Literature Review, Data Description, and Approach

Jeffrey Fitzpatrick

500728133

Supervisor: Ceni Babaoglu

October 21, 2024

Table of Contents

[Abstract 3](#_Toc179193150)

[Problem 3](#_Toc179193151)

[Themes and Research Questions 3](#_Toc179193152)

[Data 3](#_Toc179193153)

[Techniques and Tools 4](#_Toc179193154)

[Literature Review 4](#_Toc179193155)

[Descriptive Statistics of the Selected Dataset 5](#_Toc179193156)

[Univariate Analysis 5](#_Toc179193157)

[Data Dictionary 5](#_Toc179193158)

[Dependent (target) Variable 6](#_Toc179193159)

[5 Number Summary 6](#_Toc179193160)

[Frequency of Categorical Variables 9](#_Toc179193161)

[Bivariate Analysis 10](#_Toc179193162)

[Pairwise Visualizations 10](#_Toc179193163)

[Correlation Analysis 13](#_Toc179193164)

[GitHub Repository 15](#_Toc179193165)

[Overall Methodology 15](#_Toc179193166)

[References 16](#_Toc179193167)

# Abstract

## Problem

To provide comfort for occupants, commercial buildings rely on heating, ventilation, and air conditioning (HVAC) systems and lighting systems. Large amounts of energy are wasted, especially during non-working hours (Masoso and Grobler, 2010).

To optimize these systems, a research paper (Singh et al., 2018) described an experiment for accurately estimating the number of occupants in a room using “non-intrusive” environment sensors and machine learning (ML) models. Multiple sensor nodes were placed throughout a 6m by 4.6m test room in a wireless sensor network (WSN). Low-cost sensors were deployed at each of the four desks to measure temperature, light, and sound. A CO2 sensor was deployed in the middle of the room to provide the most accurate reading. Additional motion detection sensors were deployed on the ceiling, above the door and large window. Because of privacy concerns, video-based systems are not considered appropriate for detecting occupancy.

Previous research focused on using occupancy detection (i.e., determining whether a room is occupied or not) to save energy. On the other hand, the goal of ML occupancy estimation research is to design adaptive systems that can detect the exact number of occupants, resulting in additional energy savings and improved comfort for occupants.

## Themes and Research Questions

For my project, I have chosen the theme of Classification for building predictive models.

To reduce energy consumption in buildings, I will investigate research questions such as:

* Which of the implemented supervised learning techniques perform the best in predicting occupancy?
* Which types of sensor data show the most promising results? Are there limitations in using certain types of sensor data such as light?
* Based on the research, what alternative techniques could be used for ML-based occupancy estimation?

## Data

For my project, I will be using the Room Occupancy Estimation dataset, available from the [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/). The Room Occupancy Estimation dataset can be downloaded from <https://archive.ics.uci.edu/dataset/864/room+occupancy+estimation>.

The dataset contains over 10000 data points and 18 features. Each feature represents data (temperature, light, sound, motion, or CO2) from a particular sensor. Measurements were recorded over several days in 30 second intervals. The actual occupancy was established by having participants register and record the exact time each time they entered or left the room. At any given time, up to three participants were in the room at the same time. In about 80% of the data points, the room was unoccupied.

## Techniques and Tools

To solve the stated problem, I propose to implement supervised learning algorithms such as random forests and support-vector machines (SVM). In addition, I plan to investigate how an unsupervised learning algorithm such as Principal Component Analysis (PCA) can be used for dimensionality reduction.

For this project, I plan to use R for exploratory data analysis (EDA) and Python for implementing the machine learning algorithms. All the supervised and unsupervised learning algorithms can be implemented using the scikit-learn open-source library. To evaluate the models, I will use multiple performance metrics including accuracy and F1 score.

# Literature Review

At least six pages answering the questions listed earlier in the introduction section:

* What do you already know about the topic?
* What do you have to say critically about what is already known?
* Has anyone else done anything that is exactly the same?
* Has anyone else done anything that is related?
* How does your work fit in with what has gone before?
* Why is your research worth doing in the light of what has already been done?

