is more important than false positive error rate. Then the adjusted error rates were calculated if we accept false positives as a correct result. In Table 8, the adjusted error rates of SVM and k-NN drop to 0.08 and 0.05 respectively. However, in this case, the logistic model still has a high adjusted error rate. Overall, SVM has the best performance of predicting the occurrence of wildfires.

Table 10. Confusion matrix error rate and adjusted error rate (only consider false negatives) of different classification models.

	Logistic Model	SVM	Random Forests	k-NN
Error rate	0.59	0.39	0.49	0.38
Adjusted error rate	0.57	0.08	0.21	0.05

Stage two: size of burned area (Linear regression)

The second stage is to predict the size of the burned area if forest fire happens. The data with burned area greater than 0 was used in this part. Area was transformed as log(area+1). Model selection was performed to find the best model to run linear regression. The result showed the model with wind, RH, DMC and ISI has the lowest AIC (Akaike's Information Criteria), which can be used to measure the goodness of fit.

Reduced model: log(area+1) ~ DMC + ISI + RH + wind

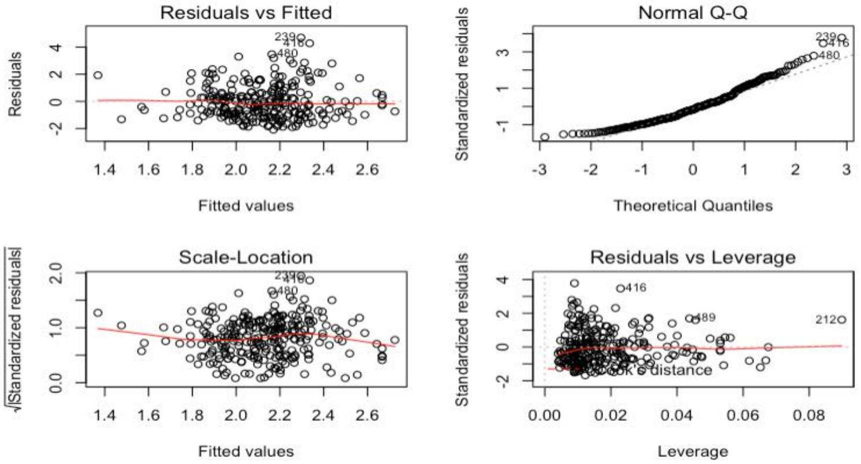

Figure 5. Plots of reduced linear regression model

From the Normal Q-Q plot (Fig. 5), the model does not satisfy the normality assumption because some points deviate from the regression line and form positively skewed graph plots. From the Residual vs Fitted plot, some points are not distributed randomly around 0. Moreover, from the Scale-Location plot, the variance does not keep unchanged as the fitted value increases, which means the equal variance assumption may be violated. From the Residuals vs Leverage plot, there is still one high leverage point on the right. There are still some outliers (eg. #416). From the summary (Table 11), the predictor ISI is significant in predicting the log transformed size of burned area. Due to the small R^2 and high p-value, the linear relationship between predictors and response variables cannot be proved.

Table 11. Summary of reduced model

	Estimate	**Standard error**	**t value**	**P**
(intercept)	2.395547	0.344855	6.847	2.89e-11 ***
wind	0.059001	0.041733	1.414	0.1586
RH	-0.008167	0.005200	-1.571	0.1175
DMC	0.002144	0.001334	1.607	0.1091
ISI	-0.043542	0.019970	-2.180	0.0301 *

F. Conclusion

There are two stages in this project. Several binary classification methods were used to predict the occurrence of the forest fire. However, the performance of models was not satisfactory. Overall, SVM and k-NN have the best performance of predicting the occurrence of wildfires. SVM has an error rate of 39% and false negative rate of 8%. The k-NN model has an error rate of 38% and false negative rate of 5%. Then linear regression was performed to find the relation between burned area with other variables. Based on the result, ISI has a significant impact on the total burned area. However, the linear regression model is not significant. Due to the nature property of the data, a suitable linear regression model cannot be obtained.

Which environmental attributes have significant effects on the occurrence of forest fire?

Based on the summary of full logistic regression, none of the environmental attributes are significant at significance level 0.01. Among all predictors, DMC, DC and wind are more likely to be the significant.

Are there any environmental factors that can reduce the likelihood of a forest fire?

The estimates of ISI, temp, RH and rain are negative, which means the response variable will decrease as these factors increase.

