Using Sensor Data and ML to Estimate Room Occupancy

Jeffrey Fitzpatrick

500728133

Ryerson University

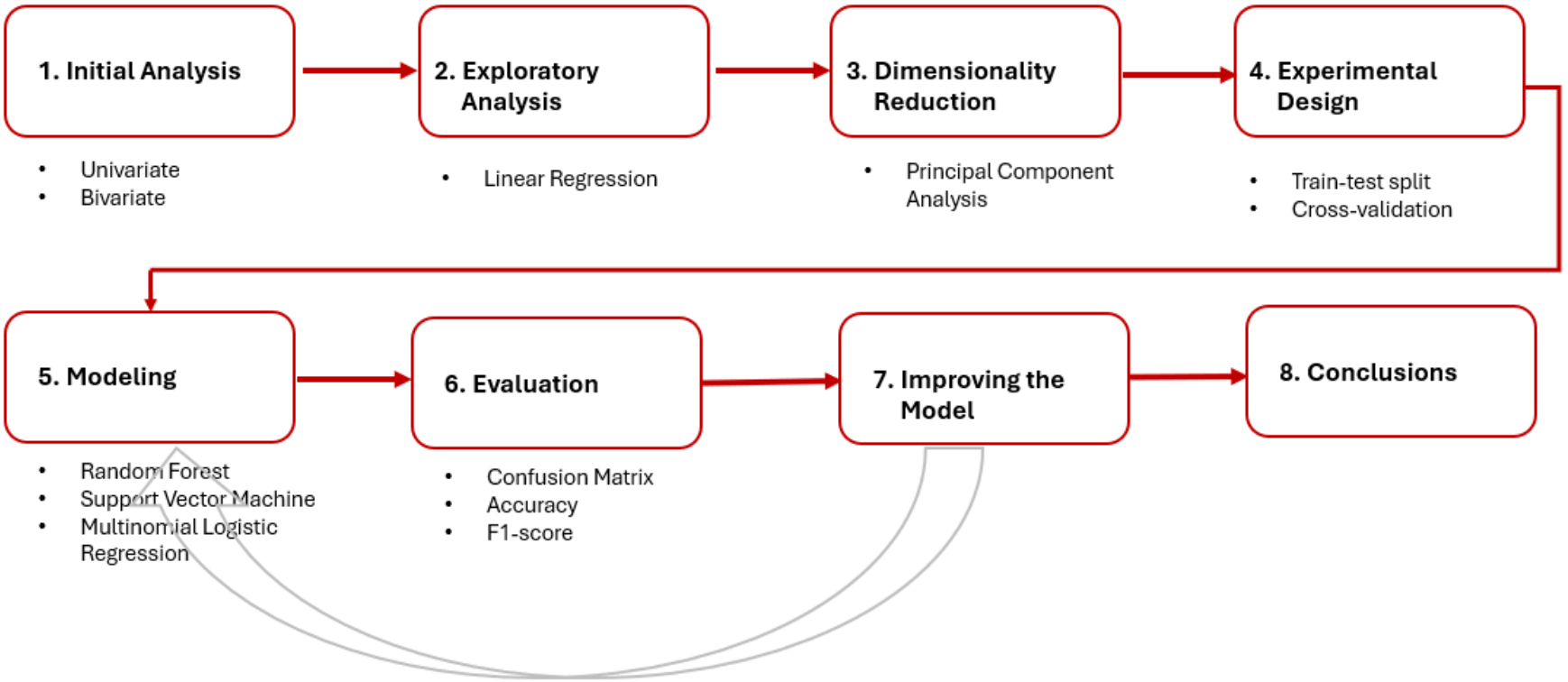
Room Occupancy Estimation dataset

- Available from the UC Irvine Machine Learning Repository
- Contains 10129 datapoints and 16 features
- Each feature represents data from a particular sensor
  - Temperature
  - Light
  - Sound
  - Motion
  - CO2
- Measurements were taken over several days in 30 second increments
- No missing values in the dataset

Research Questions

- For my project I chose the theme of Classification for building predictive models
- To reduce energy consumption in buildings, I investigated these research questions:
  - Which of the implemented supervised learning techniques perform the best in predicting occupancy?
  - Which types of sensor data (temperature, light, sound, motion, CO2) show the most promising results?
  - Based on the research, what alternative types of sensor data could be used for ML-based occupancy estimation?
- As part of my analysis, I compared my results with the results of earlier studies on the same dataset (Singh et al., 2018 and Mao et al., 2023)

