Documentation- MicroClimate Project

A. Project's Target

The purpose of the "MicroClimate" Project is to develop a reliable machine learning model that identifies occupancy patterns and analyzes indoor environmental quality using multivariate sensor data collected in February 2015 in a European region.

Main Objectives

1. Predict whether a room is occupied or unoccupied based on sensor readings (temperature, humidity, light, CO_2, and humidity ratio).
2. Analyze environmental interactions and how they evolve during different occupancy periods.
3. Detect daily temporal patterns, such as break times or environmental stabilization rates.
4. Provide actionable insights to improve energy efficiency and comfort optimization in smart buildings.

Motivation

Accurate occupancy detection enables intelligent energy management, allowing automatic adjustments to lighting, ventilation, and temperature control systems.
By learning patterns of human activity from environmental sensors, this project contributes to sustainable and adaptive building systems.

B. Dataset Detailed Explanation

Dataset Overview

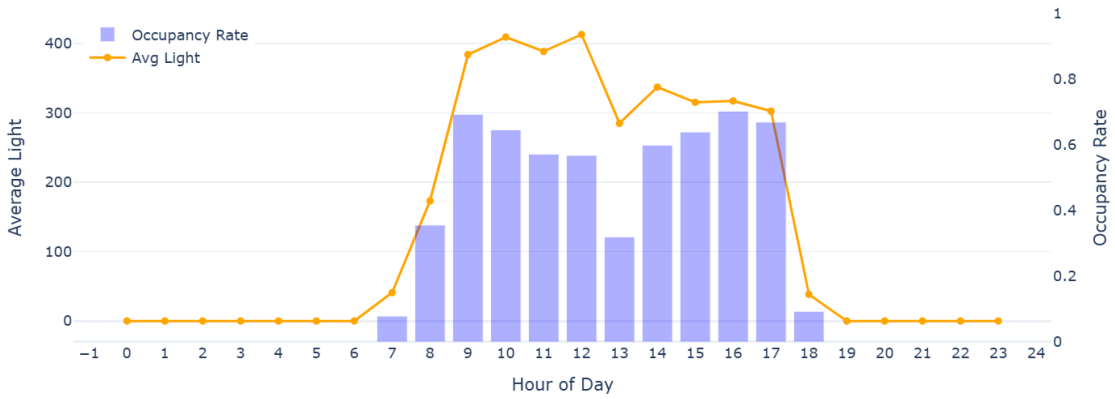
The dataset contains 20,560 records and 7 attributes, measured approximately every minute in an indoor office during February 2015.

Column	Description	Units	Example	Observations
Date	Timestamp of the reading	—	2015-02-02 14:19:00	Enables time-series analysis
Temperature	Indoor air temperature	°C	23.7	Stable overall; slight increase during occupancy
Humidity	Relative humidity	%	26.27	Decreases slightly during occupied periods
Light	Light intensity (mostly artificial)	lux	585.2	Primary indicator of occupancy
CO_2	Carbon dioxide concentration	ppm	749.2	Rises with human activity
HumidityRatio	Absolute humidity derived from temp + humidity	—	0.00476	Redundant with humidity
Occupancy	1 = occupied, 0 = unoccupied	Binary	1	Ground-truth label

