UNIVERSITÀ DI PISA

Master's degree in Artificial Intelligence and Data Engineering

Data Mining and Machine Learning project presentation

ROOM COUNTING PEOPLE

Antonino Patania

Github repository:
https://github.com/tonipata/room-people-counting.git

A.Y. 2024/2025

Contents

1. Introduction

Nowadays, work positions in offices are greatly increasing and along with these are increasing energy consumption that ensures proper quality in work, such as heating, ventilation and air conditioning systems.
One way to optimize their usage is to make them demand-driven depending on human occupancy.
Room Counting People is an application that estimates the number of people inside a room by leveraging multiple heterogeneous sensor nodes and a machine learning model.
It exploits data from 5 different types of non-intrusive sensors: CO2, temperature, illumination, sound and motion.
The machine learning models used are:

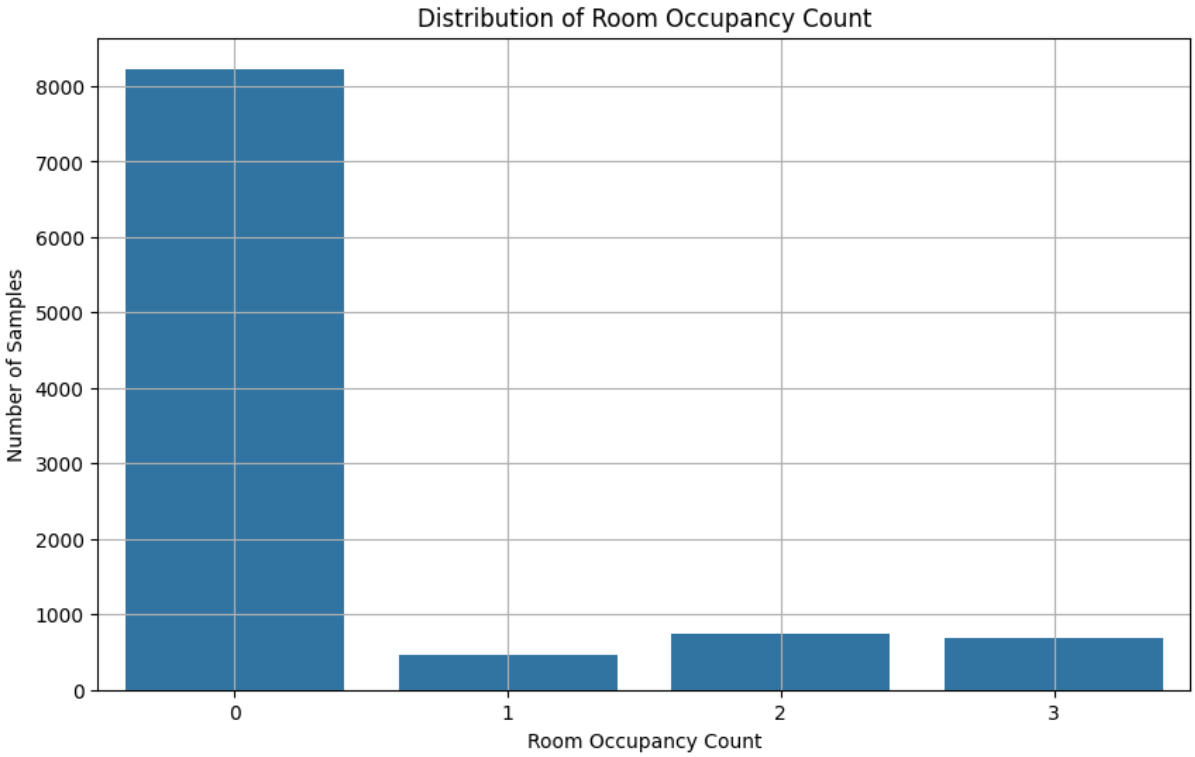
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Random Forest (RF)
- Support Vector Machines (SVM)

2. Dataset

The dataset used for this application is called “Room Occupancy Estimation” and come from https://doi.org/10.24432/C5P605
It’s composed by 18 features and 1 column for the ground truth.

