Classifying Room Occupancy data using classification algorithm

7072CEM Machine Learning Coursework

Student Name: Aditya Gujar
Student Id: 13366670

***Abstract*—This study proposes some classification algorithm which can be used for studying and analyzing the Room occupancy using the sensor data. In this study we have applied three classification algorithm which is Logistic regression, Random Forest classification and KNN algorithm. Based on the sensor data, the goal is to use machine learning techniques to predict room occupancy. To get the dataset ready for analysis, pre-processing procedures including data cleaning, normalization, and feature selection are carried out. These models' effectiveness is measured using parameters including accuracy, precision, recall, and F1-score. The study contributes to the field of intelligent building management systems by offering insights on the efficacy of various algorithms for estimating room occupancy using sensor data.**

Keywords—KNN, Random Forest, Logistic regression, exploratory data analysis

I. INTRODUCTION

Applications for room occupancy estimate include building automation, energy management, security, and the identification of human activities. The number of people in a space may be precisely estimated, providing useful information that can be used to optimise resource use, strengthen security, and enhance occupant comfort. In several areas, such as building automation, energy management, security, and the identification of human activities, room occupancy estimate is essential. Thanks to developments in sensor technology, it is now possible to gather comprehensive environmental data that may be used to optimise resource use and improve occupant comfort. Based on the sensor data, the goal is to use machine learning techniques to predict room occupancy. To get the dataset ready for analysis, pre-processing procedures including data cleaning, normalisation, and feature selection are carried out.

Recent developments in sensor technology have made it feasible to gather extensive and varied environmental data. Temperature, light intensity, sound levels, CO2 levels, and occupancy counts are just a few examples of the data that are included. Machine learning methods may be used to forecast occupancy trends over time or to estimate room occupancy in real-time using this sensor data.

This study's focus is on the issue of the necessity for efficient classifying the room occupancy. The study contributes to the field of intelligent building management systems by offering insights on the efficacy of various algorithms for estimating room occupancy using sensor data. It is now simpler to gather high-resolution data from multiple sensors installed in rooms because to advances in sensor technology. These sensors record data on temperature, light intensity, sound levels, CO2 levels, and other variables, giving useful insights into a room's occupancy status. The project's findings may have practical effects on occupant comfort, energy efficiency, and building management systems. An accurate estimate of room occupancy can aid with resource allocation, security enhancement, and general building management.

This study's main goal is to demonstrate how classification algorithms may be used to efficiently classify the data points. A building's ability to manage its energy needs effectively depends on accurate room occupancy prediction. Energy systems may be optimised to deliver heating, cooling, and lighting only when and where needed by understanding the occupancy state of rooms. As a result, energy is saved, the carbon impact is decreased, and money is saved. This automation increases operational effectiveness, increases occupant comfort, and decreases manual intervention. By determining the presence of inhabitants in certain locations, it provides efficient surveillance of restricted areas, assures adherence to occupancy limitations, and facilitates quick reaction during crises or disasters.

II. LITERATURE REVIEW

A. Comparative study of classification algorithms

The use of temperature and humidity sensors for room occupancy estimate is explored in the research article titled "Room Occupancy Estimation Using Temperature and Humidity Sensors: A Comparative Study". The effectiveness of several machine learning methods in precisely estimating room occupancy based on sensor data is compared by the authors. The methods comprise k-means clustering, k-nearest neighbours (KNN), and logistic regression. The authors provide information on each algorithm's implementation, including feature selection and parameter adjustment. The research article offers a thorough examination of temperature and humidity sensors' use in room occupancy estimate. It provides insights on how various machine learning algorithms perform and how well they may be applied to precisely estimate room occupancy from sensor data. The results add to the body of knowledge in the area and offer insightful advice for scholars (R et al., 2022).

B. Indoor occupancy with IoT devices

The paper focuses on using machine learning algorithms and an Internet of Things (IoT) LoRa-based monitoring system for interior occupancy detection and estimate. The importance of indoor occupancy detection for numerous applications, including energy management, security, and resource optimisation, is covered in the opening section of the study. It emphasises the requirement for precise and current occupancy data to allow smart building technologies.