Applied Methodology and Study Design

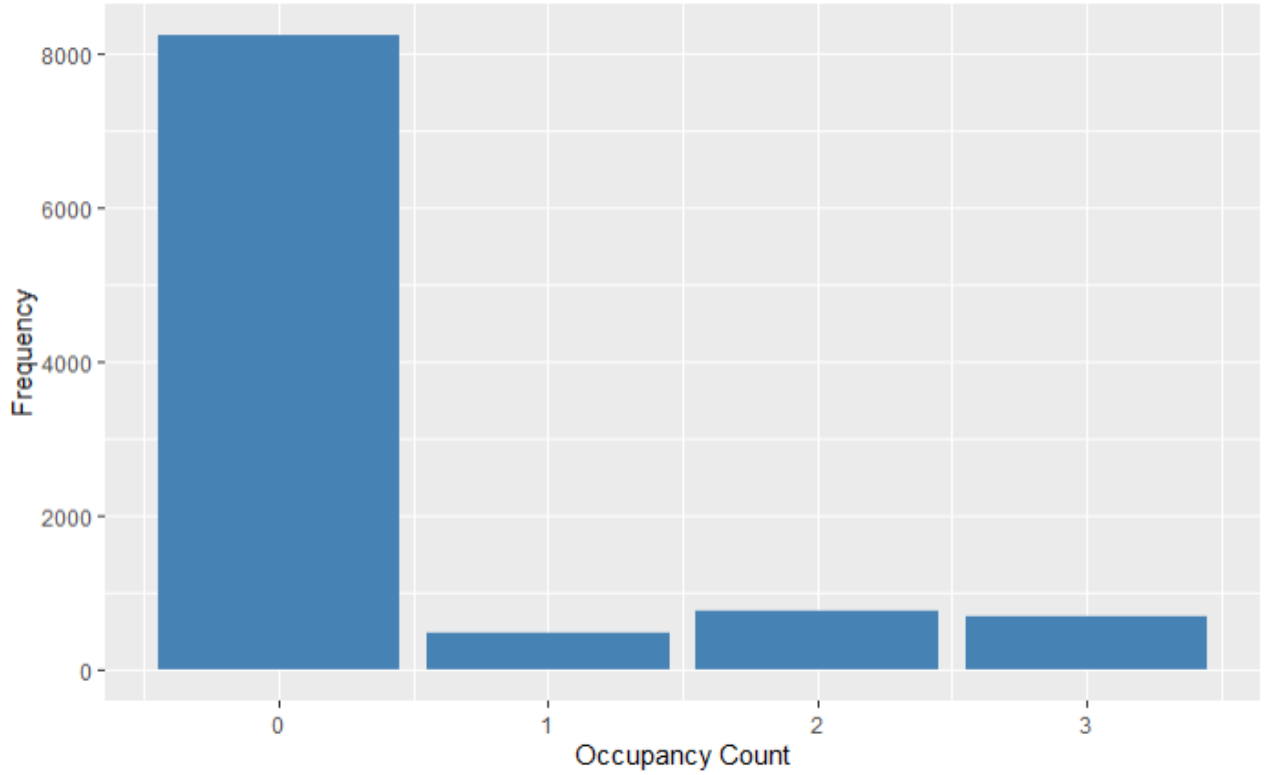


Initial Analysis: Data Dictionary

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, S2_Temp, S3_Temp, S4, Temp	Continuous	Temperature reading in degrees Celsius
S1_Light, S2_Light, S3_Light, S4_Light	Integer	Light reading in lux
S1_Sound, S2_Sound, S3_Sound, S4_Sound	Continuous	Sound reading in volts (amplifier output read by ADC)
S5_CO2	Integer	CO2 reading in ppm
S5_CO2_Slope	Continuous	Derived slope of C02 values
S6_PIR, S7_PIR	Binary	Binary value conveying motion detection
Room_Occupancy_Count	Integer	Number of occupants in room (0 to 3)

