**Exploratory Data Analysis**

Attributes:

1. Date
2. Temperature
3. Humidity
4. Light
5. co2
6. HumidityRatio
7. Occupancy

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **datatest.txt** | **datatest2.txt** | **datatraining.txt** |
| **Rows** | 2665 | 9752 | 8143 |
| **Columns** | 7 | 7 | 7 |

**Boxplot for the attributes**

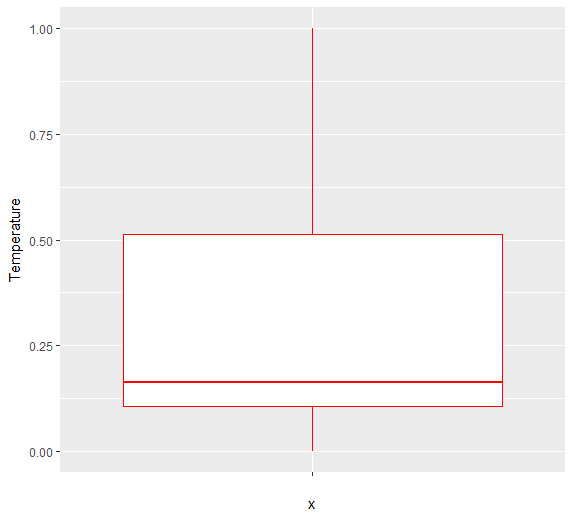


Figure 1: Temperature

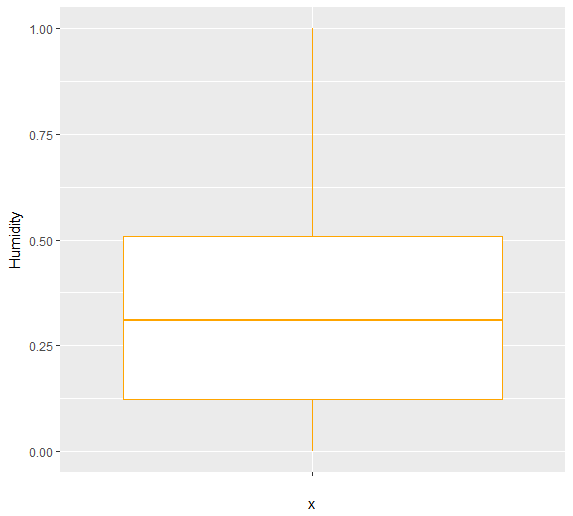


Figure 2: Humidity

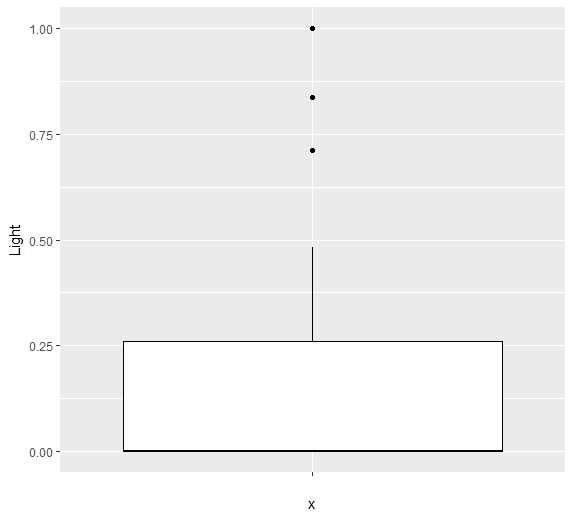


Figure 3: Light

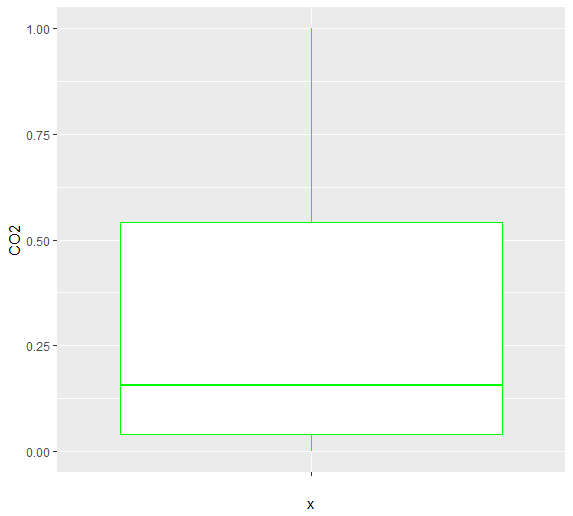


Figure 4: CO2

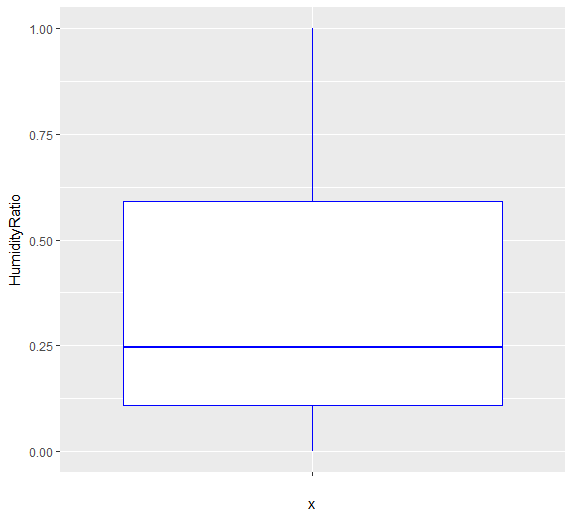


Figure 5: Humidity Ratio

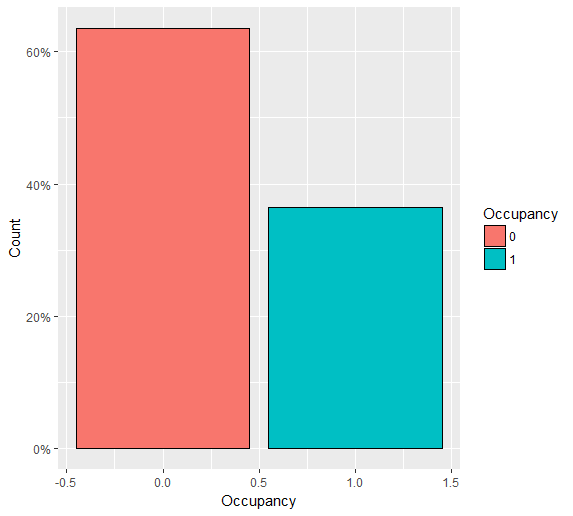


