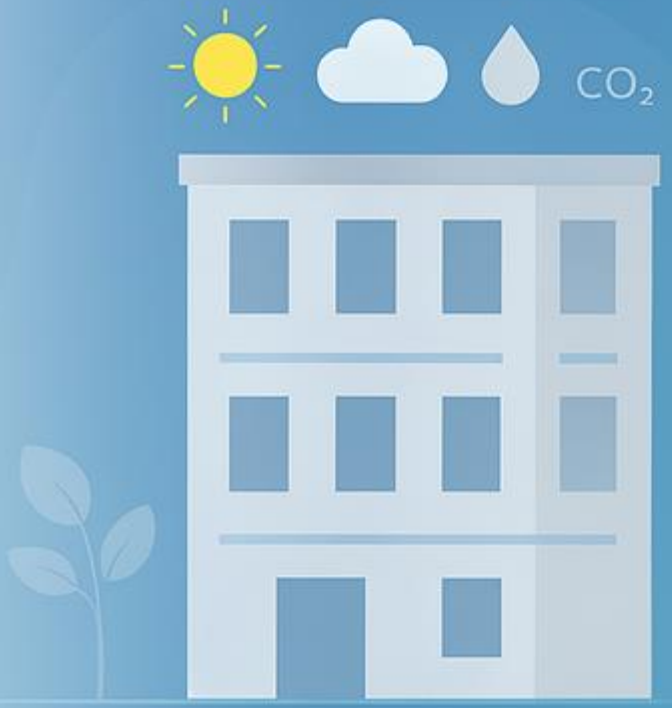
Room Occupancy Estimation

Seth Freni, Ahsan Nadeem, Uyen Nguyen, Dheeraj Kaul



Problem

- ***High energy usage:*** Temperature control systems like HVACs consume significant energy even when rooms are unoccupied.

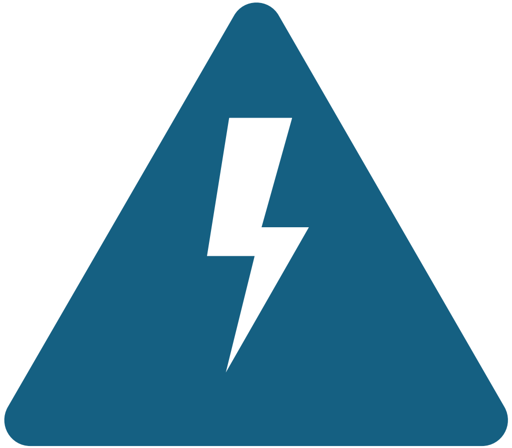
- ***Inefficient scheduling:*** Traditional systems rely on fixed schedules or manual controls, lacking adaptability to real-time occupancy.

- ***Energy wastage:*** This mismatch between system operation and actual room usage leads to unnecessary energy consumption.

- ***Need for occupancy awareness:*** Efficient operation requires systems to detect and understand the number of occupants in a room.

- ***Demand-based control:*** With real-time occupancy data, systems can adjust temperature settings dynamically.

- ***Goal***: Enable automated, energy-efficient, and demand-responsive temperature control tailored to actual room usage.



Related Work

Occupancy Detection using Environmental Features (Candanedo & Feldheim, 2016)
• Predicted room occupancy (binary) using temperature, humidity, light, CO_2, humidity ratio.
• **Models:** Logistic Regression, SVM, Random Forest.
• **Accuracy**: ~95% (Random Forest best).