FEATURES	**UNIT**
Date	YYYY/MM/DD
Time	HH:MM:SS
S1_Temp	C
S2_Temp	C
S3_Temp	C
S4_Temp	C
S1_Light	Lux
S2_Light	Lux
S3_Light	Lux
S4_Light	Lux
S1_Sound	Volts
S2_Sound	Volts
S3_Sound	Volts
S4_Sound	Volts
S5_CO2	PPM
S5_CO2_Slope	
S6_PIR	
S7_PIR	

Ground truth has 4 categories which are: ‘0’, ‘1’, ‘2’ and ‘3’.
Pir sensor (Passive InfraRed) is used for motion detection and set the value to ‘1’ when it detects a motion, so it can only have 2 values, ‘0’ and ‘1’.
S5_CO2_Slope is a feature derived from real-time CO2 values. It’s calculated by fitting a linear regression in a window of 25 points at each instance and calculating the slope of the line.
About ground truth, we can observe the distribution by means of a histogram and note that it is unbalanced in favor of category ‘0’.



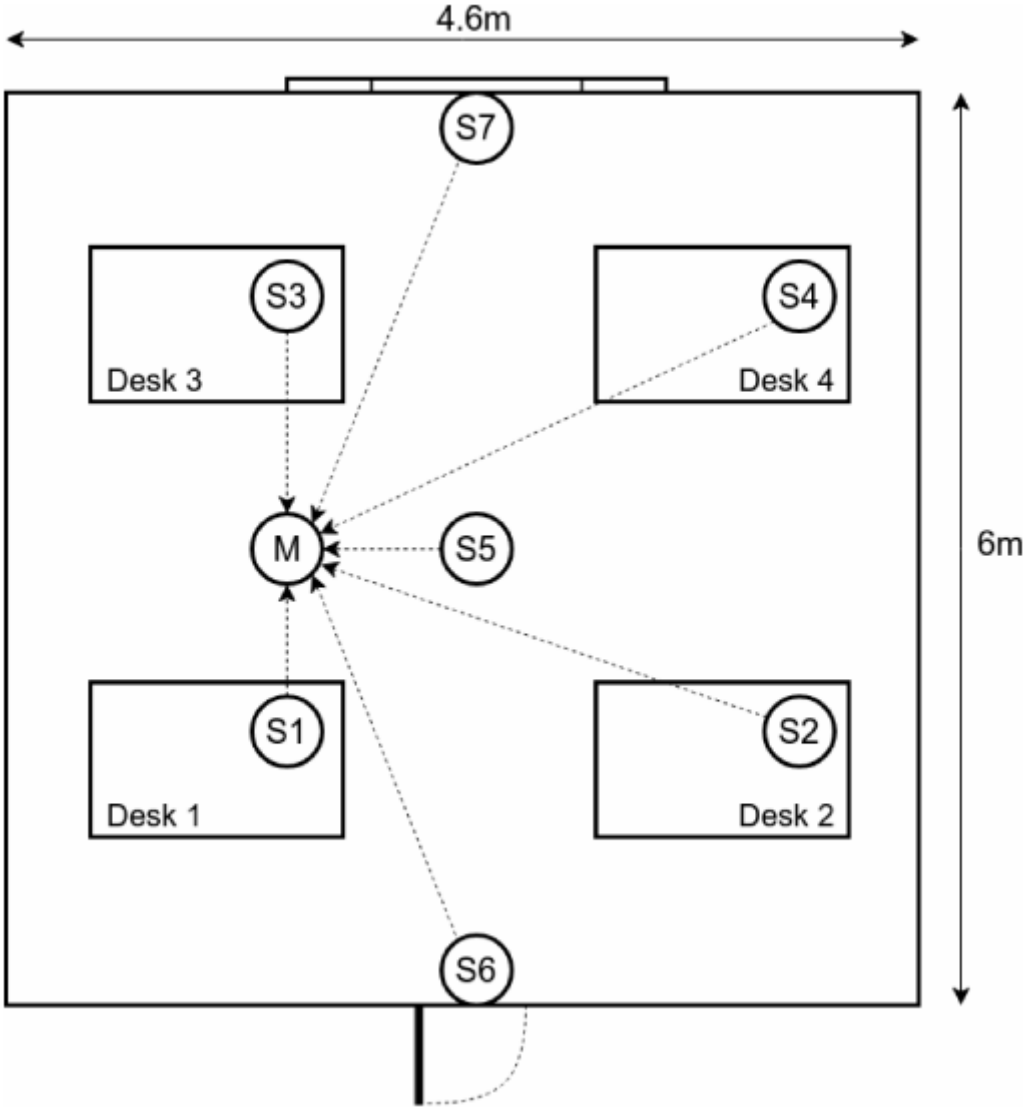
This can be explained by the fact that the system of recording sensor measurements remains active throughout the day, so during the hours when the company or office is closed, no relevant data is recorded and the number of people is obviously 0.