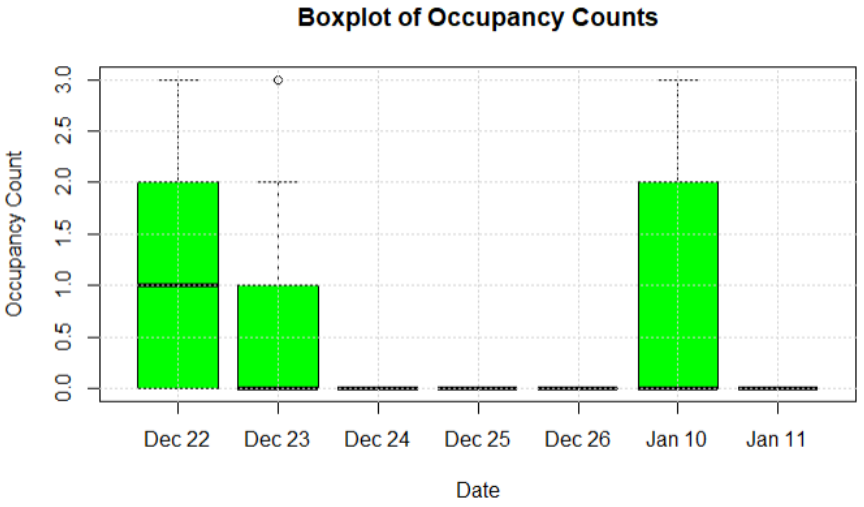
Initial Analysis: Frequency of target variable

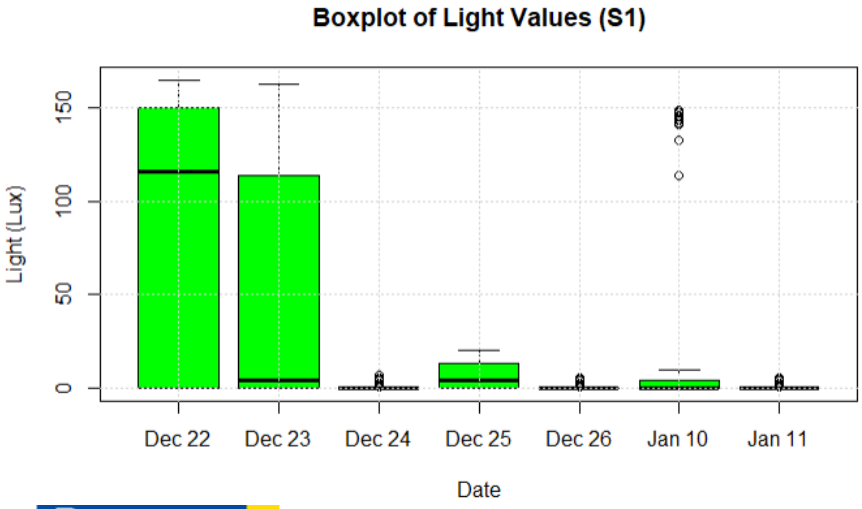
- In about 80% of the data points, the room was unoccupied

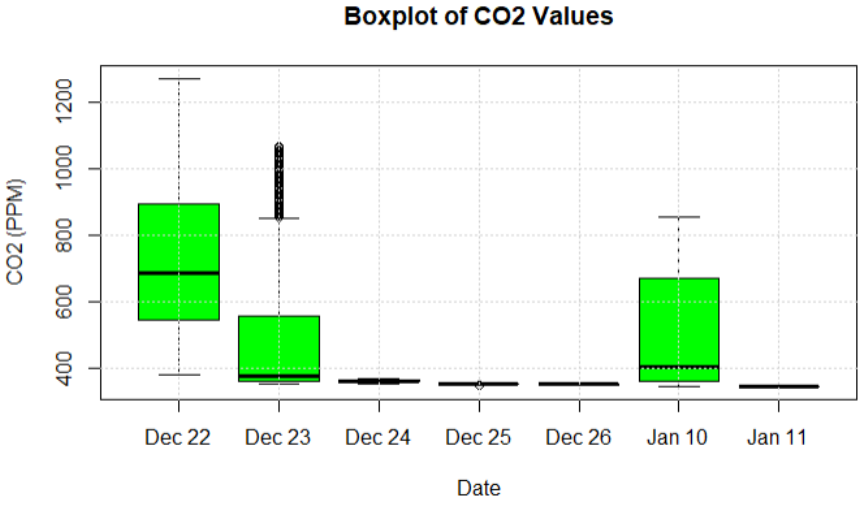


Initial Analysis: Time-Series Analysis

- Occupants on Dec. 22, Dec. 23, and Jan. 10
- Highest sensor values occurred on days with occupants