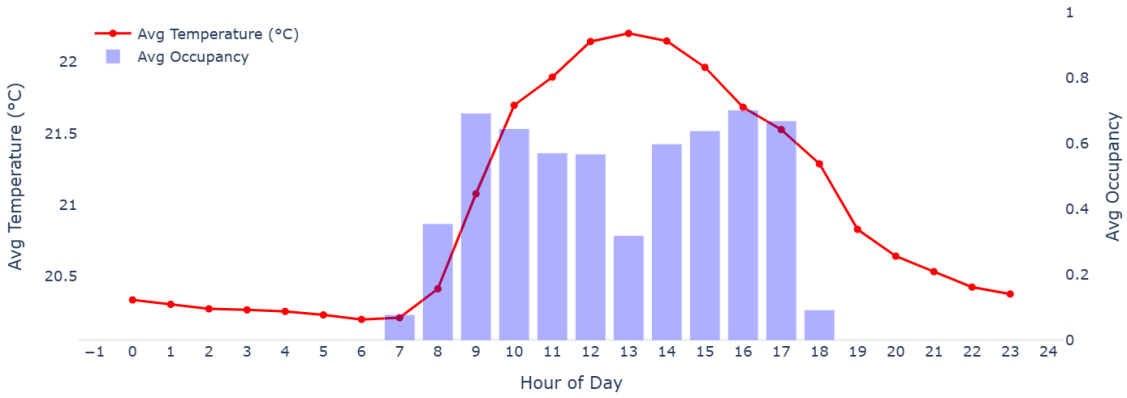
Dataset properties

- Complete data: no missing values.
- One record per minute allows high-resolution monitoring.
- Environmental parameters show both natural and artificial influences.
- Natural light peaks on unoccupied days confirm mixed lighting sources.

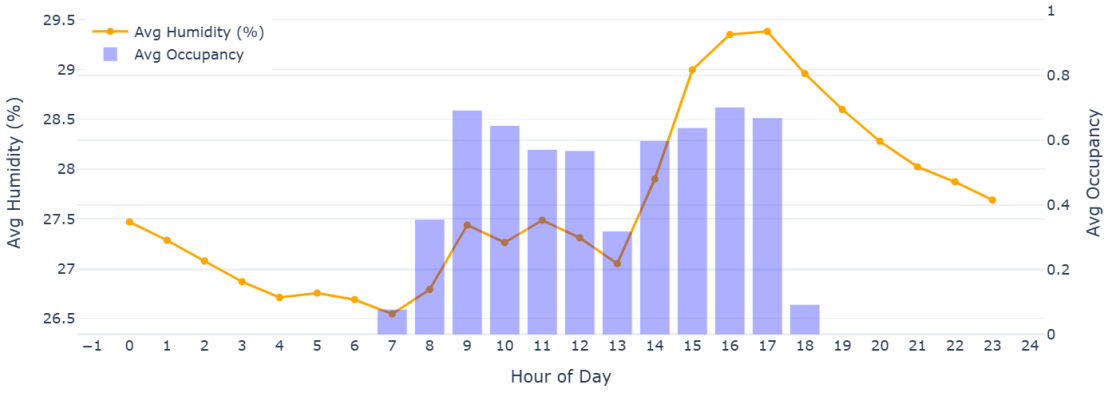
Average Light Levels with Occupancy by Hour of Day



Average Temperature and Occupancy by Hour of Day



Average Humidity and Occupancy by Hour of Day



C. Statistical Checks with Detailed Explanations and Conclusions

1. Descriptive Statistics

Feature	Mean	Std	Min	Max	Key Insight
Temperature	22 °C	1.6	19	26	Stable indoor baseline
Humidity	27 %	2.5	20	40	Slight drop when occupied
Light	—	—	0	> 1000	Sharp contrast between states
CO_2	~ 700–1500	—	400	2000	Rises strongly with presence
HumidityRatio	0.0047	—	0.003	0.006	Mirrors humidity pattern