After the forest fire occurs, which environmental attributes have the greatest impact on the total burned area?

After the performance of the stepAIC function, the four factors which are relatively significant among all factors are chosen to form the new model. Moreover, the factor ISI has the smallest p-value 0.0301, which indicates factor ISI has the greatest impact on the total burned area.

Are there any environmental factors that can reduce the size of the burned area?

By comparing the estimated coefficients of environmental factors, we can conclude that RH and ISI, which have negative estimates, can reduce the size of the burned area.

The overall performance of this project indicates that wildfire is an event that is difficult to predict. Even though the accuracy is low in all the models, the data that the models used are easy and cheap to collect. We can collect the data from the weather stations in real-time and run the models in real-time. By applying the models, we can use the predictions to help with fire prevention management and try to save more lives before this natural hazard happens. The size of wildfire and the spread of wildfire is still an unsolved question, and it required more research to build a sufficient model.

Reference

Bowman, D. M., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., D'Antonio, C. M., DeFries, R. S., Doyle, J. C., Harrison, S. P., Johnston, F. H., Keeley, J. E., Krawchuk, M. A., Kull, C. A., Marston, J. B., Moritz, M. A., Prentice, I. C., Roos, C. I., Scott, A. C., … Pyne, S. J. 2009. Fire in the earth system. Science, 324(5926), 481–484. https://doi.org/10.1126/science.1163886

Cortez P. and Morais A.2007. A Data Mining Approach to Predict Forest Fires using Meteorological Data..In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523

