Room Occupancy Estimation

Machine Learning Project

Azzam Abdulrahman AlFurhud

Dataset Overview

10,129 observations | 16 attributes

Attribute Name	Description
Date	YYYY/MM/DD
Time	HH:MM:SS
Temperature S1 – S4	In degree Celsius
Light S1 – S4	In Lux
Sound S1 – S4	In Volts (amplifier output read by ADC)
CO2 S5	In PPM (parts per million)
CO2 Slope S5	Slope of CO2 values taken in a sliding window
PIR S6 – S7	Binary value conveying motion detection
Room_Occupancy_Count (Target)	Number of people in the room (Manually Noted Ground Truth)

EDA

- Due to some features not only correlated but giving the same information through different sensors I believed it is better to aggregate them as follows:
  - Temperature – Sound – Light : Average
  - PIR : Summation
- I could not find appropriate use for the date\time so I did not include it in the models
- The dataset did not have any null or duplicates

Defining the features and target

- Features included: All the aggregated values + CO2 and CO2 slope
- Target: Room occupancy count
- Data was split into training (80%) and testing (20%)
- The features were also standardized for comparison of model performance with non-standardized features

Logistic Regression

```
LR confusion matrix
[[1613    0    4    2]
 [   9   84   10    0]
 [   0    2  122   40]
 [  10    6   22  102]]
#######
LR accuracy score
0.9481737413622903
#######
LR report
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1619
           1       0.91      0.82      0.86       103
           2       0.77      0.74      0.76       164
           3       0.71      0.73      0.72       140

    accuracy                           0.95      2026
   macro avg       0.85      0.82      0.83      2026
weighted avg       0.95      0.95      0.95      2026

#######
LR cross validation with f1 weighted average
Cross validation results is [0.94774191 0.95089792 0.94831601 0.95264111 0.94800591]
Average of cross validation results is 0.9495205722402724
```

```
LR confusion matrix with standardized features
[[1612    1    0    6]
 [  10   83    6    4]
 [   0    3  131   30]
 [   9    7   20  104]]
#######
LR accuracy score with standardized features
0.9526159921026653
#######
LR report with standardized features
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1619
           1       0.88      0.81      0.84       103
           2       0.83      0.80      0.82       164
           3       0.72      0.74      0.73       140

    accuracy                           0.95      2026
   macro avg       0.86      0.84      0.85      2026
weighted avg       0.95      0.95      0.95      2026

#######
LR with standardized features cross validation with f1 weighted average
Cross validation results are [0.95223682 0.9540699  0.95497631 0.95932083 0.9551636 ]
Average of cross validation results is 0.9551534912981046
```

Gaussian Naïve Bayes

```
GNB confusion matrix
[[1583    1   10   25]
 [   0   97    4    2]
 [   0    0  146   18]
 [   6    6   23  105]]
#######
GNB accuracy score
0.9531095755182626
#######
GNB report
              precision    recall  f1-score   support

           0       1.00      0.98      0.99      1619
           1       0.93      0.94      0.94       103
           2       0.80      0.89      0.84       164
           3       0.70      0.75      0.72       140

    accuracy                           0.95      2026
   macro avg       0.86      0.89      0.87      2026
weighted avg       0.96      0.95      0.95      2026

#######
GNB cross validation with f1 weighted average
Cross validation results are [0.95445217 0.96377101 0.963722   0.96339622 0.96838648]
Average of cross validation results is 0.9627455761497805
```

XGboost

```
XGB confusion matrix
[[1618    0    0    1]
 [   0  103    0    0]
 [   0    1  161    2]
 [   2    0    2  136]]
#######
XGB accuracy score
0.9960513326752222
#######
XGB report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1619
           1       0.99      1.00      1.00       103
           2       0.99      0.98      0.98       164
           3       0.98      0.97      0.97       140

    accuracy                           1.00      2026
   macro avg       0.99      0.99      0.99      2026
weighted avg       1.00      1.00      1.00      2026

#######
XGB cross validation with f1 weighted average
Cross validation results are [0.99604279 0.99703742 0.99753129 0.99654521 0.99556268]
Average of cross validation results is 0.9965438770934242
```

Models Comparison

Metrics	Logistic Regression (Standardized features)	Gaussian Naïve Bayes	XGboost
F1 Score (Macro Average)	0.85	0.87	0.99
F1 Score (Weighted Average)	0.95	0.95	1.00
Accuracy	0.9526	0.9531	0.9960
Cross-Validation F1 Score (Weighted Average)	0.9551	0.9627	0.9965

XGboost won on every metric as shown in the table above

Features Importance using the winning model (XGboost)