The IoT LoRa-based monitoring system the authors created is described by the authors. It comprises of wireless sensors placed inside of buildings to gather environmental variables including temperature, humidity, light intensity, and CO2 levels. In order to offer accurate occupancy information, the system also has occupancy sensors(Adeogun et al., 2019)

III. PROBLEM DEFINITION AND DATASET

Our research and study focus on the issue of estimating room occupancy using sensor data. The goal is to create a model that can reliably predict whether a room is inhabited or vacant using data from several sensors, including those that detect CO2 levels, temperature, humidity, and light intensity. Accurate occupancy data must be accessible in order to optimize building energy use, strengthen security measures, and raise occupant comfort levels generally. We can put in place clever technologies that automatically alter lighting, warmth, and ventilation based on the actual number of people in the space by properly calculating room occupancy. This can result in considerable energy savings as well as more cozy and environmentally friendly interior settings.

Utilizing the sensor data to build a trustworthy model for occupancy estimate is a problem. In order to create a prediction model that can effectively generalize to unobserved data, we must examine the patterns and correlations between the sensor readings and the occupancy state. To ensure the resilience and accuracy of our model, we must also handle any concerns with data quality, missing values, and noise.

Our overall objective is to create an occupancy estimate algorithm that can properly predict, based on sensor data, whether a room is filled or not. This model will advance sustainable indoor environmental practices, improve occupant comfort, and support energy-efficient building management systems.

A. Dataset

The dataset is being downloaded from UCI machine learning repository. The name of the dataset is Room occupancy estimation dataset. It appears to have readings for a room's temperature, sound level, carbon dioxide (CO2) concentration, occupancy count, and motion detection(Singh et al., 2018). A list of the columns in the dataset is provided below:

Name of column	Description
Date	It shows the date of the measurement
Time	It shows the time
S1_Temp, S2_Temp, S3_Temp, S4_Temp	Temperature from four sensors
S1_Light, S2_Light, S3_Light, S4_Light	Intensity of light measure from four sensors
S1_Sound, S2_Sound, S3_Sound, S4_Sound	Sound intensity from four sensors
S5_CO2	CO2 level measure
S5_CO2_Slope	CO2 level slope measure
S6_PIR, S7_PIR	Motion detection
Room_Occupancy_Count	Count of room occupancy

TABLE 1: Dataset description

IV. METHODOLOGY

We are attempting to estimate the room occupancy using a dataset that includes data from multiple sensors. Our goal is to create models that can precisely predict, based on sensor data, whether a space is filled or not. We start by looking at and comprehending the dataset, together with its characteristics and intended variable. The data is then pre-processed by managing missing values, outliers, and, if necessary, data normalisation. We divide the dataset into training and testing sets after choosing pertinent characteristics. On the training data, we next train three distinct algorithms: K-Nearest Neighbours (KNN), Random Forest, and Logistic Regression. These algorithms are able to anticipate outcomes based on unobserved data and discover patterns from sensor readings. Finally, we compare the models' accuracy in forecasting room occupancy and assess their performance using suitable assessment criteria. To maximise energy efficiency and resource allocation in buildings, the most precise and trustworthy technique for room occupancy estimate must be found.

A. Logistic Regression

A popular approach for binary classification issues is logistic regression. The logistic function is used to represent the connection between the independent variables (sensor data) and the dependent variable (occupancy). P(Y=1|X) reflects the probability of occupancy being 1 given the input characteristics X, and z is the linear combination of the input features and their corresponding coefficients(Premanand, 2021) (Li, 2017). It makes use of Sigmoid function for building the model. This equation for logistic regression is provided by:

$$P(Y = 1|\, X) = \frac{1}{1+e^{-z}} \qquad (1)$$

B. Random Forest Classifier

An ensemble learning technique called Random Forest uses many decision trees to provide predictions. On several randomly selected subsets of the training data, it builds a number of decision trees and then combines their predictions. The final forecast is made by the algorithm using the majority decision or average. The building and aggregation of decision trees are addressed in the equations for the Random Forest method, which go beyond the purview of a single equation(E R, 2021). The algorithm has the following steps:

- For each particular tree in the forest, it would randomly select a small group of the dataset.
- It would also select a small group of features randomly.
- And on the selected data and the attributes it would build a decision tree. The same process is followed again sub dividing the dataset further till all the data is being covered. Basically, in Random Forest we

build a number of decision tree and then make the classification.

- When we are making the prediction, we are passing the input through each decision tree.
- The final prediction is made using the aggregation method.