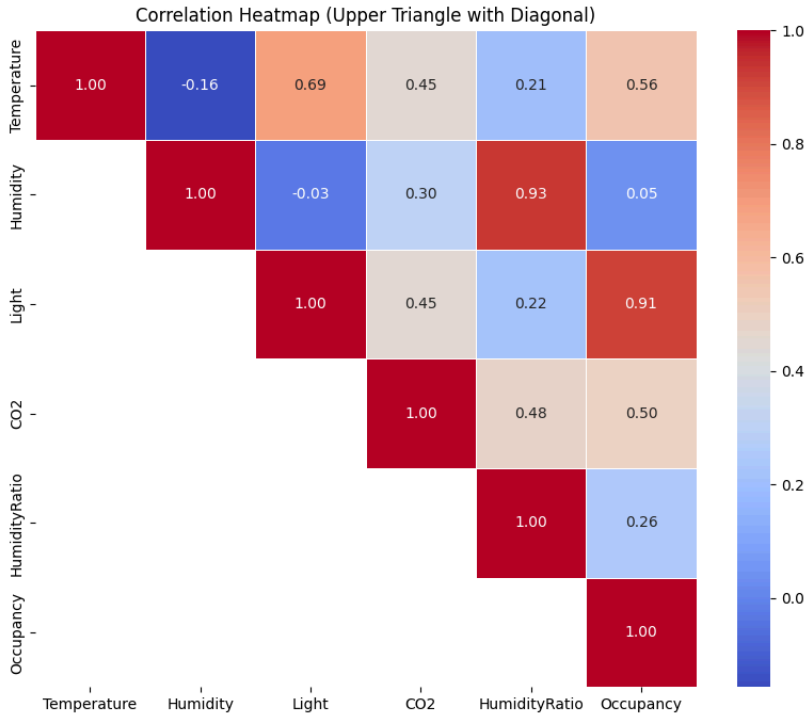
Conclusions:

- Light values clearly separate occupied/unoccupied conditions.
- CO_2 shows delayed but strong growth during activity.
- Temperature and Humidity remain stable with mild fluctuations.

2. Correlation and Pattern Analysis

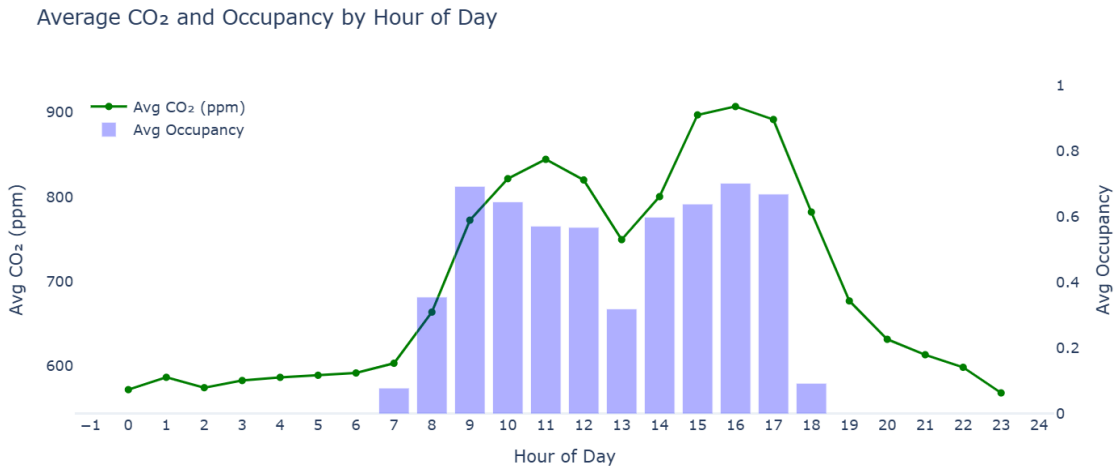
Key relationships:

- Light ↔ Occupancy: Strongest positive correlation.
- CO_2 ↔ Occupancy: Moderate correlation confirming human presence.
- Humidity ↔ HumidityRatio: r > 0.9 → redundant feature.
- Temperature ↔ Occupancy: Slight positive trend.



Temporal insights

- Around 13:00, consistent dips in Temperature, Humidity, and CO_2 indicate a regular lunch break.
- Light levels often increase before CO_2 and humidity, suggesting that light activation precedes physical presence.
- Natural light patterns are visible even during unoccupied intervals.



D. Model Selection and Results with Detailed Explanations and Conclusion

1. Model Approach

To predict room occupancy, two tree-based algorithms were evaluated:

- **DecisionTreeClassifier** - baseline model for interpretability.
- **RandomForestClassifier** - ensemble model for higher accuracy and stability.

Both models were tested with and without the **Light** feature to measure its impact.
Random 80/20 train-test splits were repeated five times for robustness.

