# Evaluating Room Occupancy Classifiers

*Raaka Mukhopadhyay*
*rmukhopa*

*Due Wed, Dec 2, at 8:00PM (Pittsburgh time)*

## Contents

```r
set.seed(151)
library("knitr")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("MASS")
library("klaR")
library("tree")
library("rpart")
library("rpart.plot")
```

```r
occupancy_train <- readr::read_csv("http://stat.cmu.edu/~gordonw/occupancy_train.csv")
occupancy_test <- readr::read_csv("http://stat.cmu.edu/~gordonw/occupancy_test.csv")
```

## Introduction

In this paper, we hope to use classification techniques for classifying the type of occupancy, either occupied or not occupied, of a room based on several characteristics about the room and time. If the predictions are reasonably accurate, this tool could be utilized for indoor emergency situations. For example, according to the National Fire Protection Association, as of 2018, a structure fire occurred every 63 seconds and a home fire occurred every 87 seconds; this tool could be useful in these situations for deciding how to handle and plan for fire emergency scenarios, as well as other emergency situations.

Citation: https://www.iii.org/fact-statistic/facts-statistics-fire

# Exploratory Data Analysis

## Overview of Dataset and Variables

The *predictor* variables are as followed: `Temperature`: temperature of the room (measured in degrees C) `Humidity`: relative humidity of room (measured in percent) `CO2`: carbon dioxide in room (measured in ppm) `Hour`: hour of the day (measured using hour, 0 to 23)

The *response labels* that we want to predict with our classifiers: `Occupancy`: whether or not the room is occupied (0 for not occupied, 1 for occupied status)

Data from: Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Luis M. Candanedo, Véronique Feldheim. Energy and Buildings. Volume 112, 15 January 2016, Pages 28-39.

## EDA on Y (`Occupancy`)

In the training set, we have a total of 5700 observations. Of these observations, 4497 (or 78.89%) of the rooms were not occupied and 1203 (or 21.20%) of the rooms were occupied. This is shown in the tables below:

```
table(occupancy_train$Occupancy)
```

```
##
##    0    1
## 4497 1203
```

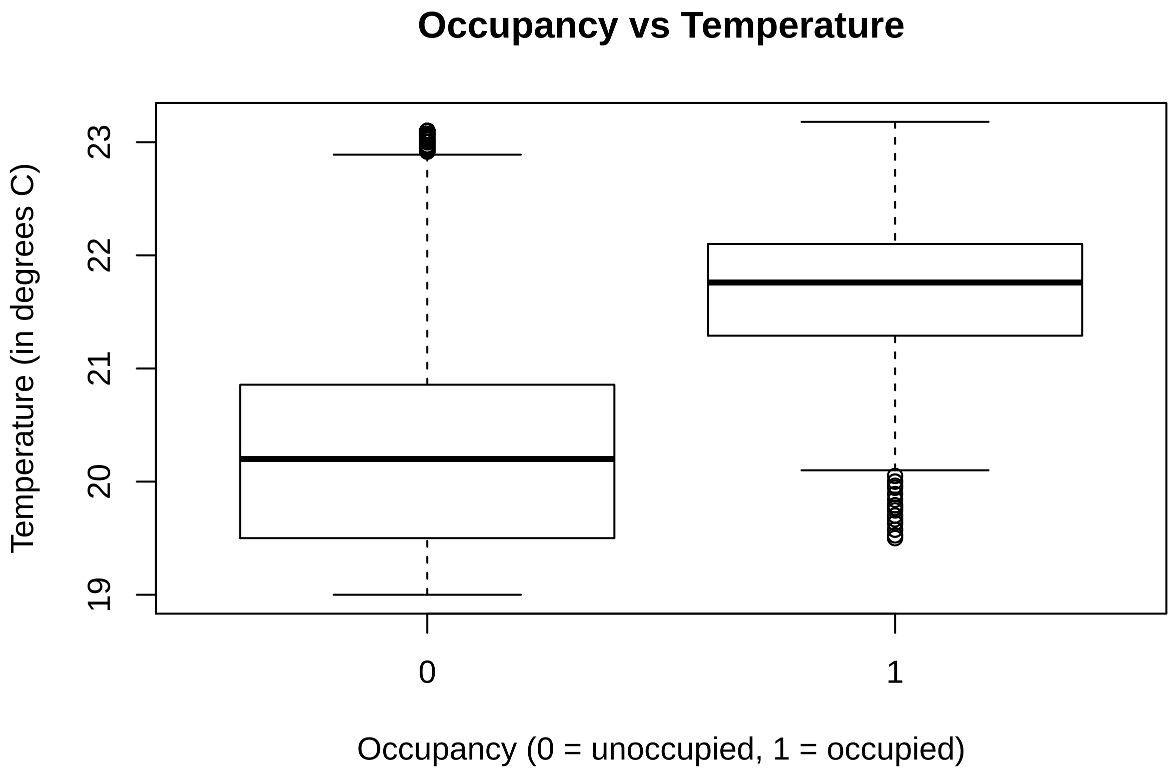
```
prop.table(table(occupancy_train$Occupancy))
```

```
##
##         0         1
## 0.7889474 0.2110526
```

