Machine Learning - Software Project

Component 2

Authors: Liviu-Ștefan Neacșu-Miclea, Răzvan-Gabriel Petec
Specialization: Applied Computational Intelligence
Group: 246/2

1 Data distribution

Before applying any preprocessing steps, the data distribution, as shown in Figure 1, reflects a total of 10,129 measurements. It is evident that the room occupancy distribution is highly imbalanced, with the majority of observations corresponding to 0 people in the room.

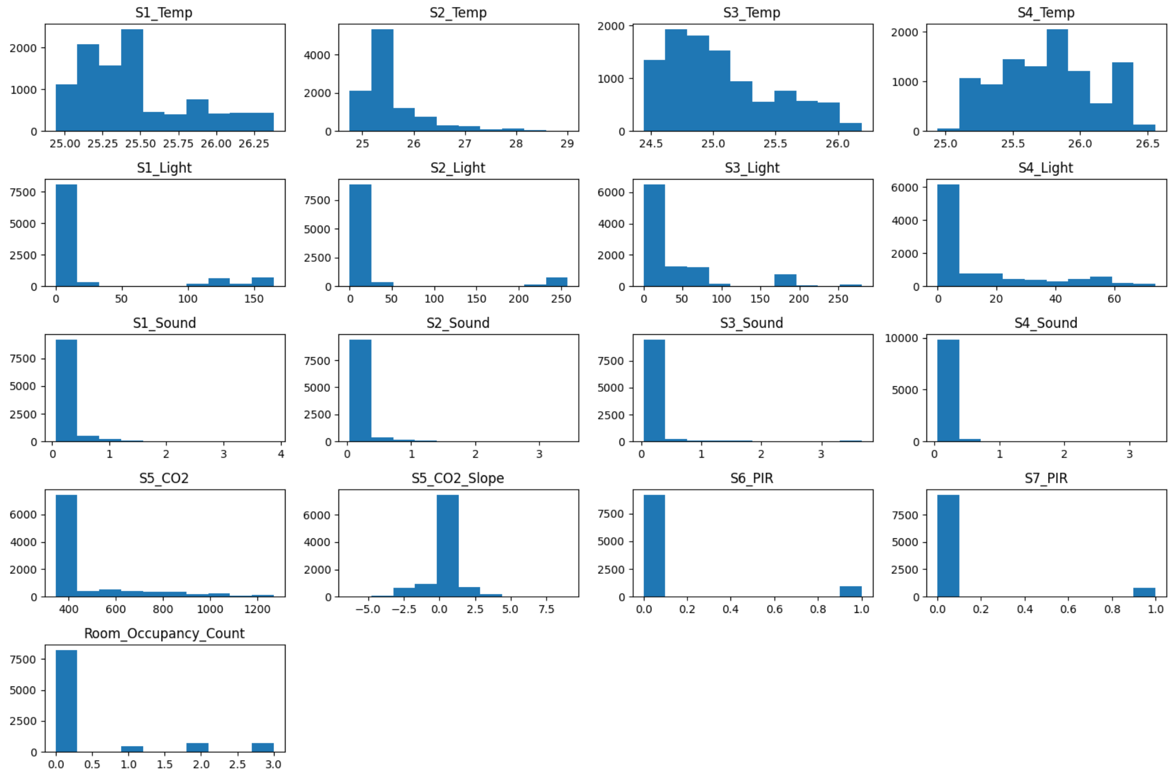


Figure 1: Data distribution before preprocessing. Two of the fields (Date, Time) were not plotted because they were not numerical data at this moment

To mitigate this, we decided to take a look at the dates where the measures were done in the dataset. We observed that there were 6 dates in total in which the measures were taken, from which 3 of them only had measures corresponding to 0 people in the room. We decided to drop the data corresponding to the measurements in the days where only data corresponding to 0 people in the room were measured, because otherwise it would most probably lead to overfitting the model. The 3 remaining dates were modelled by giving each one a label from 1 to 3.

Additionally, the ”Time” data was converted into categorical labels according to the intervals presented in Table 1.

Table 1: Mapping the time intervals to some label

Interval	Meaning	Numerical Label
00:00:00 - 05:59:59	Night	1
06:00:00 - 08:59:59	Morning	2
09:00:00 - 11:59:59	Early Morning	3
12:00:00 - 13:59:59	Noon	4
14:00:00 - 16:59:59	Afternoon	5
17:00:00 - 18:59:59	Evening	6
19:00:00 - 21:59:59	Night	7
22:00:00 - 23:59:59	Late Night	8

Following these preprocessing steps, the updated data distributions are shown in Figure 1, with 5,238 remaining measurements. While there are still some imbalances in certain columns, further data processing techniques can be applied in future steps to address these issues.