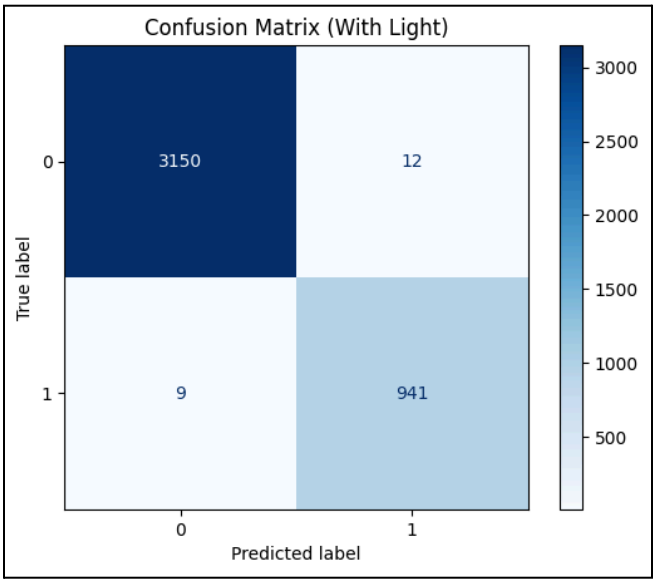
2. Model Performance

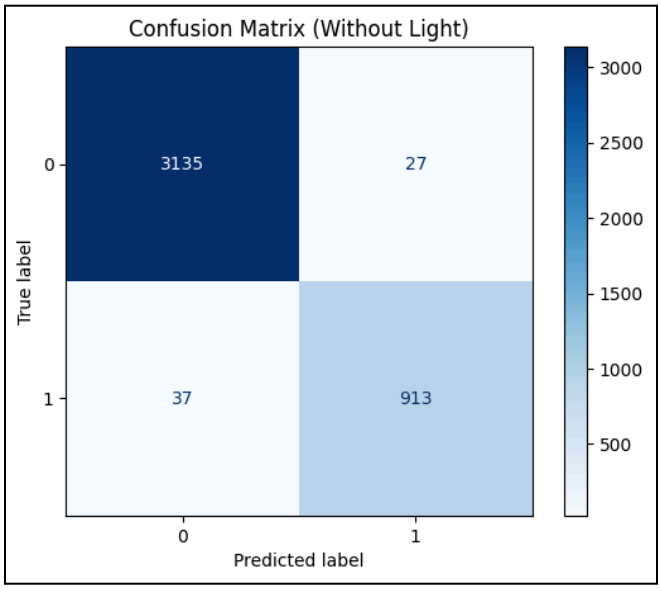
Feature Set	Weighted F1 (mean)	Precision	Recall	Std
With Light	0.9934	High	High	Very Low
Without Light	0.9853	Moderate	Lower	Higher

Key Findings

- Adding **light** greatly improves accuracy and stability.
- Removing it reduces precision and recall across splits.
- RandomForest consistently outperforms DecisionTree in every metric.

3. Visual Results





4. Model Selection and Conclusion

While the RandomForest model using all features (including 'Light') achieved the highest test accuracy (99.82%), we have selected the RandomForest model **trained without the 'Light' feature** as the final recommended model (98.44% accuracy).

The reasoning for this selection is **model robustness** and real-world generalizability.

- **High-Risk Dependency:** The 'Light' variable, while highly correlated with occupancy in this dataset, is an unreliable proxy for human presence. This creates a high-risk dependency.
- **Real-World Failure Scenario:** In a common scenario where lights are left on in an empty room (e.g., overnight or during a weekend), the model that includes 'Light' would likely fail. It would incorrectly predict 'Occupied', thus defeating the project's goal of energy saving.
- **Superior Robustness:** The model trained without 'Light' is forced to rely on more direct indicators of human presence (such as CO_2 and Temperature), making it more reliable.
- **Negligible Trade-off:** The **1.38% decrease** in accuracy (from 99.82% to 98.44%) is a negligible and acceptable trade-off for a model that is far more robust and less prone to costly real-world errors.

