Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019

Math 189 Final Project

1 Abstract

The collection of data has become a crucial part of a Data Scientists day-to-day work. This is primarily due to the innovations in data collections methods over the past few years. The sensors available to us today are much more scientifically advanced and can detect data more accurately. This has led to a rise in the use of Machine Learning methods, especially classification algorithms that have proved to be extremely effective on large datasets. Luis M. Candanedo and Véronique Feldheim have compared different supervised learning algorithms in a research paper named "Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models" [1]. In this paper, we will perform an analysis of their work by implementing the various classification algorithms brought about in their work, and compare the performance of these algorithms on the primary dataset used in their work. However, we do not exactly replicate the author's algorithms but perform a similar algorithm on the same dataset.

2 Introduction

The goal of Candanedo and Feldheim's experiment was to detect the accuracy of prediction of occupancy in a room using data collected from light, humidity, CO_2 and temperature sensors [2]. The authors implemented various supervised learning algorithms on the data using the programming software, R. By comparing the performance of these supervised learning algorithms using metrics such as training and testing accuracy, the authors could determine that some classifiers performed better than others.

Our report is an exploration of a subset of classifiers in the original research paper, using the same data set. Python will be the programming language used to implement these classifiers on the data set. Since we take only a subset of classifiers, we leave out a few of the algorithmic techniques and the reasons for doing so will be covered in the next sections.

Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019

3 Data Wrangling and Cleaning

The first step in our project is to clean the dataset. This would involve removing any rows/columns that contain NaN values, since those do not contribute to our overall analysis. We ran the following line of code in our notebook and arrived at the following output:

```
In [7]: # Further check the existence of Nan value
        training_df.isna().sum()

Out[7]: Date             0
        Temperature      0
        Humidity         0
        Light            0
        CO2              0
        HumidityRatio    0
        Occupancy        0
        dtype: int64
```

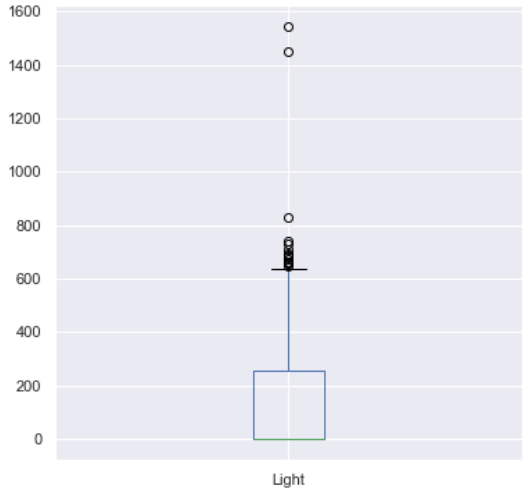
This shows that there is no missing data in our original dataset, and hence we do not need to drop any rows or columns that contain missing data.

4 Exploratory Data Analysis

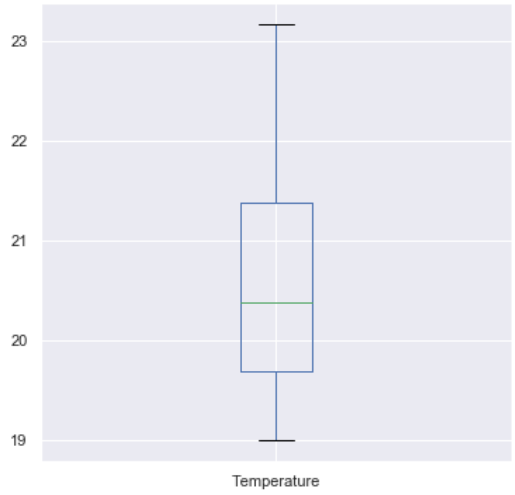
The following table displays the descriptive statistics of the training dataset, such as the central tendency, dispersion and shape of the distribution.