C. *K-nearest neighbour*

Data is categorised using the non-parametric technique K-Nearest Neighbours depending on how close it is to other data points. It determines the distance between each training instance and the test instance before choosing the K closest neighbours. The majority vote among the K neighbours is used to decide the class designation. The Manhattan distance or Euclidean distance are two distance metrics that are used in the KNN equation to calculate the distances between data points. The algorithm works in the following way:

Each time it appears in the test data:

- Determine the distance (e.g., using the Euclidean distance) between each instance in the training data and the test instance.
- Using the obtained distances, choose the K closest neighbours.
- Assign the categorization label that appears the most often among the K neighbours. (Srivastava, 2019)

V. Experimental Setup

A. *Pre-processing*

The dataset is initially analyzed and pre-processed so that we could get a clear idea regarding the dataset and the attributes present in the dataset. The dataset is loaded and viewed. Next, we have checked the shape of the dataset. There are 10129 rows and 19 columns in the dataset. It is very important to clean the dataset and make it suitable to be used as an input to the Machine Learning model. So, for that its important to check if there are any null values present in the dataset. As there are no null values present in the dataset, we can proceed further. As we have checked the data type of the columns in the dataset, we found that there are some columns which are of object datatype. These columns are date and time so we would convert these two columns into date-time datatype of the panda's library format. We have also checked the statistical parameters of the data.

Out[3]:

	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
0	20/7/2022	10:49:41	24.94	24.75	24.56	25.38	121	34	53	40	0.08	0.19	0.06	0.06	39
1	20/7/2022	10:50:12	24.94	24.75	24.56	25.44	121	33	53	40	0.93	0.05	0.06	0.06	39
2	20/7/2022	10:50:42	25.00	24.75	24.50	25.44	121	34	53	40	0.43	0.11	0.08	0.06	39
3	20/7/2022	10:51:13	25.00	24.75	24.56	25.44	121	34	53	40	0.41	0.10	0.10	0.09	39
4	20/7/2022	10:51:44	25.00	24.75	24.56	25.44	121	34	54	40	0.18	0.06	0.06	0.06	39

Fig 1: Dataset

After pre-processing the dataset, we can now visualize the data to check and analyses the data. For visualization we have plot different columns in different graphs which gives us more understanding regarding the data and the pattern of the data points.

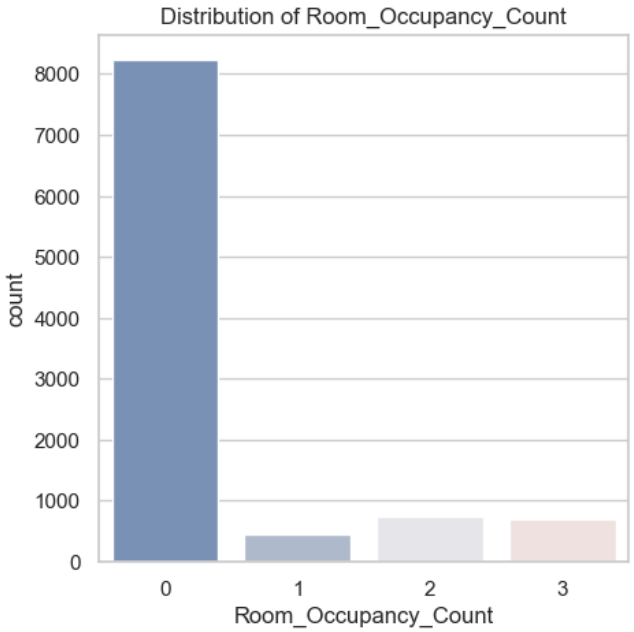

Fig 2: Distribution of Room occupancy

The figure number 2 shows the count of the room occupancy in the dataset. There are 4 numbers in the dataset which are 0,1,2,3.

- 0 means the room is unoccupied which means there are no individual in the room.
- 1 indicates that there is only single person in the room and shows a low level of occupancy of the room.
- 2 indicates there is moderate level of occupancy.
- 3 shows a higher-level occupancy in the room which represents 3 individuals in the room. From the analysis it could be noted that most of the rooms are vacant followed by equal propertion for moderate and high level of occupancy of rooms.