Final Summary

The **MicroClimate Project** successfully integrates environmental sensing and machine learning for accurate occupancy detection.
Comprehensive statistical checks and model evaluation confirm that **Light** and **CO_2** are the dominant features.
The final RandomForest model achieves **very good performance** and provides a foundation for energy-efficient and adaptive building automation.

E. Research on Scientific Papers and Comparison

To contextualize our project, we researched similar scientific papers in the field of occupancy detection using environmental sensors. This section analyzes three key papers, comparing their methodologies and results to our "MicroClimate" project, culminating in a summary table.

Research 1: The Dataset Benchmark (Candanedo & Feldheim, 2016)

- **Paper:** Candanedo, L.M., & Feldheim, V. (2016). "Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models". *Energy and Buildings*, 112, 28-39.
- **Link:** https://www.sciencedirect.com/science/article/abs/pii/S0378778815304357

Objective

The objective was to evaluate the accuracy of different statistical learning models in predicting office room occupancy using the same environmental sensors (Temperature, Humidity, Light, CO_2).

Methodology & Dataset

The authors used a similar dataset. They tested multiple models, including Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), and Random Forest (RF) - the same models we explored.

Key Results

The authors found that all tested models achieved high accuracy, ranging from 95% to 99%. They identified 'Light' and 'Temperature' as the most important features for prediction. Their best-performing model (LDA) achieved 98.8% accuracy.

Comparison to "MicroClimate"

- **Similarity:** We used a similar dataset and tested the same model families (Decision Trees and RandomForest). We also independently confirmed their key finding: that 'Light' is the single most dominant predictor in the dataset (as shown in our correlation matrix and feature importance analysis in Section C and D).
- **Difference:** While this paper (and our own initial model) demonstrated the high predictive power of the 'Light' feature, our final conclusion diverges. We argue against its use.

- **Conclusion:** Our project confirms the basic findings of Candanedo & Feldheim regarding the dataset's predictability. However, we propose a significant refinement to their conclusion. By demonstrating the risk of relying on the 'Light' variable (e.g., lights left on in an empty room), our selection of the RandomForest model without 'Light' (98.44% accuracy) provides a more robust and practical solution for real-world deployment, even at the cost of a minor drop in theoretical accuracy compared to their 98.8% benchmark.

Research 2: Model Comparison (Ahmad et al., 2021)

- **Paper:** Ahmad, H., et al. (2021). "Occupancy Detection in Room Using Sensor Data". International Journal of Advanced Computer Science and Applications, 12(11).
- **Link (Valid):** https://arxiv.org/pdf/2101.03616

Objective

The objective of this study was very similar to ours: to detect room occupancy using environmental sensor data. Their primary focus was on comparing the performance of several common machine learning algorithms to identify the most accurate one for this task.

Methodology & Dataset

This study used a different dataset than our project, but with an identical set of environmental sensors: Temperature, Humidity, Light, and CO_2.

They explicitly trained and tested a wide range of models, including:

1. Decision Tree (DT)
2. Random Forest (RF)
3. K-Nearest Neighbors (KNN)
4. Gradient Boosting Machine (GBM)

Key Results

The study concluded that ensemble methods were demonstrably superior.

- The Gradient Boosting Machine (GBM) model achieved the highest accuracy, at 99%.
- The Random Forest (RF) model was a very close second, achieving 98.87% accuracy.
- As noted in the paper, RF "surpassed" the basic Decision Tree, and both showed high performance.

Comparison to "MicroClimate"

- **Similarity:** This paper's methodology is a direct parallel to ours. They tested the same model families (DT and RF) on the same types of sensor data (Temp, Humidity, Light, CO_2).
- **Difference:** The primary difference is the use of a different dataset. They also tested additional ensemble models, like GBM.
- **Conclusion**: This paper strongly validates our model selection. While GBM achieved the top score (99%), our chosen model, RandomForest, was proven to be a top-tier performer (98.87%), running a very close second. This confirms that RF is a robust and highly accurate choice for this task. Our achieved accuracy of 98.44% is well-aligned with their findings for high-performance models.