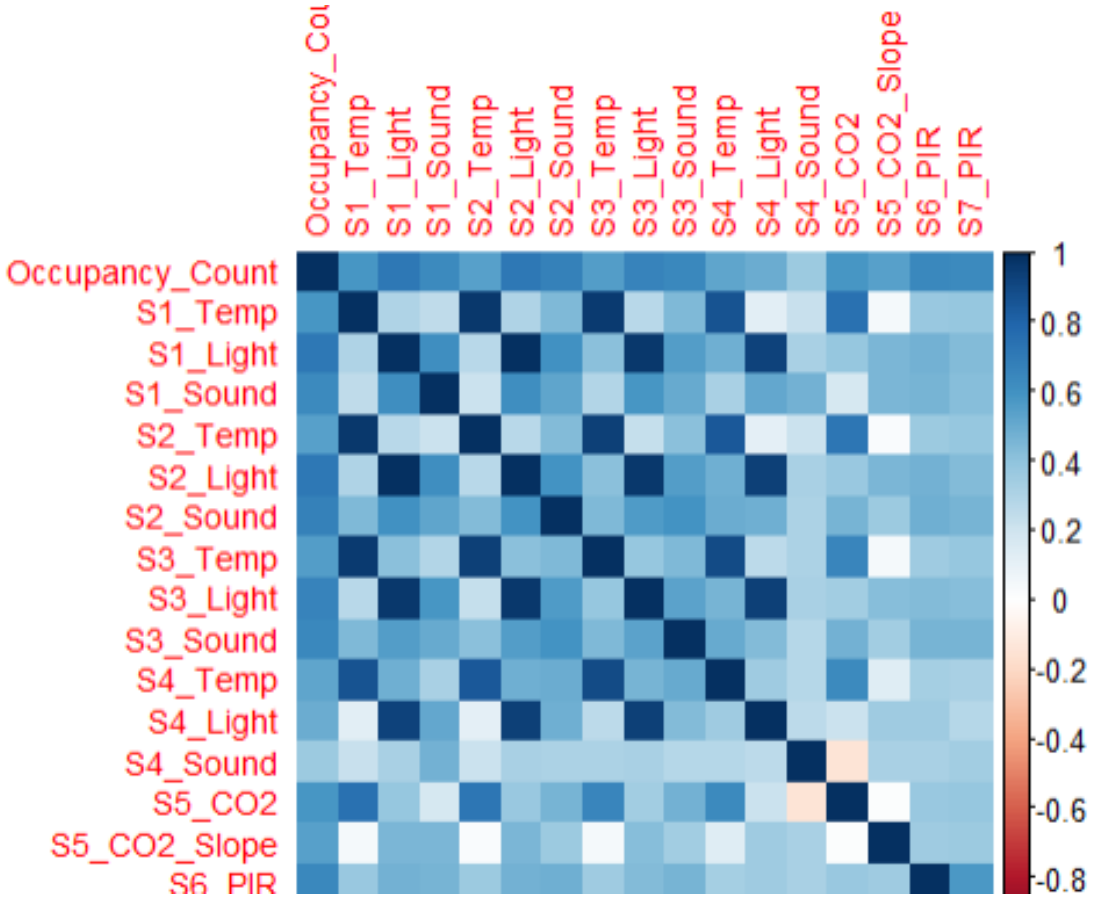




Ryerson University

Initial Analysis: Correlation Analysis

- Moderate correlation between sensor values and room occupancy
- Strong correlation of light values between sensors
- Relatively weak correlation of sound values between sensors



