RESEARCH ARTICLE | JANUARY 17 2024

Occupancy detection through machine learning and environmental data

Abdi Suryadinata Telaga; Salmaa Rihhadatul 'Aisy; Sekar Putri Arumi

AIP Conf. Proc. 2926, 020120 (2024)
https://doi.org/10.1063/5.0183264

Occupancy Detection Through Machine Learning and Environmental Data

Abdi Suryadinata Telaga[1,*], Salmaa Rihhadatul 'Aisy[2], Sekar Putri Arumi[2]

[1]*Department of Building Construction Engineering, Astra Polytechnic, Jakarta, Indonesia*
[2]*Department of Informatics Management, Astra Polytechnic, Jakarta, Indonesia*

[*]Corresponding author: abdi.telaga@polytechnic.astra.ac.id

Abstract. Occupancy detection is critical during building operations, particularly in energy efficiency, space utility, heating, ventilation control, and air conditioning (HVAC) to optimize user comfort. However, applying occupancy sensors to detect occupants can violate a person's privacy, and the resulting data could be more precise. Therefore, this study aims to detect occupancy through environmental sensors by utilizing the accuracy of machine learning-based classification results. The ultimate goal of this study is to detect whether there are people in a room based on the benchmarks studied. Benchmarks for classifying whether people are in a room include CO2, humidity, lighting, temperature, and occupancy. This paper presents a method for comparing and evaluating a set of different machine learning techniques based on a given performance measure (for example, precision, accuracy, f1 score, and algorithm recall) and then using the algorithm with the best performance. This study uses five machine learning classification algorithms: K-Nearest Neighbors, Support Vector Machine, Decision Tree, Naive Bayes and Random Forest, and the method used to evaluate the results is cross-validation. Experiments were conducted using occupancy detection datasets from the UCI Machine Learning Repository. After performing the experiment, the predictive results of the five algorithms are high. That is, each has an accuracy value of more than 90%. Support Vector Machine has the highest accuracy value compared to other algorithms, namely 98.5%, and Decision Tree has the lowest accuracy value, namely 91.9%. The results show that selecting the right features and the suitable classification model can significantly impact prediction accuracy.

Keywords: *Occupancy detection, Environmental data, Machine learning, Support vector machine, Classification*

INTRODUCTION

Currently, energy is one of the main focuses. Using less energy can reduce the depletion of existing resources in the world. Energy savings can be made in several ways, such as by improving the efficiency of heating, ventilation, and air operations conditioning (HVAC) systems by providing heating and ventilation according to the number of controlled objects in the area that needs to be controlled [1]. If the occupancy of a room can be detected automatically, lighting and cooling systems air can be directed automatically. Therefore, occupancy detection is an essential strategy for reducing energy use [2].

The study shows that when the occupancy detection system is precisely and accurately used, the energy-saving ratio can be between 30% and 42% [3]. Furthermore, occupant-centric control of appliances can save energy by a maximum of 40% [4]. Moreover, other studies show that a reduction in energy consumption of between 29% and 80% is achieved when insight is linked to occupancy and is used as input for the Heating, Ventilation Air Conditioning (HVAC) control system [5]. In addition, automatically controlling the lighting and cooling systems can support the convenience of teaching and learning activities or practicum in class and laboratory, especially for the Astra Polytechnic academic community.

While environmental indicators could be used to monitor the occupant's comfort level, the change of indicators can also indicate occupant presence in a room. Occupancy detection using environmental indicators benefits over

The 7th Progressive and Fun Education International Conference 2022
AIP Conf. Proc. 2926, 020120-1–020120-8; https://doi.org/10.1063/5.0183264
Published by AIP Publishing. 978-0-7354-4782-0/$30.00

other types of detection that can concern privacy issues [6], such as a camera or infrared sensors. Room occupancy is essential for energy management in buildings [7]. Further, data from environmental sensors can be used for demand control ventilation (DMV). Moreover, DMV could save energy between 38% to 51% [8]. However, environmental sensors have a downside: the data are sensitive to the environment and prone to false negatives/positives [9]. Therefore, advanced methods are required to detect occupancy based on environmental sensors.

Along with the rapid development of technology, machine learning can use automated systems and make computers learn to do better in the future based on what happened in the past [10]. In this study, machine learning is used to determine whether people are in a room based on data from the environment obtained with the help of sensors. The data are the room's temperature, humidity, lighting, and CO2 levels. Before the data is processed, in machine learning, irrelevant features or relevant features that harm model performance are cleaned first [11]. After that, the machine learning process involves selecting a classification algorithm with the highest accuracy value. In some previous studies, room occupancy prediction experiments have been carried out, but in that studies prediction was only applied to one existing algorithm in machine learning. Meanwhile, this study will compare five machine learning algorithms: K-Nearest Neighbors, Support Vector Machine, Decision Tree, Naive Bayes, and Random Forest. From these several algorithms, an algorithm with the best accuracy value will be selected compared to other algorithms to be applied in predicting whether there are people in a room. But before it is implemented, there will be a model performance evaluation by using the 10-Fold Cross Validation method.