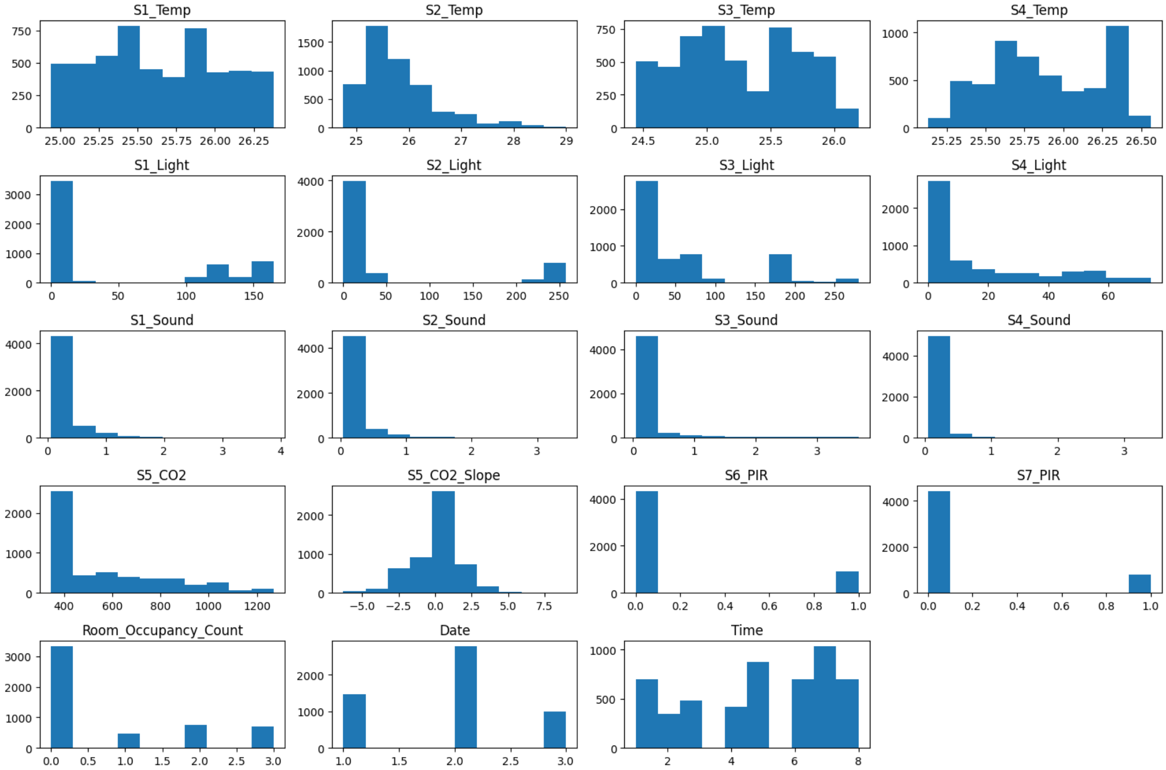


Figure 2: Data distribution after preprocessing and non-numerical fields labeling.

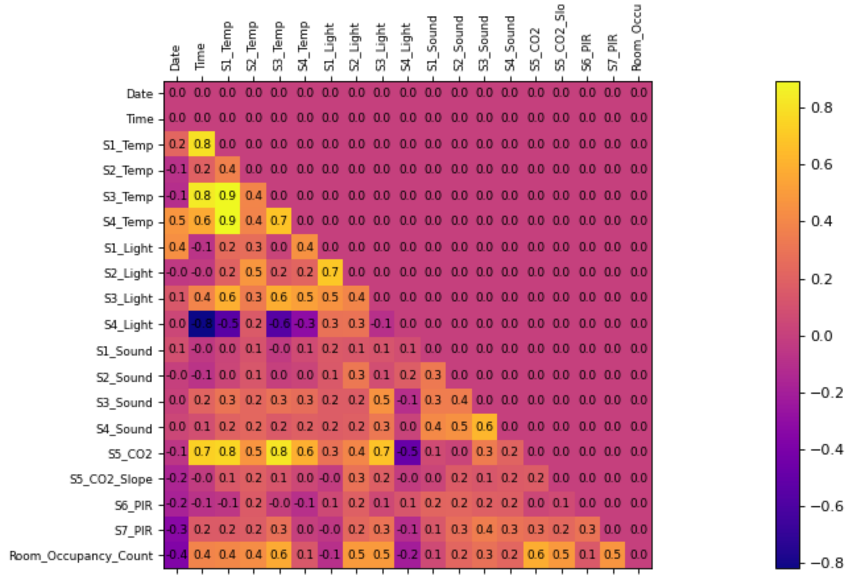
2 Feature analysis

2.1 Correlation

The Pearson correlation index is computed between each two dimensions X_i and X_j of the data:

$$\rho(X_i, X_j) = \frac{\operatorname{cov}(X_i, X_j)}{\sigma(X_i) \cdot \sigma(X_j)} = \frac{\mathbb{E}((X_i - \mu(X_i)) \cdot (X_j - \mu(X_j)))}{\sigma(X_i) \cdot \sigma(X_j)},$$

then the correlation matrix C is constructed as $c_{ij} = \rho(X_i, X_j)$ `if` $i < j$ `else` 0 — see Fig. 3.



s

Figure 3: Correlation matrix (using Pearson's [Pla83] test).