If you choose to replicate an existing study, you may need to highlight the similarities and differences between the replicated research and your analyses. For example, suppose you decided to replicate study X. In that case, you need to stress whether you will be applying the same set of research questions, using a different time frame of the datasets, following the same approach, or conducting similar analyses. Nevertheless, you might need to compare your outcome with the actual results of the replicated research.

Ceni: While you are writing your literature reviews, write a summary of the related papers that you reviewed.  Write it in your own words and don’t use something from the paper that you don’t understand. A short paragraph about each paper you reviewed should be sufficient.

A literature review is not a review of general machine-learning techniques. You should focus on your research question and similar research questions as well as your dataset and similar datasets. You should be talking about previous approaches that relate to your current project.

By literature, we mean research papers, conference papers, journal papers, books, etc.

References should be cited where you mention them in the section. You provide a reference in the bracket so that the person reading it can go to the source and get more information.

All the references should be listed at the end.

# Descriptive Statistics of the Selected Dataset

The data set contains 10129 data points and 16 features. Each feature represents data (temperature, light, sound, motion, or CO2) from a particular sensor. Measurements were recorded over several days in 30 second intervals.

There are no missing values in the dataset.

## Univariate Analysis

### Data Dictionary

This table describes the variables in the Room Occupancy Estimation data set:

|  |  |  |
| --- | --- | --- |
| Field Name | Data Type | Description |
| Date | Date | Date of observation in YYYY/MM/DD |
| Time | Date | Time of observation in HH:MM:SS |
| S1\_Temp | Continuous | Temperature reading from sensor 1 in degrees Celsius |
| S2\_Temp | Continuous | Temperature reading from sensor 2 in degrees Celsius |
| S3\_Temp | Continuous | Temperature reading from sensor 3 in degrees Celsius |
| S4\_Temp | Continuous | Temperature reading from sensor 4 in degrees Celsius |
| S1\_Light | Integer | Light reading from sensor 1 in lux |
| S2\_Light | Integer | Light reading from sensor 2 in lux |
| S3\_Light | Integer | Light reading from sensor 3 in lux |
| S4\_Light | Integer | Light reading from sensor 4 in lux |
| S1\_Sound | Continuous | Sound reading from sensor 1 in volts (amplifier output read by ADC) |
| S2\_Sound | Continuous | Sound reading from sensor 2 in volts (amplifier output read by ADC) |
| S3\_Sound | Continuous | Sound reading from sensor 3 in volts (amplifier output read by ADC) |
| S4\_Sound | Continuous | Sound reading from sensor 4 in volts (amplifier output read by ADC) |
| S5\_CO2 | Integer | CO2 reading from sensor 5 in ppm |
| S5\_CO2\_Slope | Continuous | Derived slope of C02 values taken in a sliding window  **Note:** This was calculated by fitting a linear regression in a window of 25 points at each instance and calculating the slope of the line. The factor of 25 was chosen by trial and error with respect to the classification accuracy metric. |
| S6\_PIR | Binary | Binary value conveying motion detection from sensor 6:  **0:** No motion events detected in 30 second frame  **1:** At least one motion event detected in 30 second frame |
| S7\_PIR | Binary | Binary value conveying motion detection from sensor 7:  **0:** No motion events detected in 30 second frame  **1:** At least one motion event detected in 30 second frame |
| Room\_Occupancy\_Count | Integer | Number of occupants in the room at one time (ground truth) |

### Dependent (target) Variable

The dependent (target) variable is Room\_Occupancy\_Count.