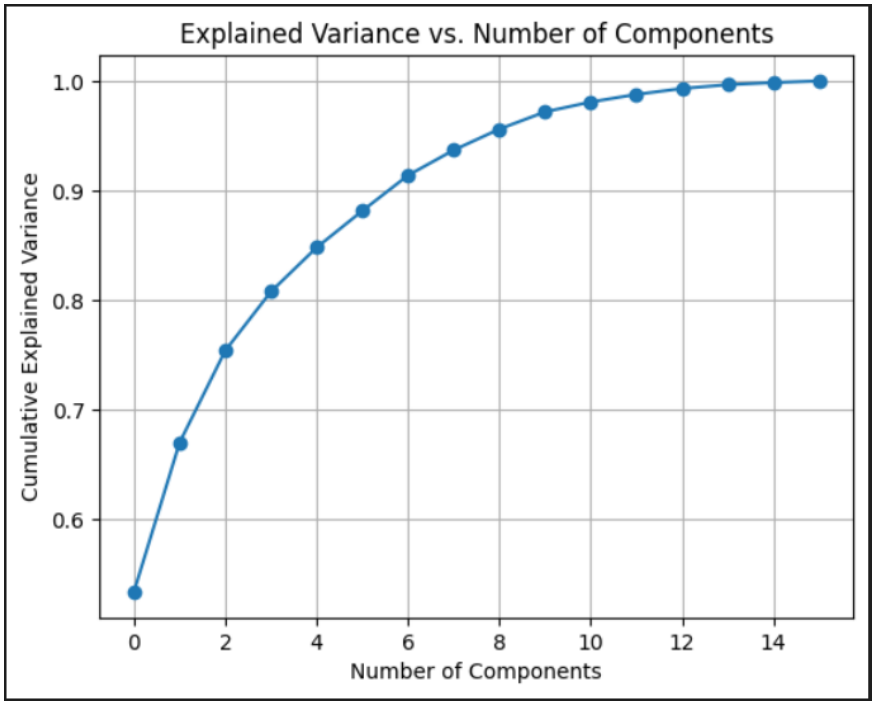
Exploratory Analysis: Linear Regression

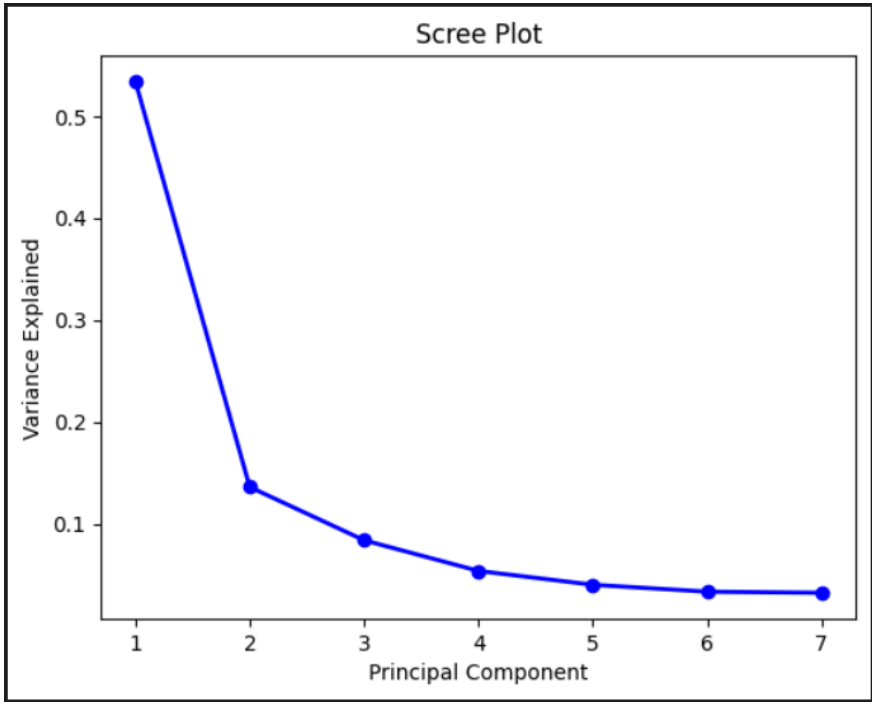
- Using all sensor data provided the best model fit (89.4%)
- Based on the remaining models, light data had the best fit (79.2%) followed by CO2 data (74.6%)
- Consistent with the introductory paper (Singh et al., 2018) which found light features performed the best overall
- Sound data (44.4%) and motion data (40.1%) had the worst fit