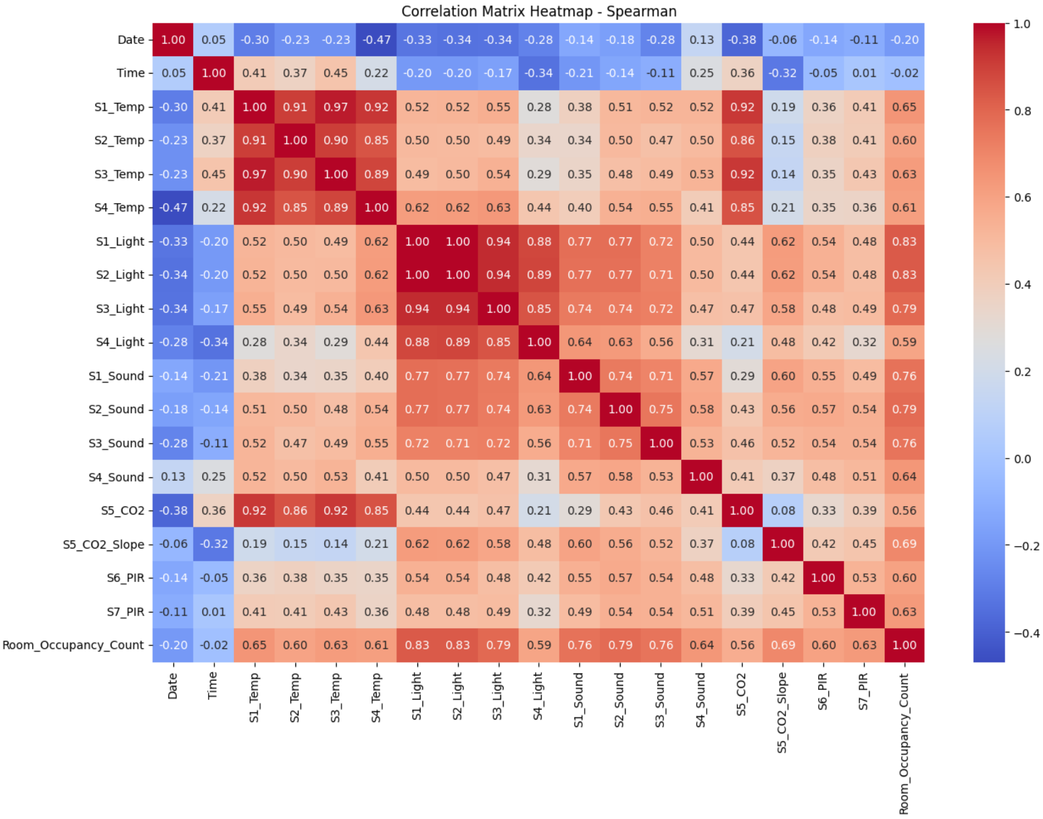


Figure 4: Correlation matrix (using Spearman's [Spe61] test).

2.2 Independence

For any two features X_i and X_j, if both are sampled from categorical distributions, the Pearson's χ^2 test [Pla83] is used to measure their independence. If at least one feature is continuous data, then the Spearman correlation [Spe61] is computed instead. Numerical results in matrix form can be consulted in Fig. 6, while Fig. 5 shows an independence matrix computed using only the χ^2 test, where each continuous observation is assigned to the class of the largest integer lower than or equal to the measured value.