	Temperature	Humidity	Light	CO2	HumidityRatio	Occupancy
count	8142.000000	8142.000000	8142.000000	8142.000000	8142.000000	8142.000000
mean	20.619025	25.730222	119.479153	606.519904	0.003862	0.212233
std	1.016965	5.530334	194.733941	314.331194	0.000852	0.408914
min	19.000000	16.745000	0.000000	412.750000	0.002674	0.000000
25%	19.700000	20.200000	0.000000	439.000000	0.003078	0.000000
50%	20.390000	26.222500	0.000000	453.500000	0.003801	0.000000
75%	21.390000	30.533333	255.875000	638.375000	0.004352	0.000000
max	23.180000	39.117500	1546.333333	2028.500000	0.006476	1.000000

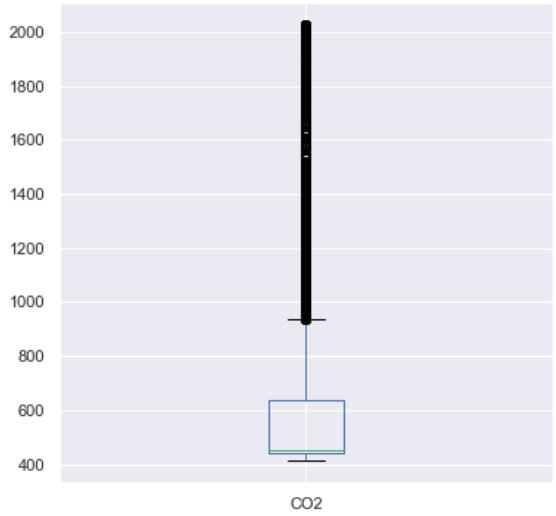
Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019



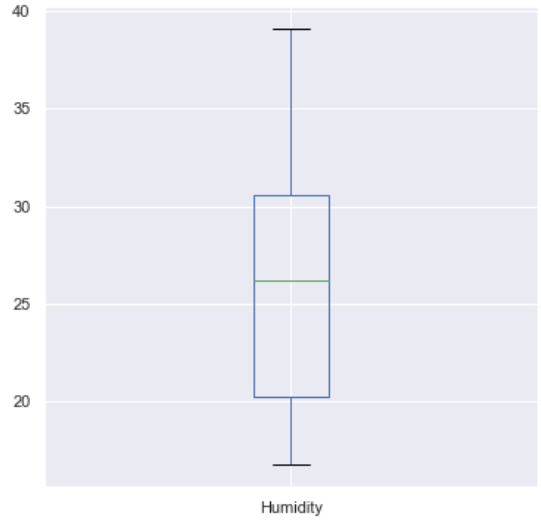
This boxplot represents that for the light variable in our dataset.



This boxplot shows us the boxplot for the temperature variable in our dataset.

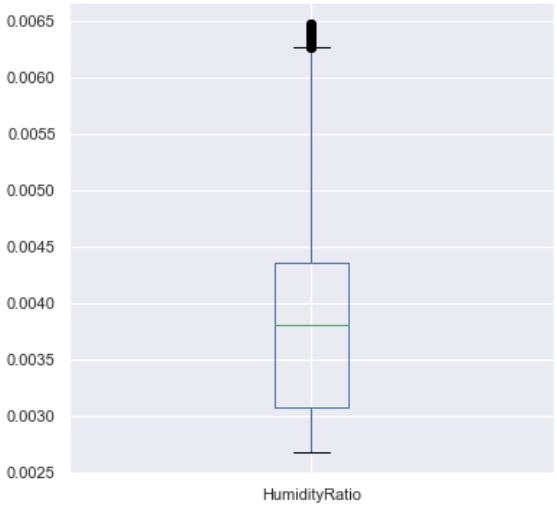


This boxplot shows us the CO2 variable in our dataset.