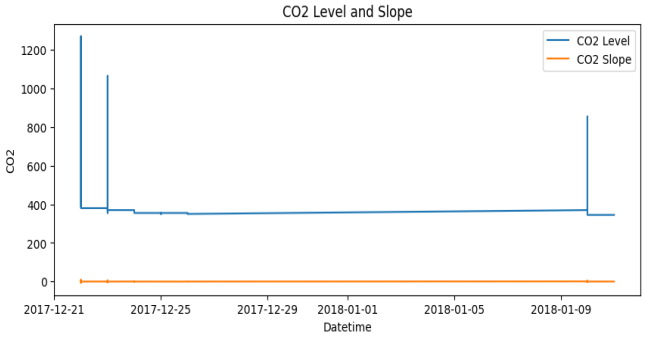

Fig 3: CO2 level and CO2 slope

. This plot of CO2 slope and CO2 level shows the relationship between these two attributes. CO2 slope and CO2 level has some pattern which each other. The sudden spike in the CO2 level shows that there is sudden increase in the CO2 level and this would ultimately affect the occupancy as well. As the CO2 slope remains stable which shows that the ventilation system is properly removing CO2 from the environment. The plot shows that there is good ventilation in the rooms and the room occupancy is also adequate.

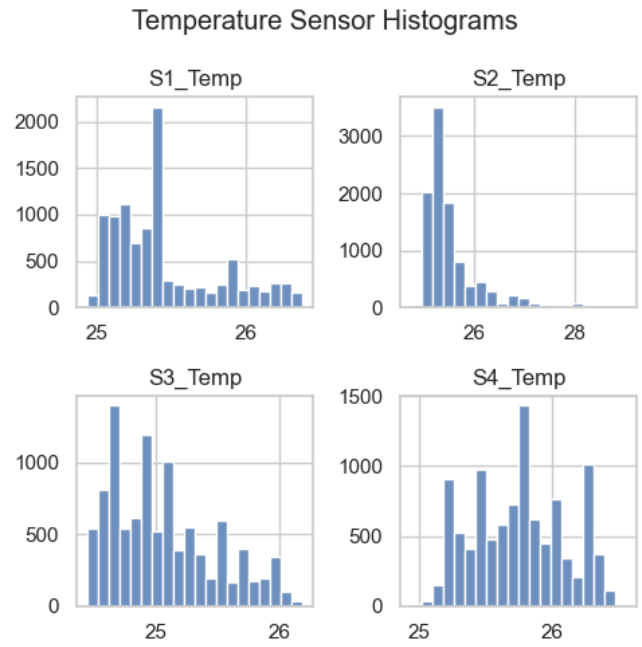

Fig 4: Temperature sensor histogram

The plot shows the histogram of four temperature sensors which are fixed in the rooms, light sensors and sound sensors in the room. By potentially analyzing the data from the three sensors we could try to understand the relation between the three sensors in the data. We could analyse that with the change in the intensity of light the temperature of the room would be also changing. So, we could analyse that with higher intensity of light the temperature has also increased. The heating and cooling systems in the room would ultimately affect the sound intensity in the room. the examination of the room's temperature, light, and sound sensors yields important information about the environment, occupant comfort, and any relationships between these variables. It aids in comprehending the room's overall ambiance and may serve as a reference for making judgements about managing the noise level, temperature, and lighting.