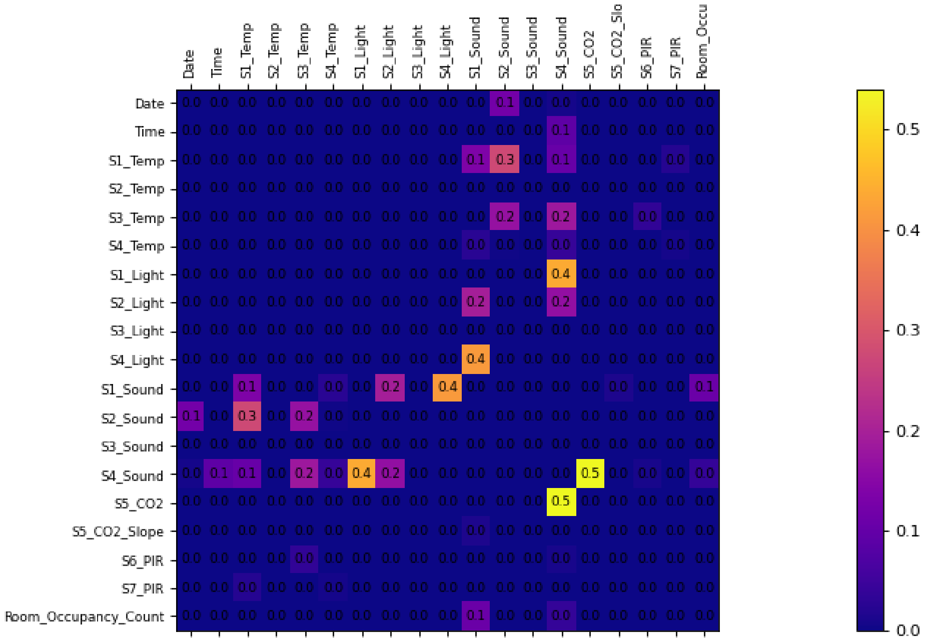
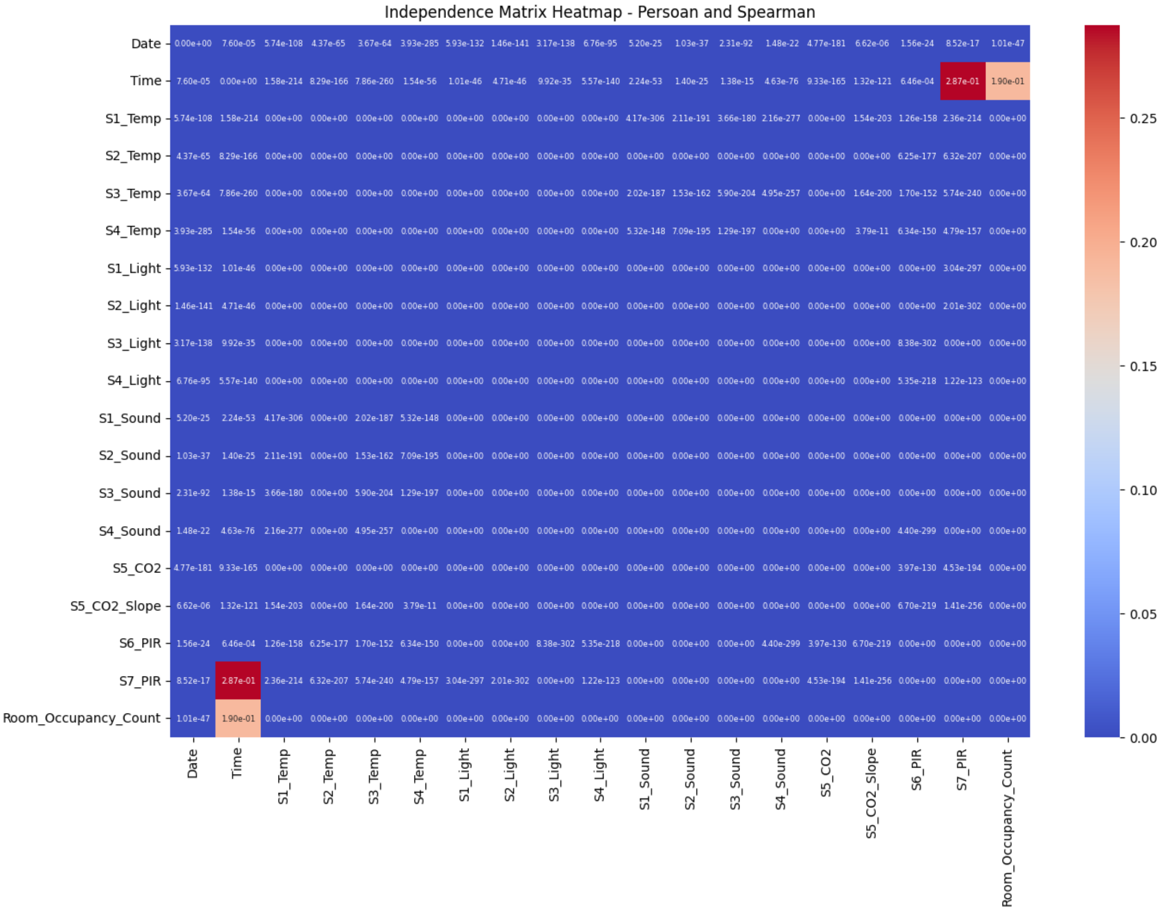


Figure 5: Independence matrix Pearson's χ^2 [Pla83] only for all the feature combinations, for continuous values, bins were selected.