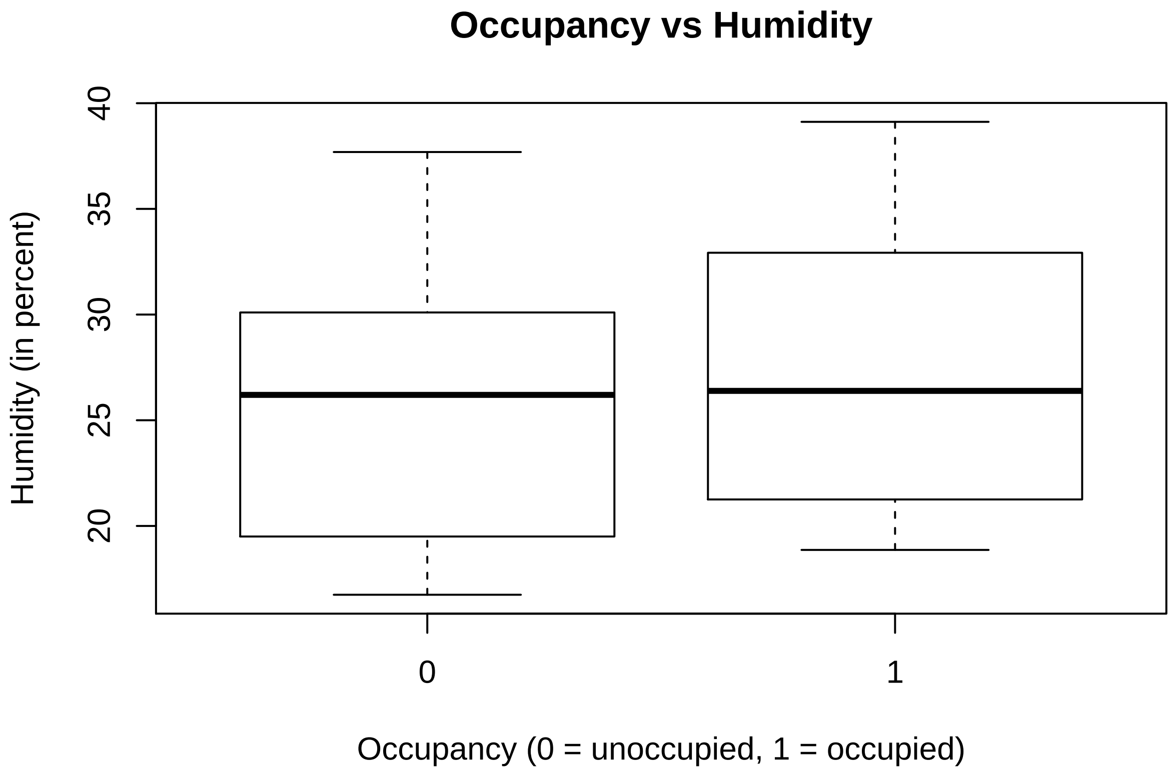
## EDA on relationship between `Occupancy` and each Quantitative Variable (`Temperature`, `Humidity`, `CO2`, `Hour`)