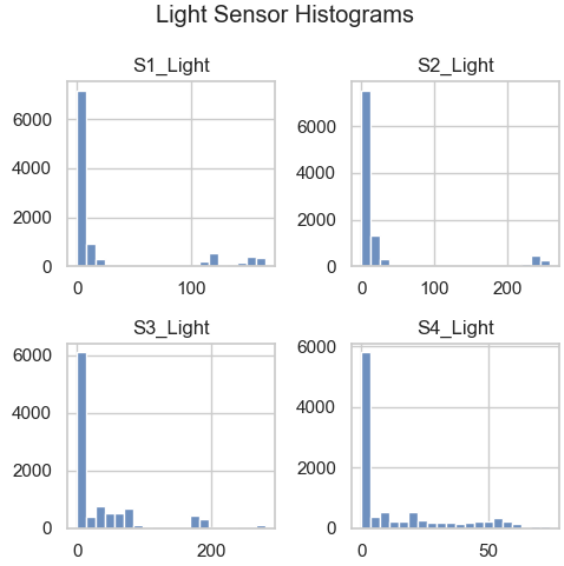

Fig 5: Light sensor histogram

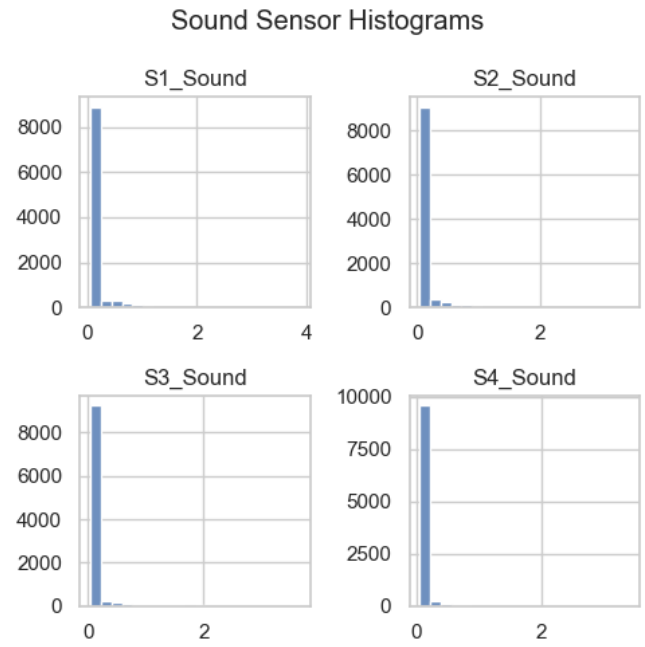

Fig 6: Sound sensors in the room

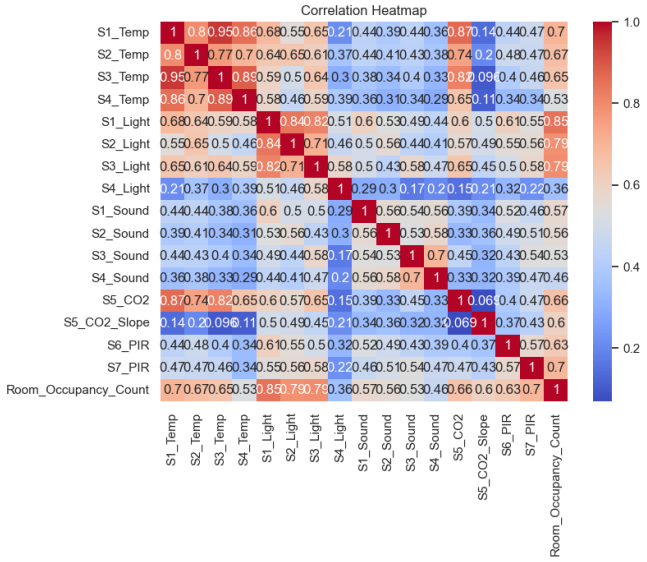

Fig 7: K-mean clustering

The heatmap shows the correlation matrix which represents the correlation between the attributes in the dataset. While the light from sensors 3 and 4 also appears to show some association with the occupancy of the room, it appears that temperature has a substantial relationship with how many people are there.

B. Classification Model building

Once the data is pre-processed, we have got a processed data which could be used as an input to the classification model and predictions and classifications could be done.
Before building the model, we have separated the dependent and independent variables. So, here the target variable is the 'Room_occupancy_count'. This is the attribute which is to be classified. After that we have performed label encoding on the given data so that we can make the data discrete.

We have implemented three classification models which are Logistic regression, Random Forest classifier and KNN classifier. All the three models are performing well.

Logistic Regression Report:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	1619
1	1.00	1.00	1.00	103
2	0.99	0.90	0.94	164
3	0.87	0.91	0.89	140
accuracy			0.98	2026
macro avg	0.96	0.95	0.96	2026
weighted avg	0.98	0.98	0.98	2026

Fig. 9: Logistic regression classification report

Random Forest classification 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	0.99	0.99	164
3	0.99	0.99	0.99	140
accuracy			1.00	2026
macro avg	0.99	0.99	0.99	2026
weighted avg	1.00	1.00	1.00	2026

Fig.10: Random forest classification report

KNN Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1619
1	0.99	1.00	1.00	103
2	0.96	0.98	0.97	164
3	0.95	0.96	0.96	140
accuracy			0.99	2026
macro avg	0.98	0.98	0.98	2026
weighted avg	0.99	0.99	0.99	2026

Fig.11: KNN classification report

Each data point is first shown as a separate cluster in the dendrogram. Clusters begin to combine as we progress up the vertical axis based on how similar or far apart, they are. The dendrogram's vertical line heights indicate how far apart or different the clusters that are being merged are from one another.