	Date	Time	S1_Temp	S2_Temp	S3_Temp	S4_Temp	S1_Light	S2_Light	S3_Light	S4_Light	S1_Sound	S2_Sound	S3_Sound	S4_Sound	S5_CO2	S5_CO2_Slope	S6_PIR	S7_PIR	Room_Occupancy_Count
Date	0.00e+00	7.60e-05	5.74e-108	4.37e-65	3.67e-64	3.93e-285	5.93e-132	1.46e-141	3.17e-138	6.76e-95	5.20e-25	1.03e-37	2.31e-92	1.48e-22	4.77e-181	6.62e-06	1.56e-24	8.52e-17	1.01e-47
Time	7.60e-05	0.00e+00	1.58e-214	8.29e-166	7.86e-260	1.54e-56	1.01e-46	4.71e-46	9.92e-35	5.57e-140	2.24e-53	1.40e-25	1.38e-15	4.63e-76	9.33e-165	1.32e-121	6.46e-04	2.87e-01	1.90e-01
S1_Temp	5.74e-108	1.58e-214	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	4.17e-306	2.11e-191	3.66e-180	2.16e-277	0.00e+00	1.54e-203	1.26e-158	2.36e-214	0.00e+00
S2_Temp	4.37e-65	8.29e-166	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	6.25e-177	6.32e-207	0.00e+00
S3_Temp	3.67e-64	7.86e-260	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	2.02e-187	1.53e-162	5.90e-204	4.95e-257	0.00e+00	1.64e-200	1.70e-152	5.74e-240	0.00e+00
S4_Temp	3.93e-285	1.54e-56	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	5.32e-148	7.09e-195	1.29e-197	0.00e+00	0.00e+00	3.79e-11	6.34e-150	4.79e-157	0.00e+00
S1_Light	5.93e-132	1.01e-46	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	3.04e-297	0.00e+00
S2_Light	1.46e-141	4.71e-46	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	2.01e-302	0.00e+00
S3_Light	3.17e-138	9.92e-35	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	8.38e-302	0.00e+00	0.00e+00
S4_Light	6.76e-95	5.57e-140	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	5.35e-218	1.22e-123	0.00e+00
S1_Sound	5.20e-25	2.24e-53	4.17e-306	0.00e+00	2.02e-187	5.32e-148	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
S2_Sound	1.03e-37	1.40e-25	2.11e-191	0.00e+00	1.53e-162	7.09e-195	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
S3_Sound	2.31e-92	1.38e-15	3.66e-180	0.00e+00	5.90e-204	1.29e-197	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
S4_Sound	1.48e-22	4.63e-76	2.16e-277	0.00e+00	4.95e-257	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	4.40e-299	0.00e+00	0.00e+00
S5_CO2	4.77e-181	9.33e-165	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	3.97e-130	4.53e-194	0.00e+00
S5_CO2_Slope	6.62e-06	1.32e-121	1.54e-203	0.00e+00	1.64e-200	3.79e-11	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	6.70e-219	1.41e-256	0.00e+00
S6_PIR	1.56e-24	6.46e-04	1.26e-158	6.25e-177	1.70e-152	6.34e-150	0.00e+00	0.00e+00	8.38e-302	5.35e-218	0.00e+00	0.00e+00	0.00e+00	4.40e-299	3.97e-130	6.70e-219	0.00e+00	0.00e+00	0.00e+00
S7_PIR	8.52e-17	2.87e-01	2.36e-214	6.32e-207	5.74e-240	4.79e-157	3.04e-297	2.01e-302	0.00e+00	1.22e-123	0.00e+00	0.00e+00	0.00e+00	0.00e+00	4.53e-194	1.41e-256	0.00e+00	0.00e+00	0.00e+00
Room_Occupancy_Count	1.01e-47	1.90e-01	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00

Figure 6: Independence matrix. Pearson's [Pla83] test was used for categorical pairs of values and Spearman's [Spe61] test was used for continuous pairs of values and continous and categorical pairs of values.)

2.3 Feature importance

A linear regression algorithm is used to compute the order of attention accorded to each individual feature. The least square method (LSM) [Mil06] was used to find the parameter β that minimizes $(Y - \beta \cdot X)^2$, as follows: $\beta = (X^t X)^{-1} X_t Y$. A bar graph of the absolute value of vector β's components, which denote the weight associated with each individual feature, is shown in Fig. 7. Analogously, a logistic regression [LaV08] algorithm was trained to fit categorical$(Y) \approx$ softmax$(\beta \cdot X)$, and the feature importance according to the weights is presented in Fig. 8.

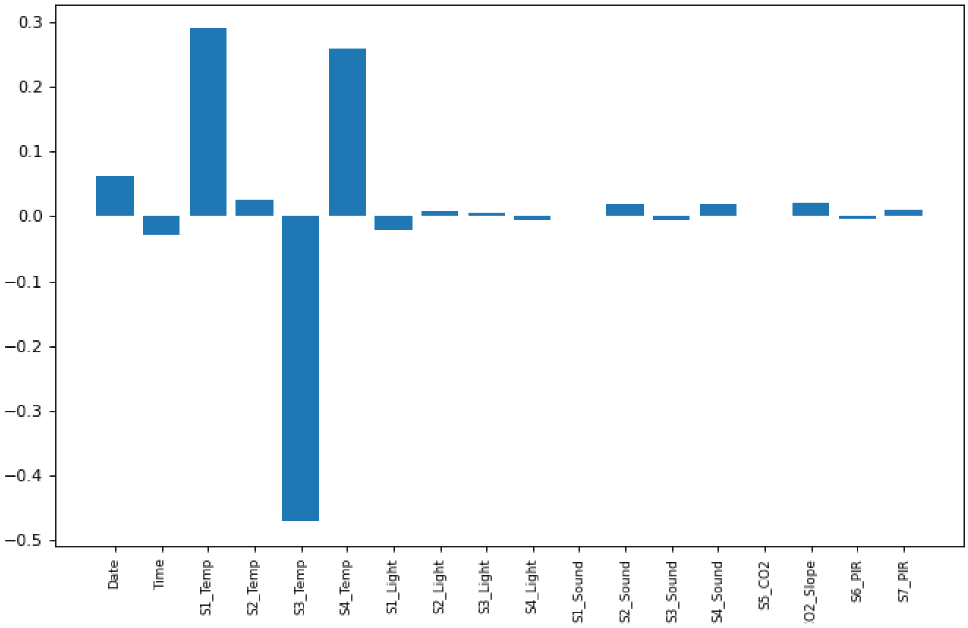


Figure 7: Feature importance using linear regression.