We will now begin to under the relationsip between each respective quantitative predictor (all of the predictors are quantitative) and the response. We will visualize these relationships using boxplots, as seen below:

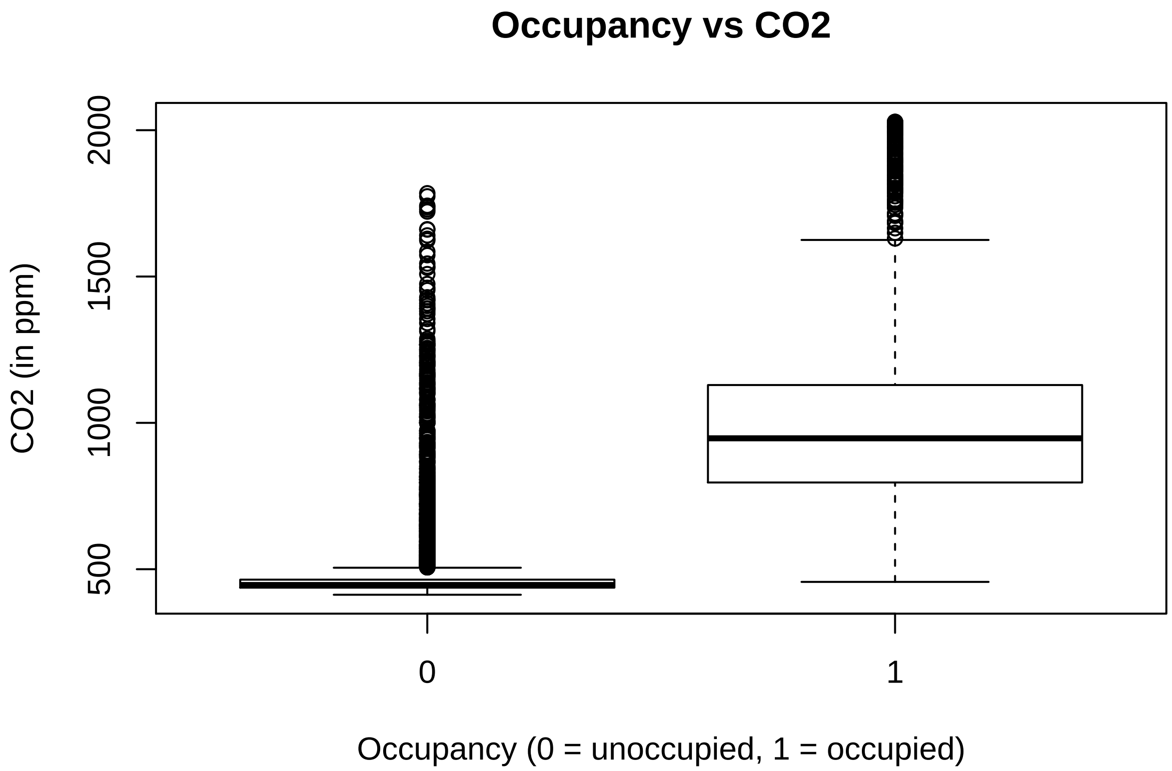
```
boxplot(Temperature ~ Occupancy,
  main="Occupancy vs Temperature",
  xlab = "Occupancy (0 = unoccupied, 1 = occupied)",
  ylab = "Temperature (in degrees C)",
  data = occupancy_train)
```