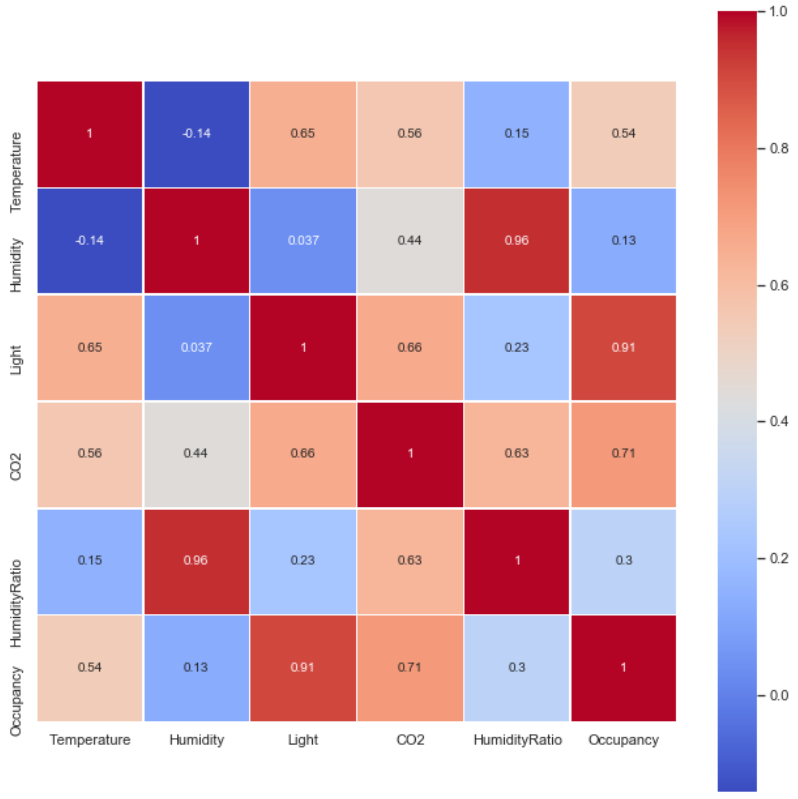


This boxplot shows us the humidity variable in our dataset.

Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019

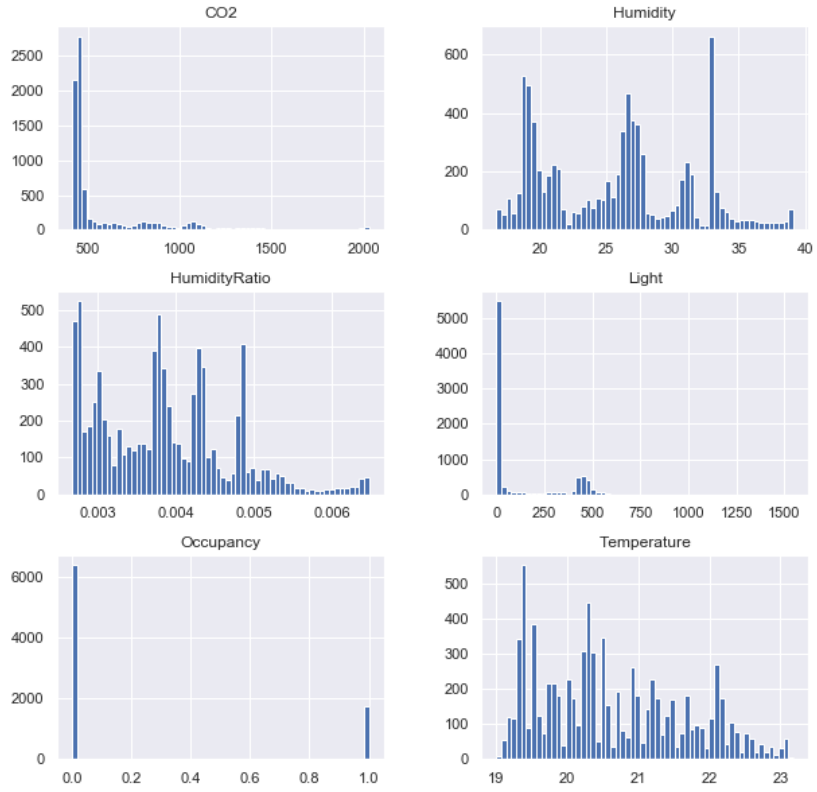


This boxplot shows us the HumidityRatio variable in our data set.