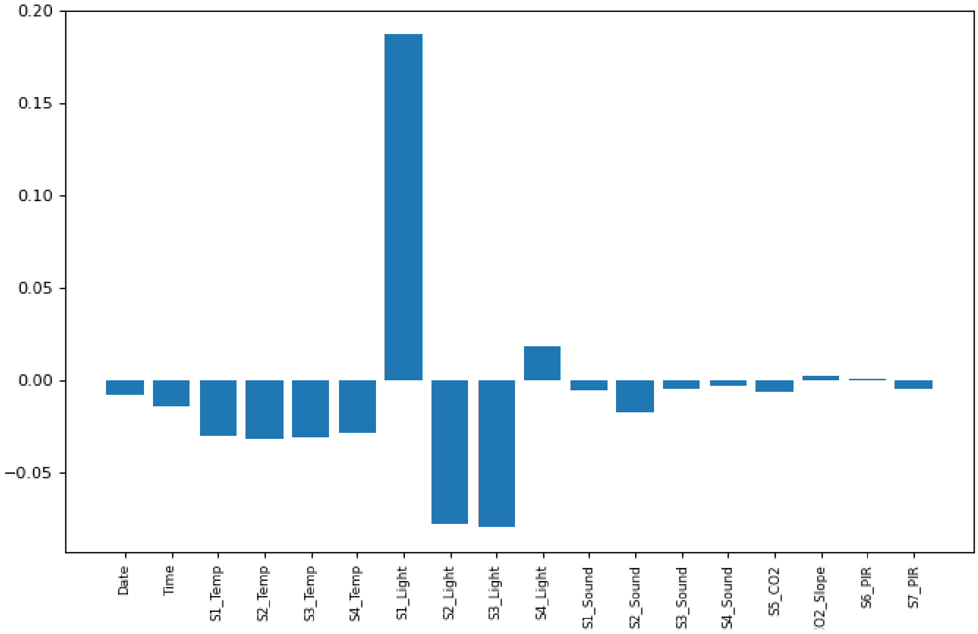


Figure 8: Feature importance using logistic regression [LaV08].

3 Data statistics

Sensor	mean	std	min	25%	50%	75%	max
S1_Temp	25.64	0.39	24.94	25.31	25.63	25.94	26.38
S2_Temp	25.81	0.71	24.75	25.31	25.63	26.13	29.00
S3_Temp	25.25	0.47	24.44	24.88	25.19	25.63	26.19
S4_Temp	25.88	0.35	25.13	25.63	25.81	26.25	26.56
S1_Light	46.54	63.92	0.00	0.00	3.00	118.00	165.00
S2_Light	47.23	88.27	0.00	0.00	4.00	25.00	258.00
S3_Light	53.26	72.95	0.00	0.00	14.00	74.00	280.00
S4_Light	16.46	21.31	0.00	0.00	7.00	29.00	74.00
S1_Sound	0.26	0.42	0.06	0.07	0.08	0.21	3.88
S2_Sound	0.19	0.36	0.04	0.05	0.05	0.13	3.44
S3_Sound	0.25	0.56	0.04	0.06	0.06	0.13	3.67
S4_Sound	0.12	0.17	0.05	0.06	0.07	0.09	3.40
S5_CO2	561.04	237.71	345.00	365.00	450.00	720.00	1270.00
S5_CO2_Slope	-0.01	1.62	-6.30	-0.53	0.00	0.63	8.98
S6_PIR	0.17	0.38	0.00	0.00	0.00	0.00	1.00
S7_PIR	0.15	0.36	0.00	0.00	0.00	0.00	1.00
Count	0.77	1.12	0.00	0.00	0.00	2.00	3.00

4 Data visualization and interpretation

From Fig. 7, it can be deduced that the sound, and temperature measurements play a greater role in estimating the number of people in the room than light and movement. On the other hand, the logistic regression (Fig. 8) accords greater average importance to light, temperature and sound than the time period or movement.

Fig. 3 reveals that the room occupancy corelates highly with light, temperature and sensor measurements and lower with the movement or sound. It also provides interesting ambiental informations (the CO2 level is higher as temperature is greater, the temperature sensors tend to highly correlate between themselves, or that CO2 and sound are loss correlated). However, the independence metrics in Fig. 5 and Fig. 6 seem to give away a high dependence relationship between the features, which may look natural since we are dealing with a real world physical system where one random variable most certainly affects another.

In Figure 9, we plotted the evolution of the means of the sensors across different time types (the ones defined in Table 1). The bottom-right plot represents the mean room occupancy count, which is the target variable to predict. Analyzing its relationship with the other sensor readings reveals clear correlations that can guide predictive modeling. Peaks in occupancy (time steps 4–6) align closely with increases in environmental indicators such as light intensity, sound levels, and CO2 concentrations, as well as motion activity detected by PIR sensors. For instance, higher CO2 levels strongly correspond to increased occupancy, likely due to human respiration. Similarly, spikes in PIR sensor activity and sound levels suggest heightened human motion and noise during periods of high occupancy, making them reliable predictors.