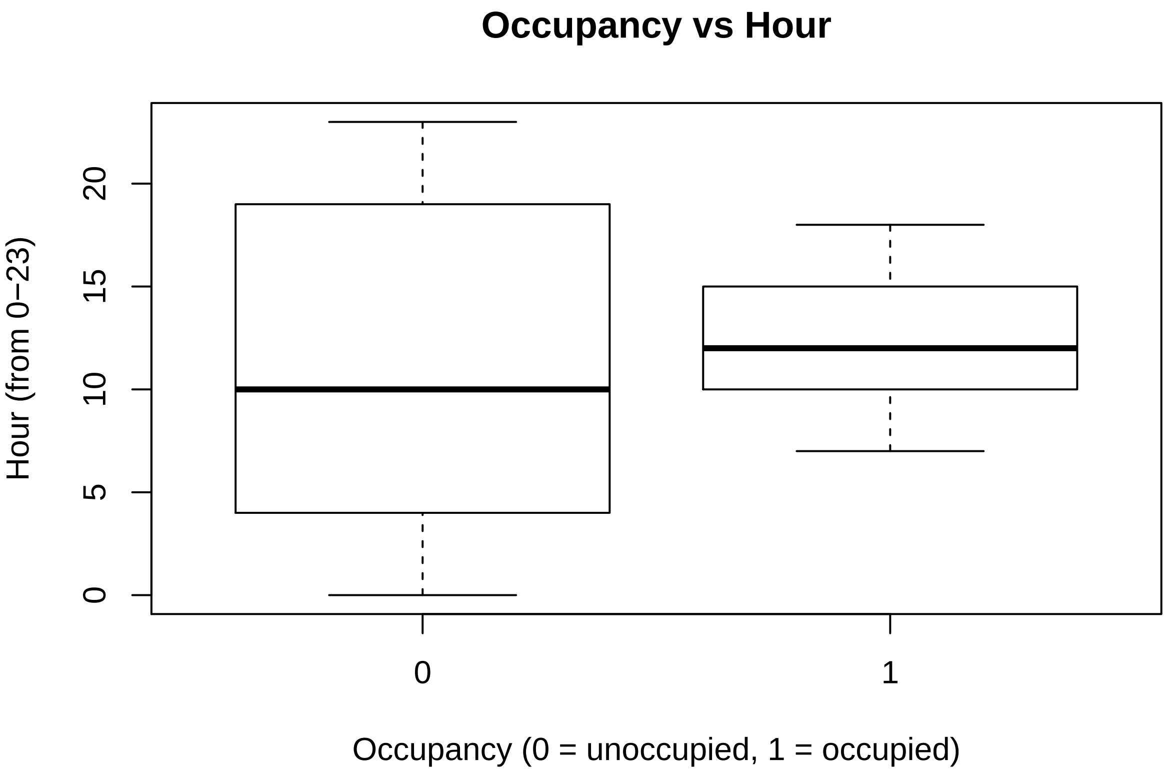
```r
boxplot(Humidity ~ Occupancy,
  main="Occupancy vs Humidity",
  xlab = "Occupancy (0 = unoccupied, 1 = occupied)",
  ylab = "Humidity (in percent)",
  data = occupancy_train)
```



```
boxplot(CO2 ~ Occupancy,
  main="Occupancy vs CO2",
  xlab = "Occupancy (0 = unoccupied, 1 = occupied)",
  ylab = "CO2 (in ppm)",
  data = occupancy_train)
```



```r
boxplot(Hour ~ Occupancy,
  main="Occupancy vs Hour",
  xlab = "Occupancy (0 = unoccupied, 1 = occupied)",
  ylab = "Hour (from 0-23)",
  data = occupancy_train)
```



Using the boxplots, we may observe differences in the predictor variable depending on the occupancy of the room. While noting these differences does not explicitly mean that there is a statistically significant relationship, we note these differences anyway as an indicator of a relationship between the particular predictor variable and the response, which may be helpful in our classifiers.

*Occupancy and Temperature:* We notice that temperature tends to be higher in an occupied room. *Occupancy and Humidity:* We notice that the median and spread of humidity are relatively similar for both occupied and unoccupied rooms. *Occupancy and CO2:* We notice that the ppm of CO2 is greater in occupied rooms; the ppm of CO2 is very low in unoccupied rooms. It is also notable that there are many outliers for both types of occupancy, and the outliers are all greater than maximum value in the fourth quartile. *Occupancy and Hour:* We notice that the median hour is relatively similar for both occupied and unoccupied. However, the spread of the unoccupied rooms is greater than the spread of the occupied rooms.

## Modeling

We now begin creating the classifiers for predicting occupancy of a room. The four classifiers we will use are linear discriminant analysis (LDA), quadratic discriminant analysis (LDA), classification trees, and binary logistic regression.

If any of the predictor variables were categorical, we would exclude them from the LDA and QDA. However, since all of our predictor variables are quantitative, we will not exclude any variables from any of the models.