Figure 6: Occupancy

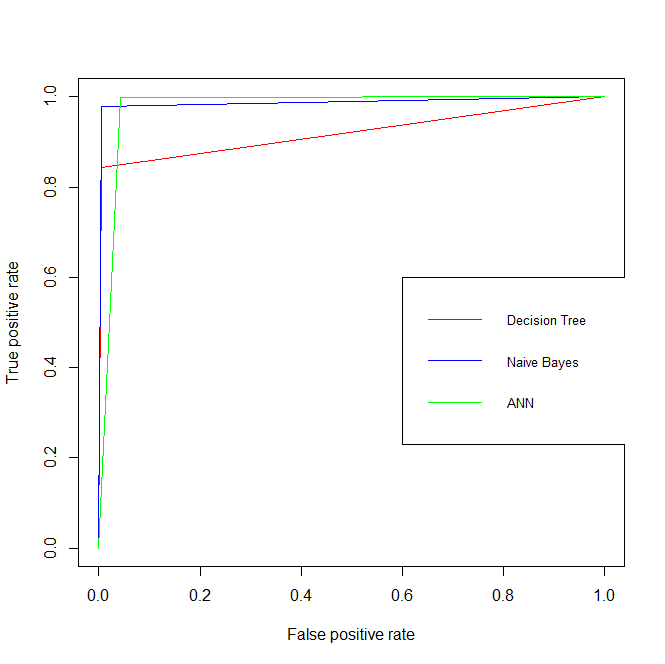
**Pre-Processing**

* No quality issue and no missing value is found in training data, data cleaning is not needed.
* “Date” variable is obviously have no relationship with other variables so it not used as predictor.
* Decision Tree does not need any pre-processing.
* Naïve Bayes, continuous values are discretized into categories.
* ANN, continuous variables must be scaled to -1 to 1 and normalized to SD = 1 and u = 0.

**Performance Measures**

ROC curve is used to measure performance.

**Performance**



Classifiers behaves differently because different algorithm are used to implement them.