Model to predict room occupancy	R-squared
All features	0.894
Light data (S1_Light, S2_Light, S3_Light, S4_Light)	0.792
CO2 data (S5_CO2, S5_CO2_Slope)	0.746
Sound data (S1_Sound, S2_Sound, S3_Sound, S4_Sound)	0.444
Motion data (S6_PIR, S7_PIR)	0.401

Dimensionality Reduction: PCA

- 7 components are needed to capture at least 90% of the explained variance
- Highest contributions (first component): S1_Light, S3_Light, S1_Temp, S2_Temp, and S3_Temp





Experimental Design

- Scaling of numerical features for Logistic Regression and Support Vector Machine (mean of approximately 0 and standard deviation approximately 1)
- Split the data into training (80%) and testing (20%) sets
- Undersampled from the majority class to rebalance the training set
  - 0: 31.6%, 1: 29.9%, 2: 19.2%, 3: 19.2%
- Applied 10-fold cross validation on the training set to test stability
  - Logistic Regression: Average accuracy of 97.9%
  - Random Forest: Average accuracy of 99.0%
  - Support Vector Machine: Average accuracy of 97.8%
- Evaluated each model on the test set using all features
- Reran each model using PCA

Model Evaluation: Logistic Regression

All Features

Class	Accuracy	Precision	Recall	F1-score	Support
0		100.0%	99.2%	99.6%	1619
1		100.0%	100.0%	100.0%	103
2		97.0%	97.0%	97.0%	164
3		88.2%	96.4%	92.2%	140
Average	98.9%	96.3%	98.1%	97.2%	

PCA

	Accuracy	Precision	Recall	F1-score
Average	94.1%	81.1%	81.3%	81.2%

Model Evaluation: Random Forest

All Features

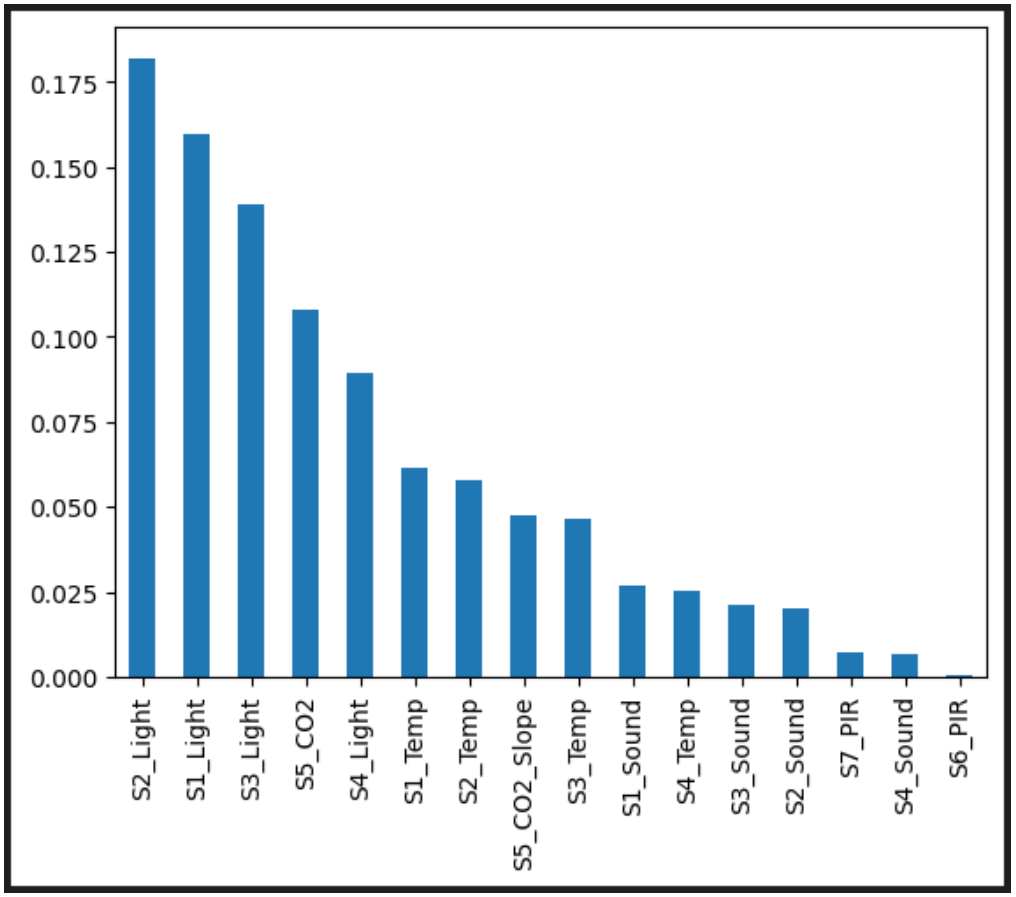
Class	Accuracy	Precision	Recall	F1-score	Support
0		100.0%	99.8%	99.9%	1619
1		99.0%	100.0%	99.5%	103
2		99.4%	99.4%	99.4%	164
3		97.9%	99.3%	98.6%	140
Average	99.8%	99.1%	99.6%	99.3%	