### LDA

The LDA classifier is created from the training data as such:

```
occupancy_lda <- lda(Occupancy~Temperature + Humidity + CO2 + Hour, data = occupancy_train)
```

We now will observe the performance of the LDA on the test data as such:

```
occupancy_lda_predict <- predict(occupancy_lda, as.data.frame(occupancy_test))
```

```
table(occupancy_lda_predict$class, occupancy_test$Occupancy)
```

```
##
##         0    1
##    0 1844  111
##    1   73  415
```

```
lda_overall_error<-(73+111)/(1844+111+73+415)
lda_0_error<- 73/(1844+73)
lda_1_error<- 111/(111+415)
lda_overall_error
```

```
## [1] 0.07531723
```

```
lda_0_error
```

```
## [1] 0.03808033
```

```
lda_1_error
```

```
## [1] 0.2110266
```

From the test data, the LDA gave an overall error rate was 7.53%. The error rate for unoccupied rooms (3.81%) was lower than the error rate for the occupied rooms (21.10%).

## QDA

The QDA classifier is created from the training data as such:

```
occupancy_qda <- qda(Occupancy~Temperature + Humidity + CO2 + Hour, data=occupancy_train)
```

We now will observe the performance of the QDA on the test data as such:

```
occupancy_qda_predict <- predict(occupancy_qda, as.data.frame(occupancy_test))
```

```
table(occupancy_qda_predict$class, occupancy_test$Occupancy)
```

```
##
##         0    1
##    0 1832   81
##    1   85  445
```

```
qda_overall_error<-(85+81)/(1832+85+81+445)
qda_0_error<- 85/(1832+85)
qda_1_error<- 81/(81+445)
qda_overall_error
```

```
## [1] 0.06794924
```

```
qda_0_error
```

```
## [1] 0.04434011
```

```
qda_1_error
```

```
## [1] 0.1539924
```

From the test data, the QDA gave an overall error rate 6.79%. The error rate for unoccupied rooms (4.43%) was lower than the error rate for the occupied rooms (15.40%).

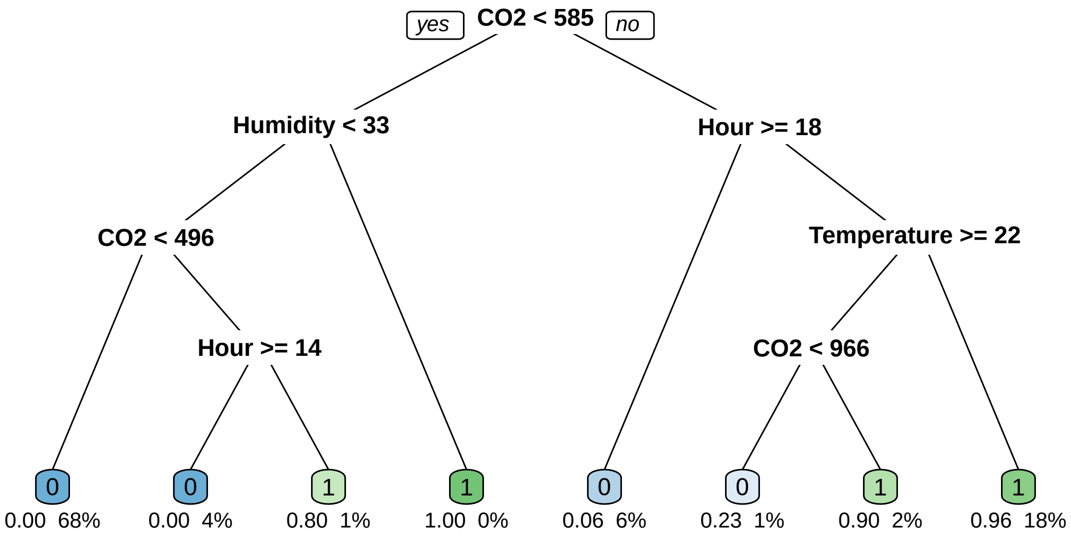
The QDA's overall error rate and error rate for occupied rooms was lower than the LDA's overall error rate and the error rate for occupied rooms. The QDA's error rate for unoccupied rooms was higher than the LDA's error rate for unoccupied room. This could suggest that the QDA may be overfitting for the unoccupied rooms.