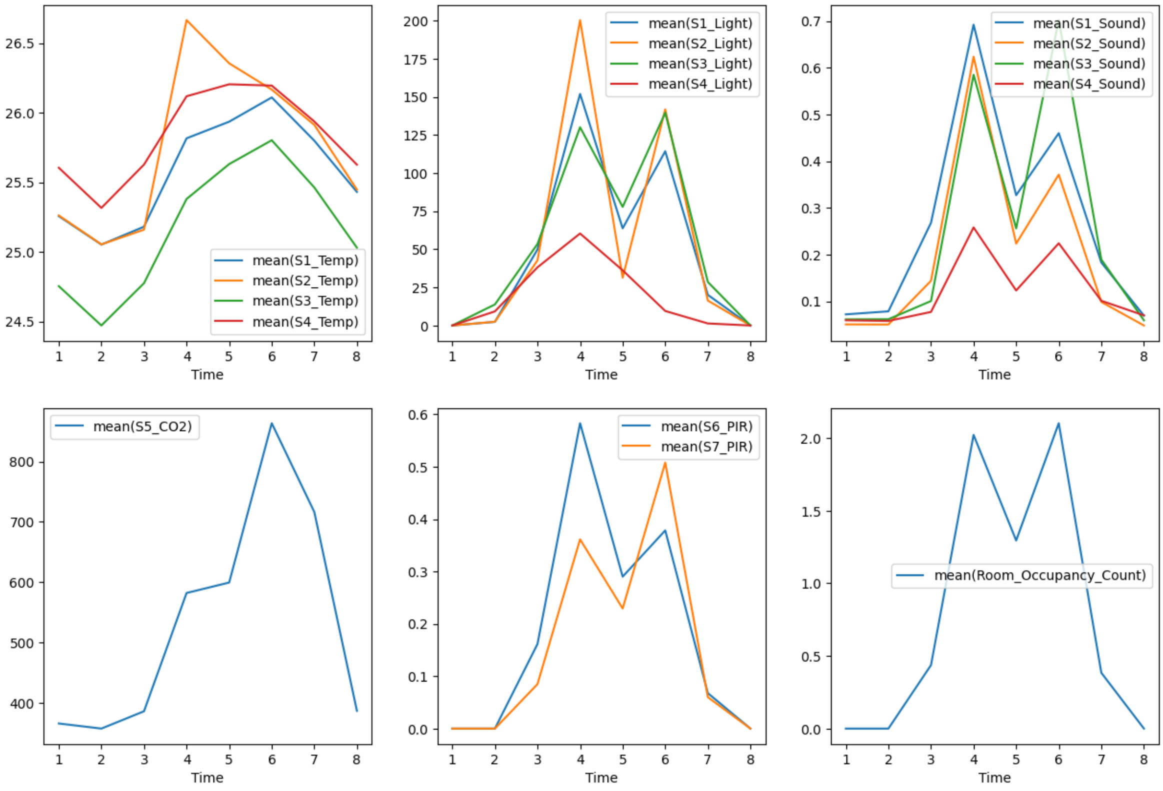


Figure 9: Sensors mean values across the times.

Environmental factors like light intensity and temperature also show trends that coincide with occupancy. Light levels rise significantly when the room is likely more active, while temperature increases can be attributed to body heat or changes in HVAC usage. However, some sensors, such as S4 for light and sound, consistently report lower values, indicating they may be positioned in less-used areas of the room and may contribute less to the overall prediction. By combining strong predictors like CO2, PIR, light, and sound data, a predictive model could effectively capture patterns tied to occupancy dynamics.

This analysis underscores the importance of both the choice of sensors and their placement for predictive accuracy. While CO2 levels and motion sensors provide direct and strong links to occupancy, features like light and sound enhance the model's robustness by capturing indirect but correlated activity. Ultimately, the integration of these features can offer a comprehensive view of occupancy patterns, enabling accurate and efficient prediction models.

In Figure 10, we reduced the dimensionality of the data points to 2 (corresponding to the two axis in the visualization) and we plotted each point with a different color for each target class, aiming to reveal its structure and relationships in lower-dimensional spaces. Specifically, the top-left panel uses Principal Component Analysis (PCA), while the remaining three panels employ t-SNE (with varying perplexity values). Each point represents a data sample, and the colors correspond to different class labels (e.g., 0, 1, 2, 3).