VI. RESULTS

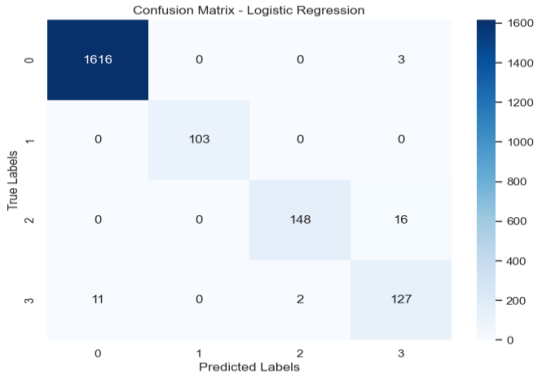

Fig.12: Confusion matrix of logistic regression

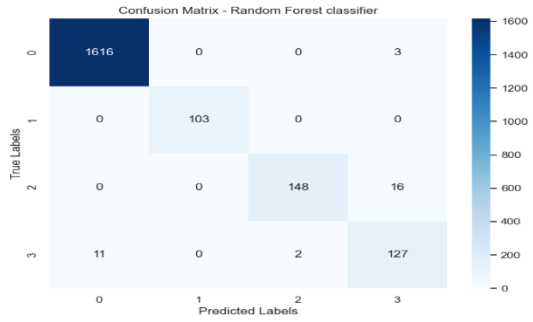

Fig.13: Confusion matrix of Random Forest

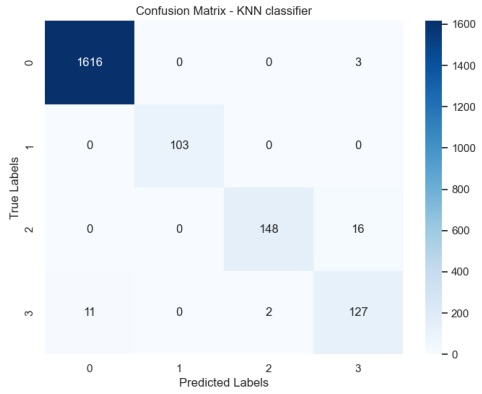

Figure 14: Confusion matrix of KNN

From the analysis and evaluation of the three models using the evaluation metrics it is found that most of the data points are properly classified into the proper class. All the three models are showing a good accuracy score. Logistic regression has an accuracy of 96%, Random Forest classifier has an accuracy of 99% while KNN classifier has an accuracy of 98%.

VII. CONCLUSION

Finally, we performed an analysis on a dataset for estimating room occupancy. To forecast the occupancy levels in a space, we looked at data from a variety of sensors, including temperature, light, sound, CO2 level, and humidity.

We carried out data pretreatment operations throughout the work, including managing missing values and normalizing the data. To better understand the connections between the various variables, we used scatter plots, correlation heatmaps, and histograms to visualize the data.

To estimate the occupancy levels, we used three classification algorithms: logistic regression, random forest, and k-nearest neighbors (KNN). In order to evaluate the models' performance, we created confusion matrices and used measures such as accuracy, precision, recall, and F1-score. All the three classification models are performing well with random forest classifier having the highest accuracy followed by KNN and then logistic regression.

References

Adeogun, R., Rodriguez, I., Razzaghpour, M., Berardinelli, G., Christensen, P. H., & Mogensen, P. E. (2019). Indoor Occupancy Detection and Estimation using Machine Learning and Measurements from an IoT LoRa-based Monitoring System. *2019 Global IoT Summit (GIoTS)*, 1–5. https://doi.org/10.1109/GIOTS.2019.8766374

R, D., Raj, K. M., Balaji, N., & K, D. (2022). Machine Learning based Estimation of Room Occupancy Using Non-Intrusive Sensors. *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, 1–5. https://doi.org/10.1109/IC3IOT53935.2022.9767992

Singh, A. P., Jain, V., Chaudhari, S., Kraemer, F. A., Werner, S., & Garg, V. (2018). Machine Learning-Based Occupancy Estimation Using Multivariate Sensor Nodes. *2018 IEEE Globecom Workshops (GC Wkshps)*, 1–6.

E R, S. (2021, June 17). *Random Forest | Introduction to Random Forest Algorithm*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Li, S. (2017, September 29). *Building A Logistic Regression in Python, Step by Step*. Towards Data Science; Towards Data Science. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

Premanand, S. (2021, October 9). *Logistic Regression | Building an End-to-End Logistic Regression Model*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/building-an-end-to-end-logistic-regression-model/

Srivastava, T. (2019, March 7). *Introduction to KNN, K-Nearest Neighbors : Simplified*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

VIII. Appendix

Dataset Link:

https://archive.ics.uci.edu/ml/datasets/Room+Occupancy+Estimation#
One Drive: Occupancy_Estimation.csv

Code Link:

https://github.com/adiguj/7072CEM_Machine_Learning_CW_Aditya_Gujar.git