Appendix

```
# Libraries
library(ggplot2)
library(MASS)
library(class)
library(e1071)
library(randomForest)
library(tidyverse)
library(colorspace)
library(ggpubr)
library("PerformanceAnalytics"
# Data visualization
raw.data <- read.csv("forestfires.csv",header = T)
data <- raw.data [5:13]
library("PerformanceAnalytics")
chart.Correlation(data, histogram=TRUE, pch=19)
data$Class <- factor(ifelse(data$area ==0, 'Absence','Presence'))
A <- ggboxplot(data, x="Class", y="FFMC", color="Class",
          palette = qualitative_hcl(n=2, palette = "Dark3"), add="jitter", xlab="Occurence of forest fire",
ylab="FFMC")
B <- ggboxplot(data, x="Class", y="DMC", color="Class",
          palette = qualitative_hcl(n=2, palette = "Dark3"),add="jitter", xlab="Occurence of forest fire",
ylab="DMC")
C <- ggboxplot(data, x="Class", y="DC", color="Class",
          palette = qualitative_hcl(n=2, palette = "Dark3"), add="jitter",xlab="Occurence of forest fire",
ylab="DC")
D <- ggboxplot(data, x="Class", y="ISI", color="Class",
          palette = qualitative_hcl(n=2, palette = "Dark3"), add="jitter",xlab="Occurence of forest fire",
ylab="ISI")
E <- ggboxplot(data, x="Class", y="temp", color="Class",
          palette = qualitative_hcl(n=2, palette = "Dark3"), add="jitter",xlab="Occurence of forest fire",
ylab="temp")
F <- ggboxplot(data, x="Class", y="RH", color="Class",
          palette = qualitative_hcl(n=2, palette = "Dark3"), add="jitter",xlab="Occurence of forest fire",
ylab="RH")
G <- ggboxplot(data, x="Class", y="wind", color="Class",
          palette = qualitative_hcl(n=2, palette = "Dark3"), add="jitter",xlab="Occurence of forest fire",
ylab="wind")
H <- ggboxplot(data, x="Class", y="rain", color="Class",
            palette = qualitative_hcl(n=2, palette = "Dark3"), add="jitter",xlab="Occurence of forest fire",
ylab="rain")

ggarrange(A,B,C,D,E,F,G,H,
        labels = c("A", "B", "C","D","E", "F","G", "H"),
        ncol = 4, nrow = 2)
# Logistic Regression
# Functions that created to calculate error rate and error rate without false positive
er.cal <- function(x)
{
  return(1 - sum(diag(x))/sum(x))
}

er.fp.cla <- function(y)
{
  return(y[2,1]/sum(y))
}
# Full model
fire = read.csv('forestfires.csv', header = T)
fire$month = as.factor(fire$month)
fire$day = as.factor(fire$day)
fire$area = ifelse( test = fire$area == 0, yes = 0, no = 1)
fire$area = as.factor(fire$area)
train.data = fire[1:400,]
test.data = fire[401:517,]
fire_model = glm(area ~  X + DC + wind, data = train.data, family = 'binomial')
y_pred_glm = predict(fire_model, newdata = test.data[c('X','DC','wind')])
summary(fire_model)
confusion <- table(ifelse(y_pred_glm  > 0.5, 1, 0), test_glm[,13])
confusion
er.cal(confusion)
er.fp.cla(confusion)
# AIC for logistic regression
stepAIC(fire_model, direction = "backward")
# Reduced model
fire_model_reduced <- glm(area ~ DMC+DC+temp+RH, data = train_glm, family = 'binomial')
y_pred_glm = predict(fire_model_reduced, newdata = test_glm[c('DMC','DC','temp','RH')])
summary(fire_model_reduced)
confusion <- table(ifelse(y_pred_glm  > 0.5, 1, 0), test_glm[,13])
confusion
er.cal(confusion)
er.fp.cla(confusion)
# SVM
classifier = svm(formula = area ~ temp + RH + wind,
                 data = train.data,
                 type = 'C-classification',
                 kernel = 'linear')
y_pred = predict(classifier, newdata = test.data[,9:11])
confusion = table(test.data[, 13], y_pred)
confusion
er.cal(confusion)
er.fp.cla(confusion)
# Random Forest
set.seed(56)
classifier_3 = randomForest(x = train.data[c('temp',  'RH' , 'wind')],
                            y = train.data$area,
                            ntree = 500)
y_pred_3 = predict(classifier_3, newdata = test.data[c('temp',  'RH' , 'wind')])
confusion = table(test.data[, 13], y_pred_3)
er.cal(confusion)
er.fp.cla(confusion
# KNN
set.seed(49)
data <- read.csv("forestfires.csv")
area.complete <- as.factor(ifelse(data$area == 0, 0, 1))
data.complete <- data.frame(data ,area.complete)
data.train <- data.complete[1:400, ] # Data 1-400 as train set
data.test <- data.complete[401:517, ] # Data 401-517 as test set
cr <- numeric(0)
fp_error <- numeric(0)
for (i in 1:100) {
  predict.test <- knn(train = data.frame(data.train[, c("temp","RH","wind","rain")]),
                    test = data.frame(data.test[, c("temp","RH","wind","rain")]),
                    cl = data.train[, 14],
                    k = i,
                    use.all = T)
  cm <- table(data.test[, 14], predict.test)
  cr[i] <- er.cal(cm)
  fp_error[i] <- er.fp.cla(cm)
}
df.cm <- data.frame(k.value = 1:100, error.rate = cr, error.rate.fp = fp_error)
knn.error = ggplot(data = df.cm) +
  geom_line(mapping = aes(x = k.value, y = error.rate)) + xlab('k value') + ylab('Confusion Matrix Error
Rate')
knn.error.fp = ggplot(data = df.cm) +
  geom_line(mapping = aes(x = k.value, y = error.rate.fp)) + xlab('k value') + ylab('Confusion Matrix Error
Rate (including false positives)')
knn.error
knn.error.fp
df.cm.min = df.cm[df.cm$error.rate == min(df.cm$error.rate),]
df.cm.min.fp = df.cm[df.cm$error.rate.fp == min(df.cm$error.rate.fp),]
df.cm.min
df.cm.min.fp
# Linear regression
raw.data <- read.csv("forestfires.csv",header = T)
data <- raw.data[5:13]
data.re <- data[data$area!=0,] # remove burn area =0
lreg <- lm(area ~ FFMC + DMC + DC + ISI + temp + RH + wind , data = data.re)
summary(lreg)
par(mfrow=c(2,2))
plot(lreg)
# log(area+1) transformation and remove log(area+1)=0
area.trans <- log(data$area+1)
data.trans <- cbind(data,area.trans)
data.trans.re <- data.trans[data.trans$area.trans!=0,]
# multiple linear regression
lreg2<-lm(area.trans ~ FFMC + DMC + DC + ISI + temp + RH + wind, data = data.trans.re)
summary(lreg2)
plot(lreg2
# AIC for linear regression
stepAIC(lreg2, direction = "backward")
# Reduced linear regression
lreg3 <- lm(area.trans ~ DMC + ISI + RH + wind, data = data.trans.re)
summary(lreg3)
plot(lreg3)
```