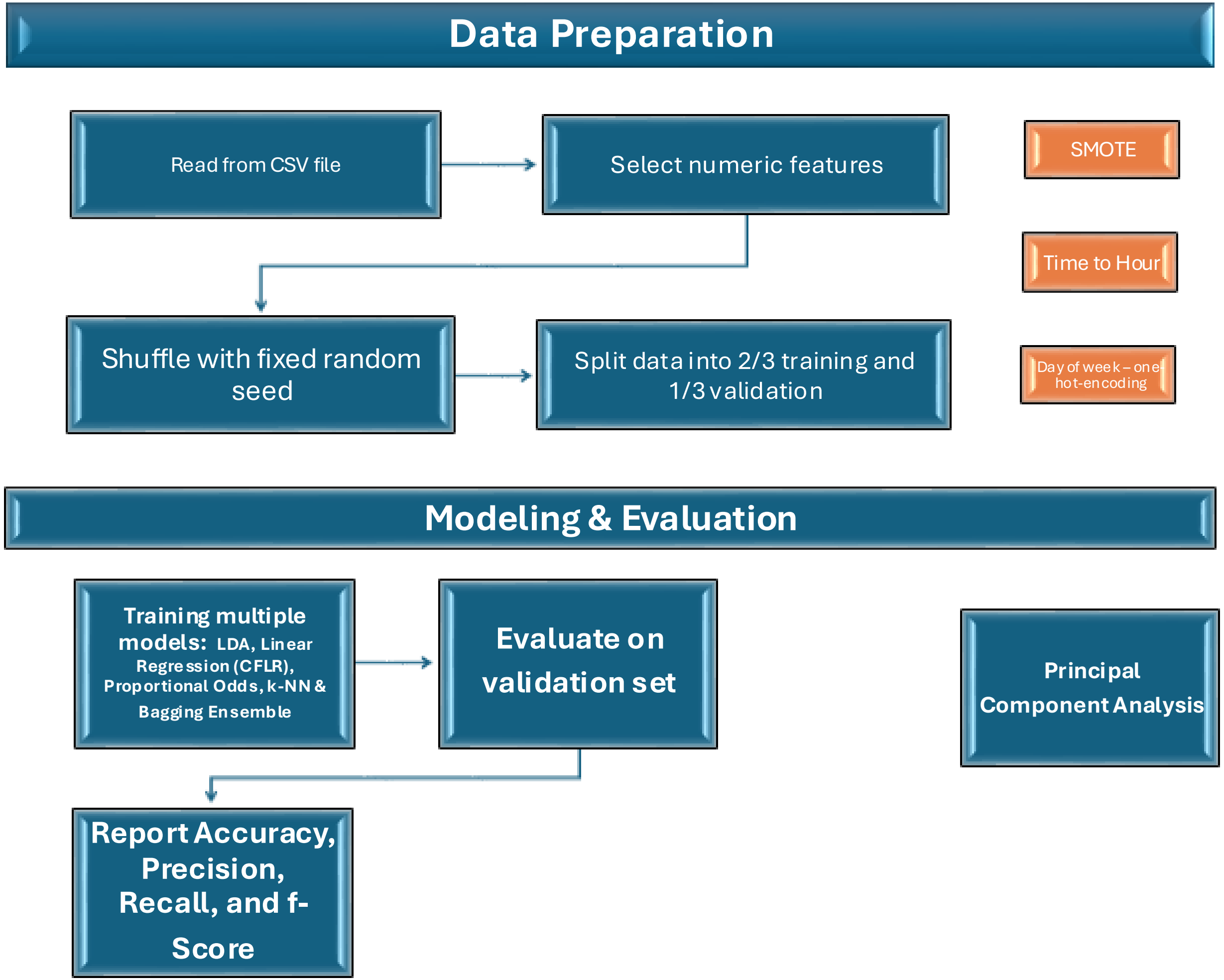
Real-Time Occupancy Estimation with Multi-sensor Fusion (Chen et al., 2018)
• Estimated occupancy counts using PIR, CO_2, temperature, sound sensors.
• **Models:** Decision Tree, k-NN, Neural Networks.
• **Accuracy:** 83–90% depending on features.

Room Occupancy Prediction (Mao et al., 2024)
• UCI Room Occupancy Estimation (0-3 occupants, 18-19 environmental sensors).
• **Model:** Logistic Regression, LDA, MSVM, MLP, LightGBM, XGBoost, Random Forest.
• **Accuracy:** Random Forrest achieved the best performance (Weighted F1 ~ 0.9985, AUC ~ 0.99996, Balanced Accuracy ~ 0.995).