2.1 Setup of environment

The room where the wireless sensors network (WSN) were deployed is 6 m x 4.6 m with 4 office desks.

The network is a Zigbee based star network[1] with 7 slave nodes feeding data to the master node.

In each sensor node, a microcontroller board sampled data from the sensors and transmitted it periodically every 30 s.



[1] The Zigbee protocol is a low-power communication system designed for IoT, sensor and automation applications. In the star topology, all nodes in the network (slaves) communicate with a central node (master, coordinator Zigbee).

Instead PIR and sound sensors need a constant polling or else the events of interest are lost. Since the output pin of the PIR sensor remains high for about 3 s in repeat trigger mode. If even a single motion event was captured in the frame of 30 s, a ‘1’ was sent to the master.
For the sound sensors, the algorithm churned out the maximum peak to peak voltage that was achieved in the time frame of 30 s.

3. Preprocessing

The dataset undergoes several preprocessing steps before getting to use the machine learning models:

- Cleaning
- Normalization;
- Discretization
- Feature engineering

3.1 Cleaning

Rows that have missing sensor values are discarded for several reasons:

- Zigbee protocol is reliable, so there is a low probability that a sensor data is not valued;
- Sensor data arrive at the system every 30 seconds and it's statistically difficult for the number of people to change in such a small time window;
- Missing sensor input data are processed faster.

3.2 Normalization

Because the SVM, LDA and QDA models can be affected by different magnitude values, it was decided to use normalization.
The normalization used is the 'min-max normalization':

$$v' = \frac{v - min_A}{max_A - min_A}$$

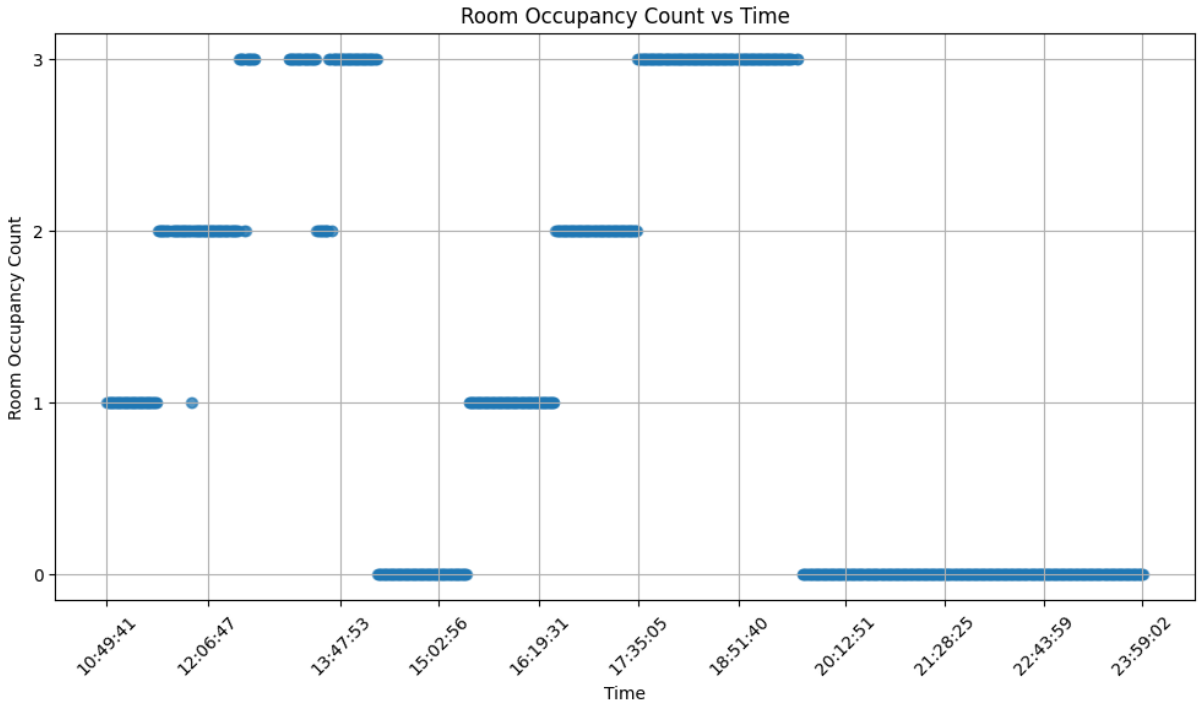
All sensor values, excluding those of PIR sensors, will have values in the range [0,1].