This heat map shows us the correlation between all the variables in our data set.

Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019



These histograms show us the distributions of each of the variables in our data set.

5 Methodology

5.1 Learning Algorithms

In this section, we are going to delve into the various supervised learning algorithms used in this experiment. As mentioned earlier, we only select a subset of the supervised learning algorithms for our experiment.

KNN

The K-Nearest Neighbors classification method measures the distance, as Euclidean or Manhattan distance, between a new data point to a data point in the training set. K-nearest data points are selected, where K can be any integer. The data point is then assigned to a class to which the majority of K data points belong.

When tuning the hyper-parameters, we adjusted the amount of neighbors utilized by the classifier. We found an optimal amount of neighbors, and calculated the accuracy of the classifier based on this optimal amount.

Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019

Decision Tree Classifier

A Decision Tree Classifier creates a tree that predicts the label given multiple numerical inputs. At each node, there is a Boolean expression evaluated based on one or a few of the input data. Depending on the values, the classifier will move to another unique node, eventually reaching a prediction.

When tuning the hyper-parameters, we adjusted the depth of the decision tree utilized by the classifier. We found an optimal depth for the decision tree, and calculated the accuracy of the classifier based on this optimal depth.

Random Forest Classifier

The Random Forest Classifier models many Decision Tree Classifiers with subsets of the original data. Using the average of the predictions of the many tree classifiers, the Random Forest Classifier controls over-fitting. Additionally, Random Forest can naturally rank the variables of a classification problem.

When tuning the hyper-parameters for the random forest, we adjusted the depth of each tree in the forest. We found an optimal depth for each tree, and calculated the accuracy of the classifier based on these values.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis classification allows for different categories or classes to be predicted upon with a high accuracy. It is used to perform dimensionality reduction on a dataset by preserving class information. What separates LDA from other classifications, such as random forest, is that both the predictions and model it provides are easy to understand and interpret.

5.2 Datasets

The overall dataset was divided into testing, training, and validation sets.
For the training dataset, there were six feature variables and one target variable. Furthermore, there are 8142 rows in this dataset. For the testing dataset, there were six feature variables and one target variable. Furthermore, there are 9751 rows in this dataset. For the validation dataset, there were six feature variables and one target variable. Furthermore, there are 2664 rows in this dataset. All three datasets have real-valued numbers with a target variable that is 1 if the room is occupied and 0 when it is not occupied.

Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019

The training dataset was used to train each of the classifiers. The validation set was used to finely tune the parameters of the models to achieve higher accuracy, hence giving us the model with the highest accuracy. Finally, the testing set was used to compute the accuracy of the best performing model on unseen data.

5.3 Experimental Procedure

The classification models chosen in this exploration were K-Nearest Neighbors, Random Forests, Regression Trees and Linear Discriminant Analysis. These specific classifiers were chosen as a subset of those explored in Candanedo and Feldheim's experiment. The different classification algorithms were implemented onto the data set using the programming language, Python.

First, each classifier was trained using the training set. Multiple models of each classifier were produced, each using different parameters. Each of these models were then tested for accuracy against the validation set. The parameters of each classifier were fine tuned and the model with the highest accuracy in each classifier was selected. Each of the fine tuned classification models were implemented on the test set to test for accuracy. The accuracy of each of the classifiers was outputted.

6 Results

Classifier	Test Accuracy
KNN	97.09%
Random Forests	96.94%
Linear Discriminant Analysis (LDA)	98.76%
Regression Tree	95.11%

The test accuracies are on average very high because we utilize all the parameters in the original dataset. Adjusting the hyper-parameters with the validation set also ensured that we had a higher accuracy on the dataset, without the risk of over-fitting the data.

Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019

7 Discussion

Our results show that the occupancy of a room can be accurately predicted utilizing light, carbon dioxide, humidity and temperature variables in a Supervised Learning algorithm. The classifiers used were all relatively accurate, all returning an accuracy of over 95%. The Regression Tree classifier gave the lowest accuracy of 95.11%, and the LDA returned the highest accuracy of 98.76%. The difference in accuracy could be due to the high levels of correlation between temperature and light (0.65), light and carbon dioxide (0.66), and carbon dioxide and humidity ratio (0.63). This would prove beneficial to LDA as predictive power increases when variables are highly correlated. However, this explains why the regression tree wasn't as effective, as it did not take into consideration the relationship between the variables. The random forest model reported an accuracy of 96.94%, which ranks third among our classifiers. The shortcomings of the classifier are that it is only as accurate as the amount of trees in the algorithm. More trees in the forest makes the algorithm have a higher computational cost, thus making the algorithm lag when the forest becomes large. KNN was the classifier that ranked second, with a test accuracy of 97.09%. The KNN performance was inferior to that of LDA due to the differences in scaling of features in the dataset. On the other hand, KNN is a good option for a classification model because it does not make any underlying assumptions about the distribution of the data set. Furthermore, there is no risk of the Curse of Dimensionality since there is a relatively small number of features in this dataset.

The implementation of any machine learning algorithm allows energy resources to be provided or withdrawn based on the amount of people in the room. In a similar study, Candanedo and Feldheim claimed that accurate occupancy detection can save energy anywhere from 30% to 80% when the machine learning algorithm was used inside a Heating, Ventilation and Cooling Machine (HVAC)[3]. Along with the algorithm, sensors are needed throughout the room, as they would take in the readings that are sent to the algorithm. This improvement of technology will lower the operating costs of running a business, as the HVAC is automated to know when, and when not to provide energy to the building.

There are negative implications to the findings however, as room occupancy information can potentially be used for surveillance or malice pursuits. While the algorithm cannot read faces, the information that someone is in a certain room can prove useful, especially when they are being targeted. Additionally, sensors can be rigged to report false values to trick the learning algorithm into believing there are people when there are not, and vice versa. This could lead to energy dissaving, and may interrupt the learning algorithm's accurate prediction.
Finally, as mentioned earlier, Artificial Neural Networks (ANNs) were left out of this experiment. The reason for this is that there are a number of design choices to make in terms of

Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019

Neural Network architecture, and each design choice has a profound impact on the final output and accuracy of room occupancy. Considering we already had high performance with other classifiers, we thought it would be apt to leave ANNs out.

8 Conclusion

Through this experiment, we have seen that it is possible to apply various Supervised Learning algorithms to predict whether a room is occupied or not. All the classifiers we have run on the dataset predict room occupancy at an accuracy of 95% or higher, showing that it is possible to attain high accuracy prediction for the occupancy of the room. Even though we have seen that LDAs perform the best and Regression Trees perform the worst, one must note that it should not be assumed that certain classifiers will always be more accurate than others. Prediction accuracy of classifiers in circumstantial to the data set and the conditions by which the data is analyzed. In other words, the results of this experiment should not be generalized in the field of Machine Learning and Artificial Intelligence.

Jaidev Mirchandani, Nithu Mathew, Samir Navani, Davis Bedingfield
Professor Wenxin Zhou
Math 189
9th June, 2019

9 References

[1] Candanedo Ibarra, Luis & Feldheim, Veronique. (2015). Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Energy and Buildings. 112. 10.1016/j.enbuild.2015.11.071.

[2] Candanedo Ibarra, Luis & Feldheim, Veronique. (2015). Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Energy and Buildings. 112. 10.1016/j.enbuild.2015.11.071.

[3] Candanedo Ibarra, Luis & Feldheim, Veronique. (2015). Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Energy and Buildings. 112. 10.1016/j.enbuild.2015.11.071.