Data

Feature	Meaning and Values
Date	Data of the measurement
Time	Time of the measurement
S1_Temp – S4_Temp	Temperature readings from 4 sensors
S1_Light – S4_Light	Light intensity readings from 4 sensors
S1_Sound – S4_Sound	Sound level readings from 4 sensors
S5_CO2	CO_2 concentration level
S5_CO2_Slope	Rate of chang in CO_2 concentration
S6_PIR, S7_PIR	Passive Infrared (motion) sensor values
Room_Occupancy_Count	Target variable: occupancy count {0, 1, 2, 3}

Basic Approach



Data Preprocessing

- Converted strings to floats
- Ignored the date and time features (results worse with these features)
    - Attempted to convert date to **one-hot-encoded** day of the week
    - Attempted to convert time to hour of the day (24-hour scale)
- Did not omit any rows (all rows complete)
- SMOTE (minimally (k-NN) or worsened the metrics)
- Shuffled and separated data (2/3rd for Training and 1/3rd for Validation)
- Synthetically generated a sample data set using AI to cross-validate results

Validation dataset Results for each model and Hyper-parameters

Model	Precision	Recall	f- Measure	Accuracy
LDA	0.9562	0.9715	0.9638	98.64%
CFLR	0.8071	0.8520	0.8289	94.73%
PO	0.7163	0.8409	0.7736	87.06%
k-NN	0.9759	0.9797	0.9778	99.32%

Closed Form Linear Regression specific:

Metric	Value
r-squared	0.9102
RMSE	0.2638
SMAPE	4.9392

K-NN Configuration

Metric	Value
k-Value	7
Weighing	Distance
Distance Metric	Euclidean

SMOTE results

Model	Precision	Recall	f- Measure	Accuracy
LDA	0.9547	0.9542	0.9544	95.42% ⇩
CFLR	0.8269	0.8157	0.8213	81.60% ⬇
PO	0.8572	0.8527	0.8549	85.28% ⇩
KNN	0.9991	0.9991	0.9991	99.91% ⇧

Closed Form Logistic Regression specific:

Metric	Value
r-squared	0.8482
RMSE	0.4350
SMAPE	17.2533

K-NN Configuration

Metric	Value
k-Value	1
Weighing	Uniform
Distance Metric	Euclidean

Bagging Ensemble Model Flow Training Stage

Perform same preprocessing as single models

→

Create a random subset from training data

- Values may be repeated
- # of samples = size of training data

↓

Train the model

→

Repeat subset creation and training for N subsets

Bagging Ensemble Model Flow Evaluation Stage

Obtain predictions for all N models on validation data

Final classification is result of majority vote among predictions

Report accuracy, recall, and f-measure

Bagging Results

Model	Subset Count
LDA	100
CFLR	45
PO	10
KNN	55

Model	Precision	Recall	F-Measure	Accuracy
LDA	0.9593	0.9691	0.9641	98.65%
CFLR	0.7782	0.8145	0.7960	94.05%
PO	0.7104	0.8408	0.7702	86.82%
KNN	0.9764	0.9774	0.9769	99.35%

Training Results

Validation Results

Model	Precision	Recall	F-Measure	Accuracy
LDA	0.9562	0.9715	0.9638	98.64%
CFLR	0.8045	0.8503	0.8267	94.64%
PO	0.7145	0.8402	0.7723	87.50%
KNN	0.9701	0.9735	0.9718	99.14%

Evaluation

k-NN: best performance (~99% acc, high precision/recall).

LDA: very strong (~98% acc), reliable across classes.

CFLR: moderate (~94%), struggles with categorical prediction.

Proportional Odds: weakest (~87%), poor fit for sensor data.