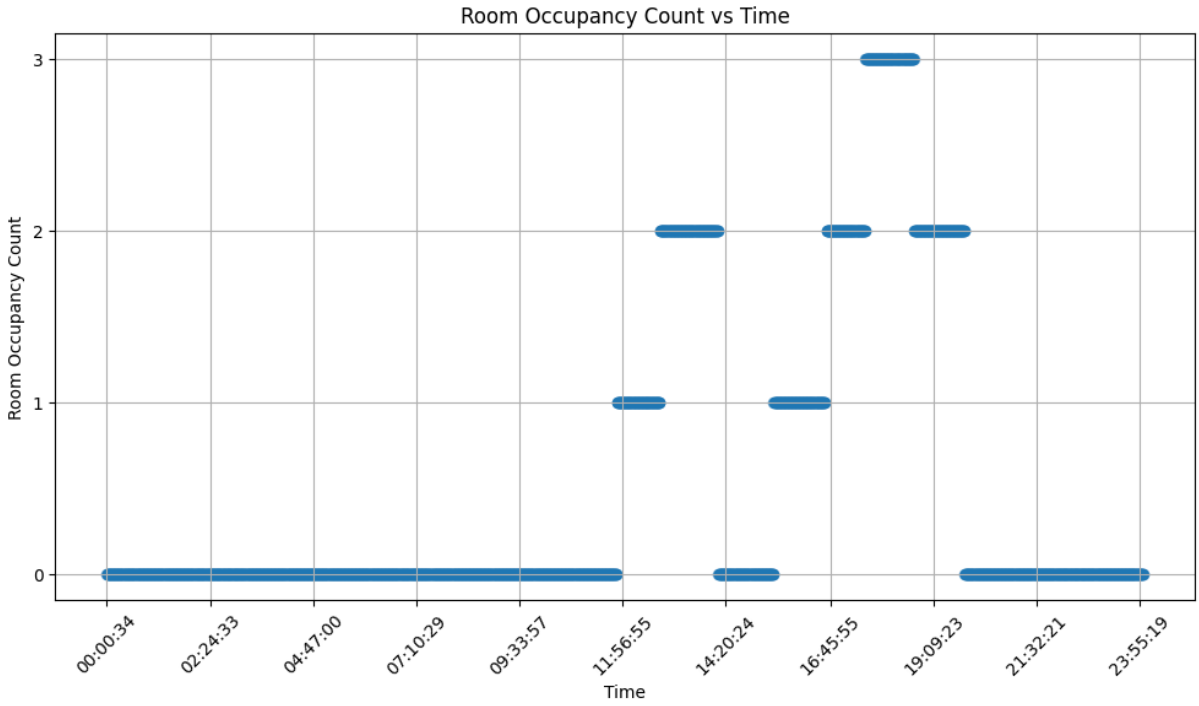
3.3 Discretization

The system remains active 24 hours a day, but its use is aimed to companies, particularly employee offices.

For this reason, it's interesting to discretize the values of the time (column 'Time') into 2 categories, 'WORK' and 'SLEEP'.

Time slots ranging from 9 a.m. to 7 p.m. fall into the first category, while the rest end up in the second.