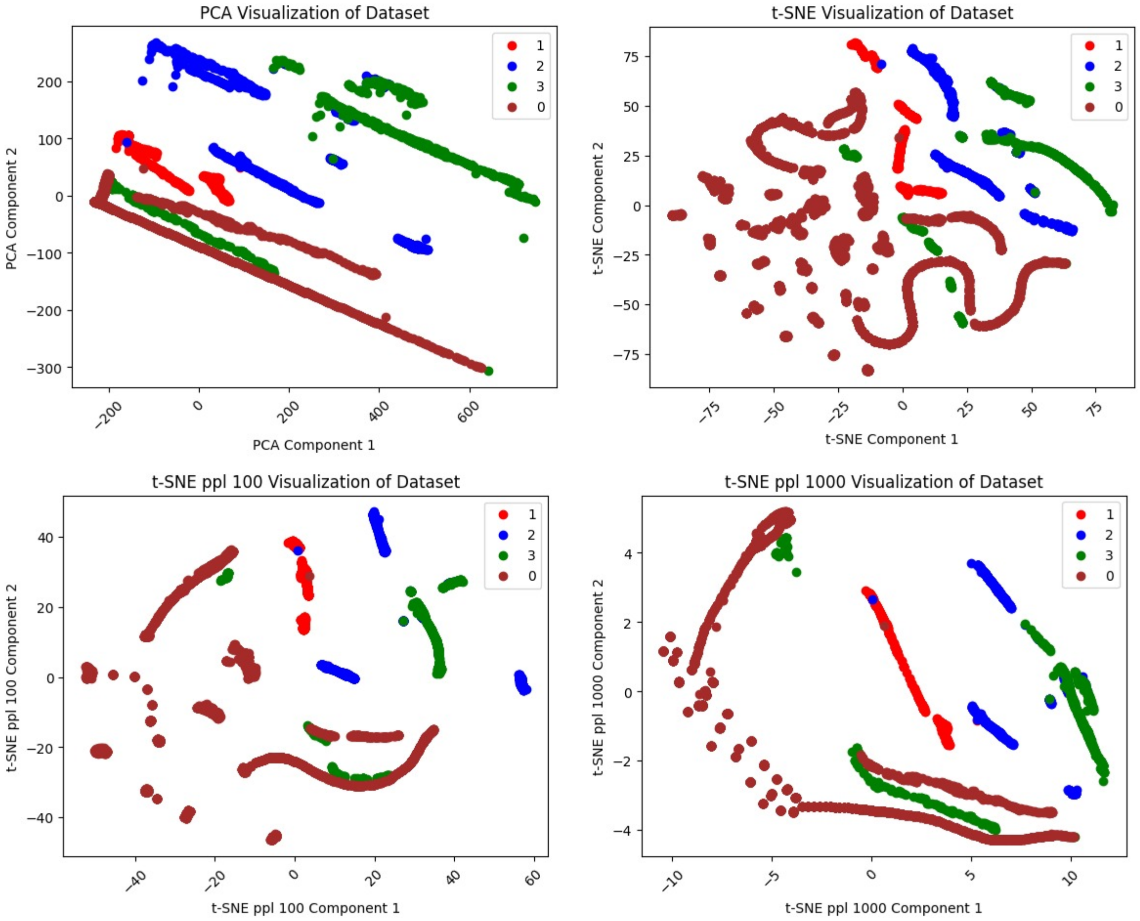


Figure 10: Dataset projections using PCA and t-SNE with different perplexities.

The PCA plot suggests that while the dataset exhibits some degree of separability between classes, there is significant overlap in certain regions. PCA relies on linear transformations and global variance maximization, which may not effectively capture complex, non-linear separations in the data. As such, the classes with more significant overlap may indicate areas where linear methods could struggle to classify accurately.

In contrast, the t-SNE visualizations, particularly in the panels with perplexities of 100 and 1000, reveal more detailed clustering patterns and better separation between classes. t-SNE excels at preserving local neighborhood structures, making it suitable for exploring non-linear relationships. The variations in perplexity highlight its impact on the balance between global and local patterns. A lower perplexity (100) focuses more on preserving local relationships, resulting in denser, more fragmented clusters. A higher perplexity (1000) provides a broader view of global relationships, with more elongated and smooth clusters.

Overall, the t-SNE visualizations suggest that the dataset exhibits non-linear separability, with clearer boundaries between some classes. These insights highlight the potential advantage of using non-linear models, such as tree-based or neural network algorithms, for classification tasks in this dataset.

5 Conclusions

The analysis highlights key aspects of room occupancy prediction based on sensor data. Initial preprocessing steps addressed the significant imbalance in occupancy distribution, with the removal of dates containing only zero-occupancy measurements. This adjustment improved data quality by reducing the risk of model overfitting. Time intervals were also converted into categorical labels, allowing for more structured analysis of temporal patterns in occupancy.

Feature analysis revealed strong correlations between room occupancy and environmental factors such as CO2 levels, temperature, light intensity, and sound. These correlations align with intuitive physical and behavioral phenomena: increased CO2 levels and sound correspond to human presence, while light and temperature trends reflect activity levels and potential HVAC adjustments. PIR sensors, though useful for detecting motion, showed slightly weaker contributions compared to other features like CO2 and sound. Notably, independence metrics indicated interdependence among features, reflecting the interconnected nature of real-world physical systems.

Overall, the integration of these features provides a robust framework for predicting room occupancy. Strong predictors like CO2, light, and sound offer direct and indirect insights into human activity, while temporal labels and sensor placement further refine the predictive model's accuracy. This comprehensive approach ensures that both spatial and temporal dynamics are effectively captured, enabling precise and scalable occupancy prediction.

References

[LaV08] Michael P LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008.

[Mil06] Steven J Miller. The method of least squares. *Mathematics Department Brown University*, 8(1):5–11, 2006.

[Pla83] Robin L Plackett. Karl pearson and the chi-squared test. *International statistical review/revue internationale de statistique*, pages 59–72, 1983.

[Spe61] Charles Spearman. The proof and measurement of association between two things. 1961.