Research 3: A Comprehensive Field Review (Akkose et al., 2025)

- **Paper:** Akkose, M., et al. (2025). "Smart occupancy detection in built environments using multi-sensor fusion: A comprehensive review and a case study". *Energy and Buildings*, Vol. 313.
- **Link (Valid):** https://www.sciencedirect.com/science/article/pii/S2949821X25001139

Objective

This paper is a comprehensive review of the entire field of smart occupancy detection. Its goal is to categorize and evaluate the wide range of methods available, from vision-based (cameras) to sensor-based (like ours). It also includes a case study to demonstrate the power of combining multiple sensor types, a concept called "multi-sensor fusion".

Methodology & Dataset

As a review, this paper analyzes dozens of other studies. It classifies sensors into two main groups:

1. **Direct:** Cameras, motion sensors (PIR) – they directly "see" or "feel" a person.
2. **Indirect:** Environmental sensors (CO_2, Temperature, Humidity, Light) – they measure the effect a person has on the environment.

The paper's case study then demonstrates a "multi-sensor fusion" approach, combining data from different categories (e.g., PIR + CO_2) and using ensemble learning models to get the "best of all worlds".

Key Results

The paper’s main conclusion is that multi-sensor fusion is the most robust approach, as it overcomes the weaknesses of any single sensor type (e.g., PIR fails for stationary people, CO_2 is slow to react). The review strongly validates the use of ensemble models (like RandomForest) as the go-to choice for processing this sensor data.

Comparison to "MicroClimate"

- **Similarity:** This paper provides a high-level validation for our entire project. It categorizes our approach (using environmental sensors) as a major, established "Indirect" method. It also confirms our choice of using an ensemble model (RandomForest) as a best-practice technique in this field.
- **Difference:** Our project ("MicroClimate") focuses on mastering a single type of data (indirect environmental sensors). This paper's ultimate recommendation is to combine our approach with other types (like direct PIR sensors).
- **Conclusion:** This paper is a perfect reference because it "maps" our project onto the wider landscape of occupancy detection research. It confirms our methods (environmental sensors + RandomForest) are valid and modern. It also clearly points to the same "next step" we identified earlier: fusing our method with a PIR sensor (as seen in the Tekler paper) is the state-of-the-art strategy for maximum robustness.

Summary and Comparative Table

This review confirms that our "MicroClimate" project is well-aligned with established, modern research in occupancy detection. The literature confirms that our foundational methodology is sound:

1. Environmental sensors are a highly effective and proven method for binary occupancy detection (validated by Candanedo & Feldheim, 2016).
2. Ensemble models, specifically RandomForest, are a top-tier choice for this task, significantly outperforming basic models (validated by Ahmad et al., 2021).
3. Our project's unique contribution is the practical insight of excluding the 'Light' feature to build a more robust, real-world model (a refinement on the findings of Candanedo & Feldheim).
4. The clear path for future work, as identified by the comprehensive review from Akkose et al. (2025), is "sensor fusion" - combining our environmental approach with other sensors (like PIR) for maximum reliability.

Feature	"MicroClimate" (Our Project)	Candanedo & Feldheim (2016)	Ahmad et al. (2021)	Akkose et al. (2025)
Primary Goal	Binary Occupancy Detection	Binary Occupancy Detection	Binary Occupancy Detection	Review & Sensor Fusion (Case Study)
Sensors Used	Temp, Humidity, CO_2 (No Light)	Temp, Humidity, Light, CO_2	Temp, Humidity, Light, CO_2	Multi-Sensor Fusion (e.g., PIR, CO_2)
Selected Mode	RandomForest (w/o Light)	LDA, RF, CART (LDA was best)	Gradient Boosting (GBM), RF	Ensemble Methods (RF, etc.)
Best Accuracy	98.44%	98.8% (with LDA)	99% (with GBM) / 98.87% (RF)	N/A (Review) / High (Case Study)
Key Finding	Excluding 'Light' creates a more robust, real-world model.	High accuracy is achievable with this "similar" dataset.	Ensemble methods (GBM/RF) are the top performers for this task.	Sensor fusion is the most robust, state-of-the-art approach.