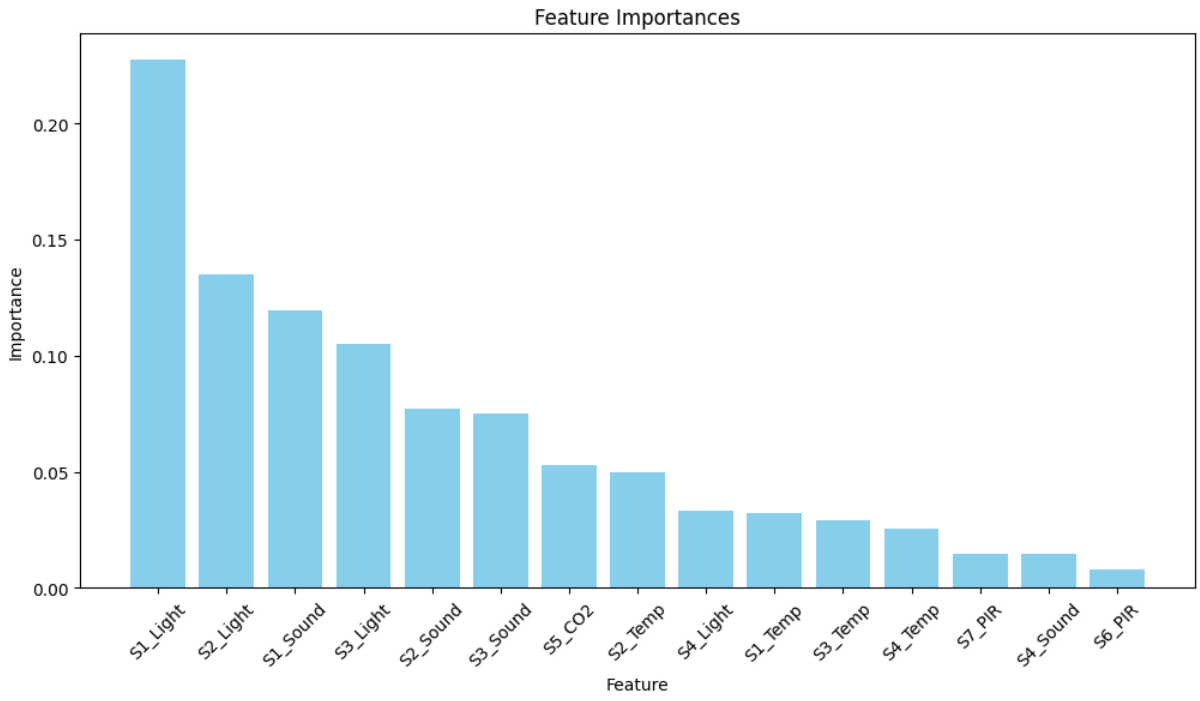




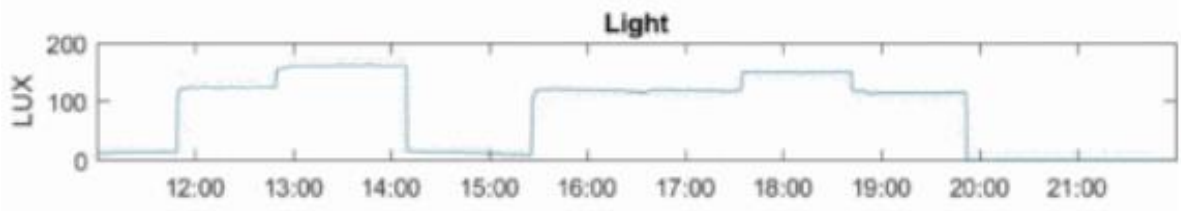
3.4 Feature Engineering

CO_2 is a good indicator for the number of people in a room, so it was decided to build a new feature based on it.

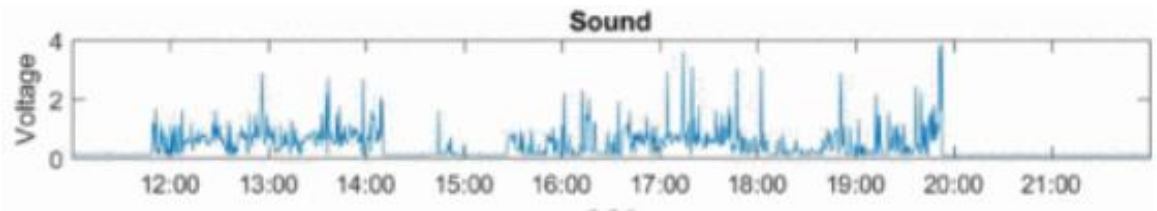
The feature represents the slope of CO_2 and is calculated by doing a linear regression on the previous 25 values of the column of interest. The value obtained is interesting because it shows the change in CO_2 over time, giving a more dynamic aspect of this column.