## Classification Tree

The Classification Tree is created from the training data as such:

```
occupancy_tree <- rpart(Occupancy~Temperature + Humidity + CO2 + Hour, data=occupancy_train, method = "
```

```
rpart.plot(occupancy_tree,
type = 0,
clip.right.labs = FALSE,
branch = 0.1,
under = TRUE)
```



We note that the classification tree selected CO2 to classify occupancy. We now will observe the performance of the tree on the test data as such:

```
occupancy_tree_predict <- predict(occupancy_tree,as.data.frame(occupancy_test),type="class")
```

```
table(occupancy_tree_predict, occupancy_test$Occupancy)
```

```
## 
## occupancy_tree_predict    0    1
##                      0 1883   15
##                      1   34  511
```

```
tree_overall_error<-(15+34)/(1883+15+34+511)
tree_0_error<- 34/(1883+34)
tree_1_error<- 15/(15+511)
tree_overall_error
```

```
## [1] 0.02005731
```

```
tree_0_error
```

```
## [1] 0.01773605
```

```
tree_1_error
```

```
## [1] 0.02851711
```

From the test data, the tree gave an overall error rate 2.00%. The error rate for unoccupied rooms (1.78%) was lower than the error rate for the occupied rooms (2.85%). All of these error rates are low and lower than all of the error rates from the LDA and QDA. Thus, this could suggest that there may be overfitting.

## Binary Logistic Regression

The binary logistic regression is created from the training data as such:

```
occupancy_logit <- glm(factor(Occupancy) ~ Temperature + Humidity + CO2 + Hour,
data = occupancy_train, family = binomial(link = "logit"))
```

```
occupancy_logit_prob <- predict(occupancy_logit, as.data.frame(occupancy_test),
type = "response")
```

We now will observe the performance of the binary logistic regression on the test data as such:

```
levels(factor(occupancy_test$Occupancy))
```

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

```
occupancy_logit_predict <-ifelse(occupancy_logit_prob > 0.5,"1","0")
```

```
table(occupancy_logit_predict, occupancy_test$Occupancy)
```

```
## 
## occupancy_logit_predict    0    1
##                       0 1849   96
##                       1   68  430
```

```
blr_overall_error<-(68+96)/(1849+68+69+430)
blr_0_error<- 68/(1849+68)
blr_1_error<- 96/(96+430)
blr_overall_error
```

```
## [1] 0.06788079
```

```
blr_0_error
```

```
## [1] 0.03547209
```

```
blr_1_error
```

```
## [1] 0.1825095
```

From the test data, the binary logistic regression (using threshold probability of 0.5) gave an overall error rate 6.79%. The error rate for unoccupied rooms (3.54%) was lower than the error rate for the occupied rooms (1.83%).

In comparison to the other models, the binary logistic regression error rate for the occupied rooms is the lowest. The binary logistic regression error rate for unoccupied rooms is higher than that of the classification tree but lower than LDA and QDA. The binary logistic regression overall error rate is higher than that of the classification tree, about the same as that of the QDA, and lower than that of the LDA.

### Final Recommendation

Ultimately, after testing all the classifiers, we would recommend choosing the classification tree because it had the lowest overall error rate, as well as the lowest error rate for unoccupied rooms and error rate for occupied rooms.

We note that all of the classifier models performed better with respect to unoccupied rooms than occupied rooms. QDA and binary logistic regression performed relatively similarly, with binary logistic performing slightly better. LDA performed the worst out of all the classifiers.

Thus, our final recommendation is the classification tree. However, given the low error rates, we consider the possibility that this may be a result of overfitting. If this is the case, the binary logistic regression classifier would be our secondary recommendation.

## Discussion

Overall, the classification tree performed the best out of all the classifiers and is our final recommendation. We take note of the possibility of overfitting and provide a secondary recommendation in that event.

In the future, we suggest identifying more predictor variables, such as size of room or seating places in a room. Thus, we can potentially create better classifiers for room occupancy. This work can then be used more accurately in the event of emergency situations and can help plan for emergency situations as well.