### 5 Number Summary

This table shows the 5 number summary of the numeric attributes in the data set:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Minimum | Q1 | Median | Q3 | Maximum |
| S1\_Temp | 24.94 | 25.19 | 25.38 | 25.63 | 26.38 |
| S2\_Temp | 24.75 | 25.19 | 25.38 | 25.63 | 29.00 |
| S3\_Temp | 24.44 | 24.69 | 24.94 | 25.38 | 26.19 |
| S4\_Temp | 24.94 | 25.44 | 25.75 | 26.00 | 26.56 |
| S1\_Light | 0.00 | 0.00 | 0.00 | 12.00 | 165.00 |
| S2\_Light | 0.00 | 0.00 | 0.00 | 14.00 | 258.00 |
| S3\_Light | 0.00 | 0.00 | 0.00 | 50.00 | 280.00 |
| S4\_Light | 0.00 | 0.00 | 0.00 | 22.00 | 74.00 |
| S1\_Sound | 0.0600 | 0.0700 | 0.0800 | 0.0800 | 3.8800 |
| S2\_Sound | 0.0400 | 0.0500 | 0.0500 | 0.0600 | 3.4400 |
| S3\_Sound | 0.0400 | 0.0600 | 0.0600 | 0.0700 | 3.6700 |
| S4\_Sound | 0.0500 | 0.0600 | 0.0800 | 0.1000 | 3.4000 |
| S5\_CO2 | 345.0 | 355.0 | 360.0 | 465.0 | 1270.0 |
| S5\_CO2\_Slope | -6.29615 | -0.04615 | 0.00000 | 0.00000 | 8.98077 |

This chart shows the boxplot of the temperature variables by sensor:

A graph showing different colored squares

Description automatically generated

This chart shows the histograms of the temperature variables:

A graph of different temperature levels

Description automatically generated

This chart shows the boxplot of the light variables by sensor. The median values of 0 suggests that the lights were turned off over half the time.

A graph of a boxplot of light values

Description automatically generated

This chart shows the histograms of the light variables:

A group of graphs showing different values

Description automatically generated

This chart shows the boxplot of the sound variables by sensor. Most of the time no sound or little sound was detected.

A graph of a box plot

Description automatically generated

This chart shows the histograms of the sound variables:

A group of graphs showing different types of sensor values

Description automatically generated

This chart shows the boxplot of the CO2 variables:

A graph of co2 values and a slope of co2

Description automatically generated

This chart shows the histograms of the CO2 variables:

A graph of co2 values

Description automatically generated

### Frequency of Categorical Variables

This chart shows the frequency of the S6\_PIR variable used to detect motion:

A graph with a bar and a number of squares

Description automatically generated with medium confidence

The S7\_PIR variable has a similar frequency.

This chart shows the frequency of the Room\_Occupancy\_Count (target) variable. In about 80% of the data points, the room was unoccupied.

A graph with blue squares

Description automatically generated

## Bivariate Analysis

### Pairwise Visualizations

This chart shows the scatter plot matrix of the temperature (S1), light (S1), sound (S1), and CO2 (S5) values:

A collage of images

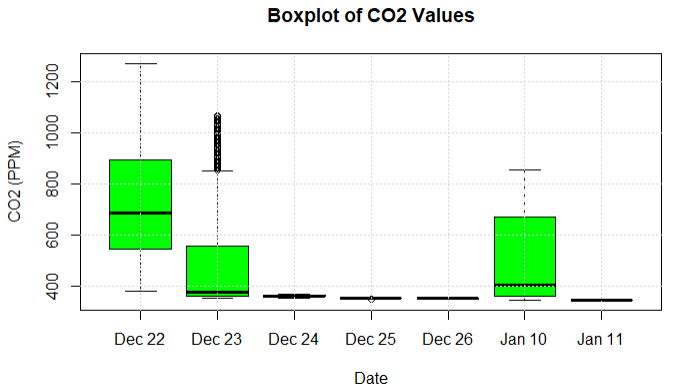
Description automatically generated

This boxplot shows the distribution of room occupancy for each day. There were occupants on December 22, December 23, and January 10 and no occupants on other days.