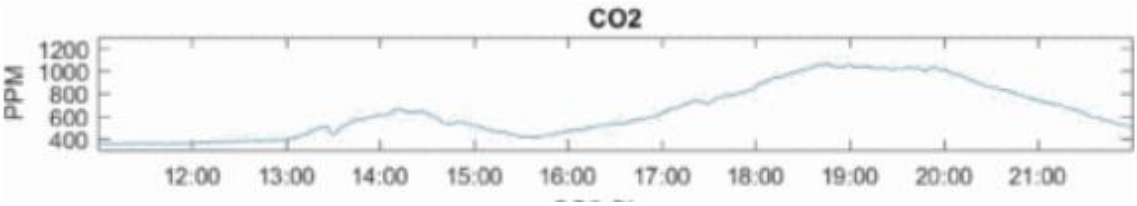
It's used Random Forest to estimate the importance of each feature before constructing a new one from 'S5_CO2'.
The most relevant features are those given by the light and sound sensors, but CO2 sensor is more relevant than temperature and motion sensors.



The change of light sensor values occurs instantaneously, generating a graph with rectangular signals.



The sound graph, on the other hand, contains abrupt changes, reaching peaks and then returning to minimum values.



While CO2 generates a more linear graph, where an increase or decrease is not recorded over a short period but over a long period (with respect to measurement sampling).

4. Classification models

As said before, the dataset is unbalanced, so accuracy cannot be relied on as a comparison parameter.
Stratified K-Fold Cross-Validation with 10 layers is used to select the best model maintaining the proportion of each class in each subdivision.
The evaluation metrics used are: precision, recall and f1.
In particular, these metrics are calculated for each class in each fold and then you do the arithmetic average (macro averaging).
After all 10 iterations have been completed, the metric calculated for each fold are averaged to obtain an overall estimate of model performance across all data.

4.1 Linear Discriminant Analysis (LDA)

LDA tries to find a linear combination of the independent features that maximizes the separation between classes.
It assumes that the data follow a Gaussian distribution within each class and have identical covariance among all classes.

$$P(X|y = k) \sim N(\mu_k, \sum)$$

μ_k: media della classe k
$\sum$: matrice di covarianza

The probability of class membership is calculated using Bayes:

$$P(y = k|X) = \frac{P(X|y = k)P(y = k)}{\sum_i P(X|y = i)P(y = i)}$$

```
Risultati della cross-validation:
Average Precision: 0.9659842896859402
Average Recall: 0.9459302888807617
Average F1-Score: 0.9549903498686344
```

4.2 Quadratic Discriminant Analysis (QDA)

QDA is a variant of LDA that allows more flexibility in classification due to one key difference: QDA does not assume classes have the same covariance matrix.

$$P(X|y = k) \sim N(\mu_k, \sum_k)$$

$\sum_k$: matrice di covarianza della classe k

```
Risultati della cross-validation:
Average Precision: 0.9733036730244489
Average Recall: 0.9734507994422025
Average F1-Score: 0.9732161504339002
```

4.3 Random Forest (RF)

```
Risultati della cross-validation:
Average Precision: 0.9930693379559978
Average Recall: 0.99243302772747
Average F1-Score: 0.9926924020452843
```

4.4 Support Vector Machines (SVM)

SVM deals with separating points in a dataset into 2 or more classes by finding a line or plane in multi-dimensional spaces that divides the classes as best as possible.

In a 2D want to find the line that is as far as possible from the nearest points of both classes.