PCA

	Accuracy	Precision	Recall	F1-score
Average	99.0%	96.7%	96.9%	96.8%

Random Forest: Importance of each feature

- Most important features:
  - S2_Light
  - S1_Light
  - S3_Light
  - S5_CO2
  - S4_Light
- Least important features:
  - S4_Sound
  - S6_PIR



Model Evaluation: Support Vector Machine

All Features

Class	Accuracy	Precision	Recall	F1-score	Support
0		100.0%	99.4%	99.7%	1619
1		100.0%	100.0%	100.0%	103
2		97.0%	97.0%	97.0%	164
3		90.6%	96.4%	93.4%	140
Average	99.1%	96.9%	98.2%	97.5%	

PCA

	Accuracy	Precision	Recall	F1-score
Average	99.3%	97.7%	98.0%	97.8%

Conclusions: Summary of Model Results

- All features: Random Forest performed the best in terms of accuracy and F1-score (consistent with Mao et al., 2023)
- PCA: Support Vector Machine performed the best in terms of accuracy and F1-score (Singh et al., 2018 concluded that good results could be achieved with only 4 components)

	All Features			PCA	
Model	**Accuracy Training Set**	**Accuracy Test Set**	**F1-Score Test Set**	**Accuracy Test Set**	**F1-Score Test Set**
Logistic Regression	97.9%	98.9%	97.2%	94.1%	81.2%
Random Forest	99.0%	99.8%	99.3%	99.0%	96.8%
Support Vector Machine	97.8%	99.1%	97.5%	99.3%	97.8%

Conclusions: Summary of Sensor Data

- Light features performed the best overall
- However, light relies on occupants turning on lights when they arrive and turning them off when they leave (Singh et al., 2018)
- CO2 data showed promising results
- Best approach is to use a combination of different types of sensor data including CO2
- Alternative types of sensor data include Bluetooth signals, Wi-Fi, camera images, GPS data, UWB radar, and electric meters
- Limitations such as high cost, limited range, false results, and privacy issues may prevent them from being widely adopted (Khan et al., 2024)

Limitations and Future Work

- Assumptions of linear regression and logistic regression
  - Linear relationship between independent variables and dependent variable
  - Independent variables are not correlated with each other
- In this dataset, the sensor data is correlated with each other (multicollinearity)
  - P-values and coefficients in linear regression models may not be reliable
- Outliers were detected in the temperature, light, sound, and C02 data
- Future Work
  - Hypothesis testing to determine significance of model results (Friedman)
  - Feature selection (e.g., removal of light features)
  - Using PCA with fewer features
  - Evaluation of more classification algorithms

References

- Singh, A.P., Jain, V., Chaudhari, S., Kraemer, F.A., Werner, S., & Garg, V. (2018). Machine learning-based occupancy estimation using multivariate sensor nodes. *IEEE Globecom Workshops (GC Wkshps)*, 1-6. https://doi.org/10.1109/GLOCOMW.2018.8644432
- Mao, S., Yuan, Y., Li, Y., Wang, Z., Yao, Y., & Kang, Y. (2023). Room occupancy prediction: Exploring the power of machine learning and temporal insights. *American Journal of Applied Mathematics and Statistics*. https://doi.org/10.48550/arXiv.2312.14426
- Khan, I., Zedadra, O., Guerrieri, A., & Spezzano, G. (2024). Occupancy prediction in IoT-enabled smart buildings: technologies, methods, and future directions. *Sensors, 24*(11). doi.org/10.3390/s24113276

Thank you