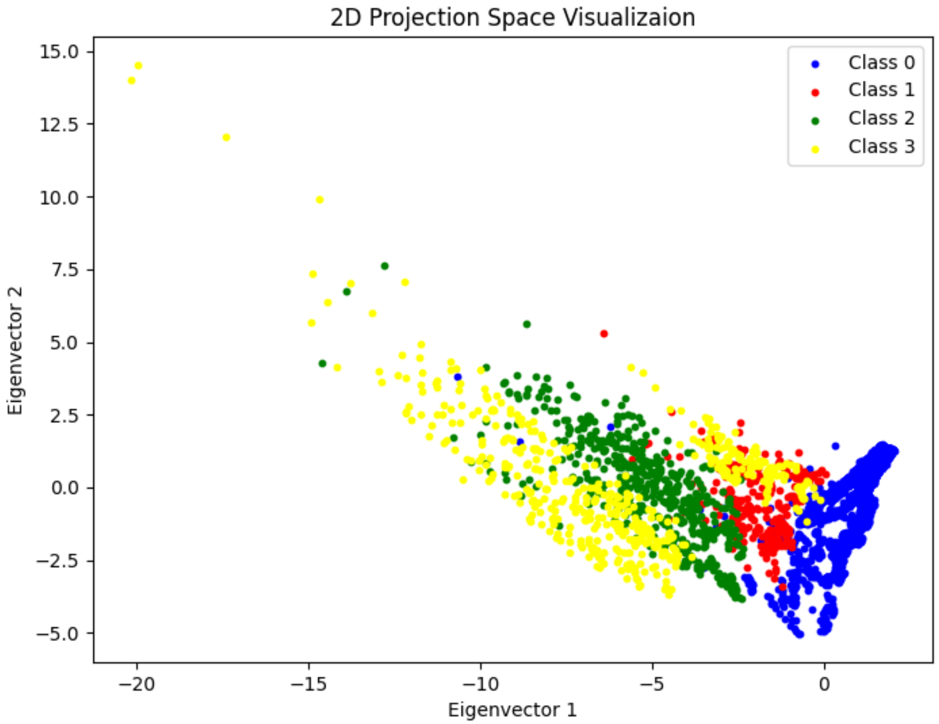
Bagging: improved PO performance, less impact on KNN/CFLR, no impact on LDA.

Class imbalance: empty vs. occupied well-separated; some overlap in counts 1–3.

SMOTE: minimal improvement, only slight gain for k-NN.

PCA Plot

- PCA reduced 19 sensor features into 2 main components for visualization.
- Class 0 (blue): clear, tight cluster, easy to detect
- Classes 1–3 (red, green, yellow): overlap, harder to separate
- Confirms evaluation: empty vs. occupied is clear, counts 1–3 less distinct
- Shows sensor features capture useful variance.



Conclusion

- Tested models: Proportional Odds, Linear Regression, LDA, k-NN, Bagging
- **k-NN with Bagging** best (~99% accuracy, very robust)
- **LDA** also strong (~95%), PO and LR weaker
- Sensor features (Temp, Light, Sound, CO_2, PIR) effective for predicting occupancy
- PCA supports results: empty vs. occupied clearly separated, overlap among counts 1–3
- Overall, system is feasible for real-time occupancy estimation
- Applications in smart buildings and energy optimization

Future Work

Expanding the dataset for practical use in real offices with larger groups of people, rather than just small spaces.

Explore more advanced models (deep learning, etc.)

Real-world deployment and scalability features

Look into more preprocessing (what things?)

Use datasets which have consistent data types which greatly eases preprocessing

Model-specific preprocessing

Bibliography

- Adarsh Pal Singh, Vivek Jain, Sachin Chaudhari, Frank Alexander Kraemer, Stefan Werner and Vishal Garg, "Machine Learning-Based Occupancy Estimation Using Multivariate Sensor Nodes," in 2018 IEEE Globecom Workshops (GC Wkshps), 2018. Room Occupancy Estimation - UCI Machine Learning Repository

- Singh, A. & Chaudhari, S. (2018). Room Occupancy Estimation [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5P605.

- Chen, Z., Jiang, C., & Xie, L. (2018). Building occupancy estimation using environmental sensors and WiFi. *Energy and Buildings, 174,* 309–322. https://doi.org/10.1016/j.enbuild.2018.06.026

- Candanedo, L. M., & Feldheim, V. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. *Energy and Buildings, 112,* 28–39. https://doi.org/10.1016/j.enbuild.2015.11.071

- Mao, S., Yuan, Y., Li, Y., Wang, Z., Yao, Y., & Kang, Y. (2024). Room occupancy prediction: Exploring the power of machine learning and temporal insights. *American Journal of Applied Mathematics and Statistics, 12*(1), 1–9. https://doi.org/10.12691/ajams-12-1-1