If data are not linearly separable, SVM uses a technique called kernel. In particular, it transforms data into a space with multiple dimensions

(e.g. from 2D to 3D). In this new space, data become separable with a hyperplane.

```
Risultati della cross-validation:
Average Precision: 0.9798193396478065
Average Recall: 0.9792614678858536
Average F1-Score: 0.9793887204165432
```

5. Model selection

The choice of model, given the results, falls into RF.
One area in which to be careful is overfitting. To check for its presence, the dataset was divided into train (0.7) and test (0.3) and the same metrics used previously were printed for each of the subsets.

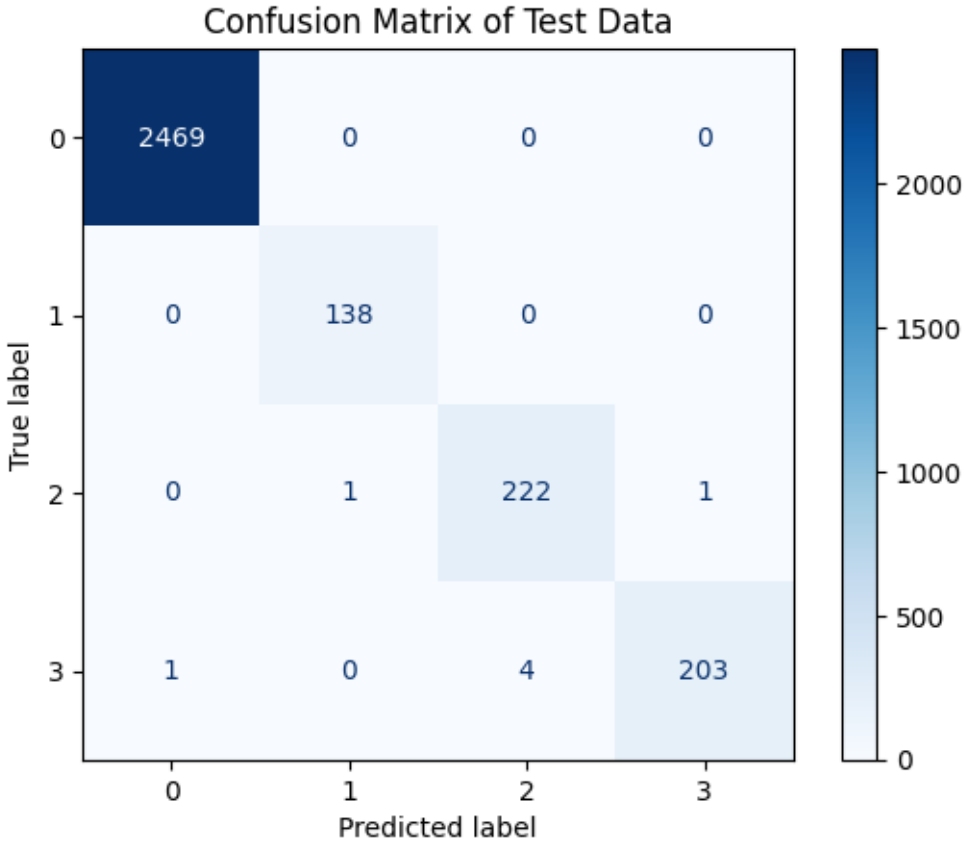
```
Performance finale sui dati di training:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      5759
           1       1.00      1.00      1.00       321
           2       1.00      1.00      1.00       524
           3       1.00      1.00      1.00       486

   macro avg       1.00      1.00      1.00      7090
weighted avg       1.00      1.00      1.00      7090
```

Performance finale sui dati di test:

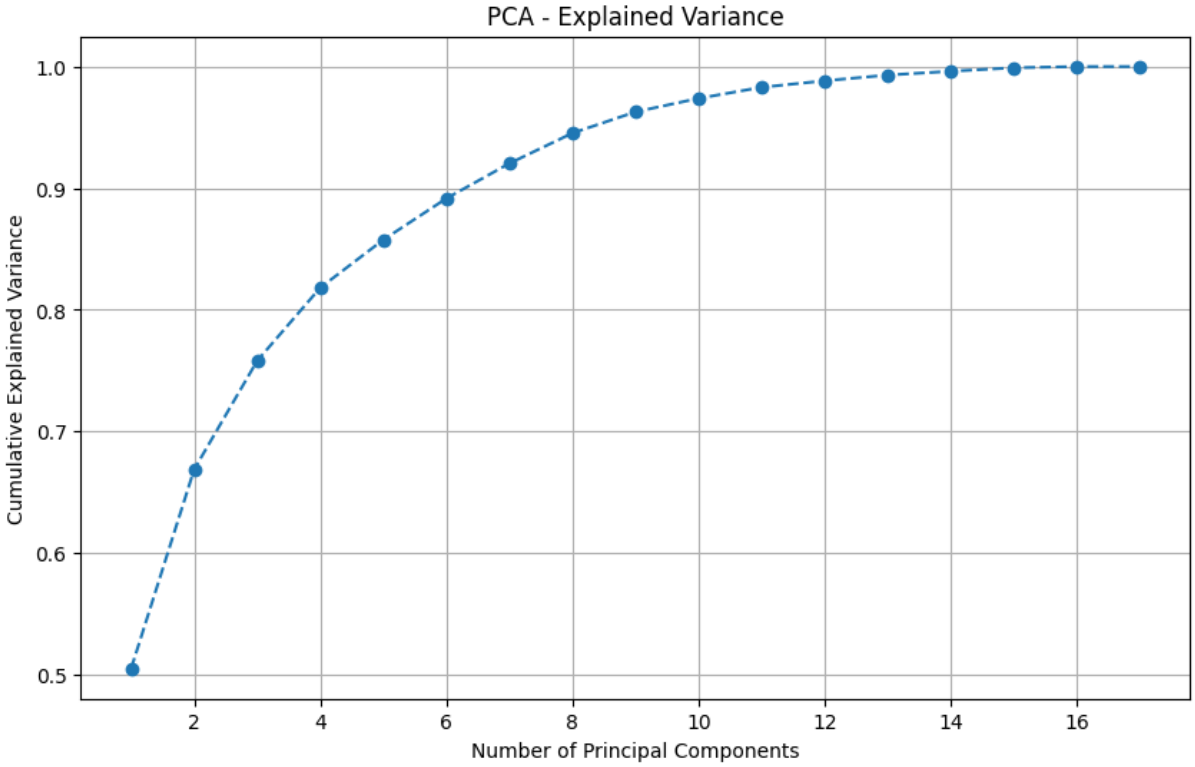
	precision	recall	f1-score	support
0	1.00	1.00	1.00	2469
1	0.99	1.00	1.00	138
2	0.98	0.99	0.99	224
3	1.00	0.98	0.99	208
macro avg	0.99	0.99	0.99	3039
weighted avg	1.00	1.00	1.00	3039



The metrics show that there is not much variation in test values compared with train values: the result is that the model is not overfitted.

5.1 PCA

Finally, PCA was applied to see if the number of features could be reduced without compromising the performance of the model.



The graph shows the percentage of cumulative explained variance as a function of the number of principal components.

It is used to determine how many principal components are needed to adequately represent the data by reducing dimensionality without losing too much information.

As can be seen, 10 components explain more than 95% of the variance.

A matrix showing how much each original feature contributes to each major component is used to choose the 10 most important features.

In the end, the 10 features with the highest absolute value for each major component are chosen.

The 10 features are:

- S2_Sound
- S4_Light
- Time_Category
- S1_Sound

- S5_CO2
- S3_Sound
- S7_PIR
- S6_PIR
- S2_Light
- S1_Light

6. Conclusion

The new reduced dataset maintains very high performance considering that 7 features are removed.

Performance finale sui dati di test:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2469
1	0.99	1.00	0.99	138
2	0.98	0.99	0.98	224
3	1.00	0.97	0.98	208
accuracy			1.00	3039
macro avg	0.99	0.99	0.99	3039
weighted avg	1.00	1.00	1.00	3039

Future developments could include the addition of new types of sensors, such as those for moisture detection to further improve the robustness and quality of the environment.

The positive results obtained demonstrate the potential of this approach in the area of energy management and improving occupant comfort.

7. Bibliography

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

Paper: By A. Singh, Vivek Jain, S. Chaudhari, F. Kraemer, S. Werner, V. Garg. 2018
https://www.semanticscholar.org/paper/e631ea26f0fd88541f42b4e049d63d6b52d6d3ac