A graph with green and black squares

Description automatically generated

This boxplot shows the corresponding distribution of CO2 values for each day:

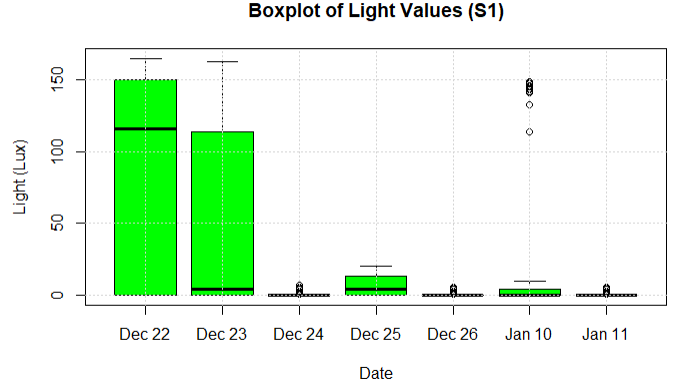


This boxplot shows the distribution of temperature values (sensor 1) for each day. The highest temperatures occurred on the days with occupants.

A graph with green and black squares

Description automatically generated

This boxplot shows the distribution of light values (sensor 1) for each day:



This boxplot shows the distribution of sound values (sensor 1) for each day:

A graph of a box plot

Description automatically generated

### Correlation Analysis

This table summarizes the performance of the linear regression models built to predict room occupancy:

|  |  |  |  |
| --- | --- | --- | --- |
| Model to predict room occupancy | R-Squared | p-values | Residual standard error (RSE) |
| S1\_Temp + S2\_Temp + S3\_Temp +  S4\_Temp | 0.5477 | S1\_Temp: < 2e-16  S2\_Temp: < 2e-16  S3\_Temp: 0.00144  S4\_Temp: < 2e-16 | 0.6011 |
| S1\_Light + S2\_Light + S3\_Light +  S4\_Light | 0.7924 | S1\_Light: < 2e-16  S2\_Light: < 2e-16  S3\_Light: < 2e-16  S4\_Light: < 2e-16 | 0.4073 |
| S1\_Sound + S2\_Sound + S3\_Sound +  S4\_Sound | 0.4441 | S1\_Sound < 2e-16  S2\_Sound: < 2e-16  S3\_Sound: < 2e-16  S4\_Sound: 1.83e-05 | 0.6664 |
| S6\_PIR + S7\_PIR | 0.566 | S6\_PIR: <2e-16  S7\_PIR: <2e-16 | 0.5888 |
| S5\_C02 + S5\_C02\_Slope | 0.7458 | S5\_CO2: <2e-16  S5\_CO2\_Slope: <2e-16 | 0.4506 |

We can compare the models using their Multiple R-Squared values, which refer to the proportion (percentage) of variation in the response predicted by the model. The model with the largest value predicts the most variation in the responses. Based on the table, we can see that light values provided the best predictor of room occupancy count, followed by CO2, motion, temperature, and sound.

This chart shows the correlation matrix between room occupancy and temperature variables:

A red squares with white text

Description automatically generated

This chart shows the correlation matrix between room occupancy and light variables:

A screenshot of a computer

Description automatically generated

This chart shows the correlation matrix between room occupancy and sound variables:

A screenshot of a computer

Description automatically generated

This chart shows the correlation matrix between room occupancy and the CO2 and motion variables:

A person standing next to a red and white square

Description automatically generated

# GitHub Repository

For this project, I’m using the following repository in GitHub:

<https://github.com/jeffreyfitzpatrick/Big-Data-Analytics-Capstone-Project>

# Overall Methodology

A graph showing the tentative overall methodology

Ceni: I'd like to share with you a few tips on the graph you need to prepare for your tentative overall methodology. I suggest you prepare a flowchart for that purpose. A flowchart is a type of diagram that represents a workflow or process. The flowchart shows the steps as boxes of various kinds, and their order by connecting the boxes with arrows.  This diagrammatic representation will illustrate your tentative methodology for your research problem. You should also provide a brief explanation of this flowchart to clarify the purpose of each step and how they contribute to your overall research methodology.

# References

A references page, listing at least six documented articles, following the APA style

1. O.T. Masoso & L.J. Grobler (2010). The dark side of occupants’ behaviour on building energy use. *ScienceDirect, 42*(2), 147-272.
2. A.P Singh, V Jain, S Chaudhari, F.A Kraemer, S Werner and V Garg (2018). Machine learning-based occupancy estimation using multivariate sensor nodes. IEEE Globecom Workshops (GC Wkshps), 1-6.