METHODS

Research stages

The research uses an occupancy detection dataset. Further, the data is tested for accuracy using a classification algorithm by importing machine learning libraries in Python. The data that has been tested is compared with its accuracy results. To produce the most accurate algorithm to be applied to the data. The research phases carried out are described in FIGURE 1 below

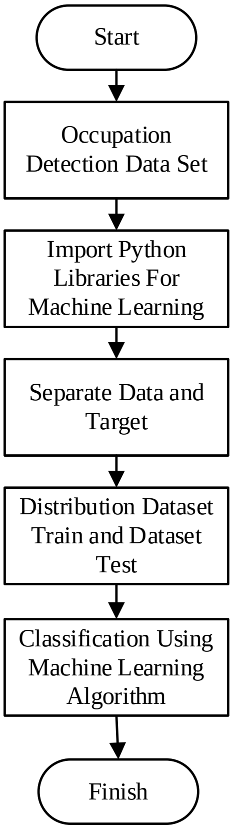
Start
Occupation Detection Data Set
Import Python Libraries For Machine Learning
Separate Data and Target
Distribution Dataset Train and Dataset Test
Classification Using Machine Learning Algorithm
Finish

FIGURE 1. Research methodology flow chart

Dataset

The dataset used in this study is the "Occupancy Detection Data Set" owned by Luis Candanedo from the University of Mons, Belgium [11]. The data is taken from the UCI Machine Learning Repository. The amount of information is 20560, with attributes, namely date, temperature, humidity, light, Co^2, humidity ratio and occupancy. Sample data are shown in table 1.

TABLE 1. Dataset sample

No	Date	Temperature	Humidity	Light	CO2	HumidityRatio	Occupancy
1	2/4/2015 17:51	23.18	27.272	426	721.25	0.004792988	1
2	2/4/2015 17:51	23.15	27.2675	429.5	714	0.004783441	1
3	2/4/2015 17:53	23.15	27.245	426	713.5	0.004779464	1
4	2/4/2015 17:54	23.15	27.2	426	708.25	0.004771509	1
5	2/4/2015 17:55	23.1	27.2	426	704.5	0.004756993	1
6	2/4/2015 18:04	23	27.125	419	686	0.004714942	1
7	2/4/2015 18:06	23	27.125	418.5	680.5	0.004714942	1
8	2/4/2015 18:07	23	27.2	0	681.5	0.004728078	0
9	2/4/2015 18:08	22.945	27.29	0	685	0.004727951	0
10	2/4/2015 18:08	22.945	27.39	0	685	0.004745408	0

Preprocessing

Preprocessing is the stage for preparing the dataset into data ready to be processed. The preprocessing stage begins with the formation of a classification model. The model is formed by configuring several parameters in the dataset. The aim is to determine the effect of parameters on the resulting performance. Building the model begins with separating the data from the target. Data are parameters - parameters that affect the prediction results.

In comparison, the target is a parameter that is classified. Data is represented by variable X, and targets are represented by variable Y. In this study, the six algorithms have the same X and Y forming parameters. X consists of temperature, humidity, light, and CO2, while Y consists of occupancy parameters.

Classification requires grouping each object. Therefore, object labelling is carried out after forming a classification model. Object labelling is divided into two classes: presence and no presence. The labelling process uses Python by looking at the value of the occupancy parameter. If the occupancy parameter has a value of 0, then "No presence"; if it has a value of 1, then "presence".

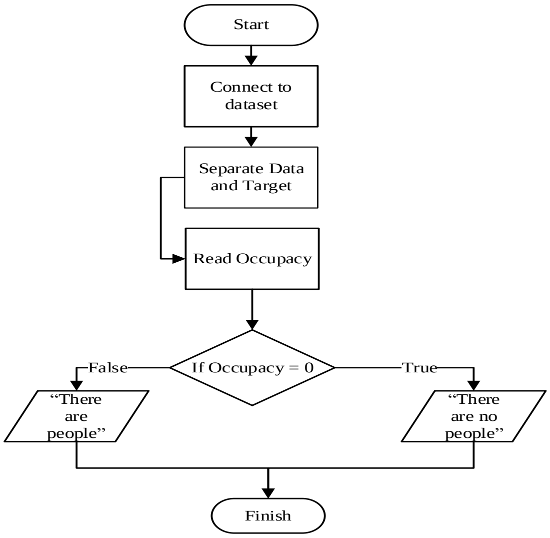

FIGURE 2. Preprocessing

Data Splitting (Training Data and Testing Data)

In machine learning, after going through data processing to become valuable data, data sharing is carried out, namely training and testing. The training data is used to train the classification model, and the data testing is used to test the classification model that has been trained. This data division, known as hold out, divides the dataset into data with a certain ratio. In this study, the authors divided the dataset into 80% for training and 20% for testing. An illustration of data sharing can be seen in FIGURE 3.

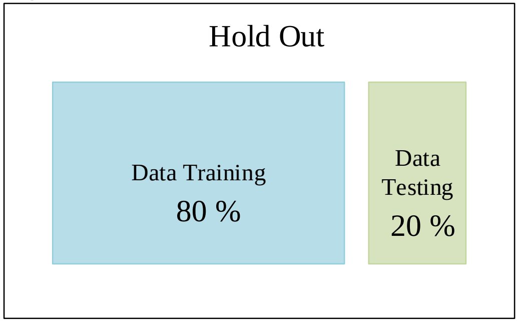

FIGURE 3. Data Splitting Illustration

Classification Model

Classification is a method of learning data in machine learning by using labelled training data (datasets). Classification models are made using five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Decision Tree, Naive Bayes, and K-Nearest Neighbors.

Support Vector Machines

The support vector machine algorithm creates the best hyperplane, the separator function. SVM will maximize the distance between two distinct sets by using a straight-shaped kernel that divides it into two classes [12] .

Random Forests

The random forest algorithm is a combination of decision trees. Each tree depends on random vector values sampled independently and with the same distribution for all trees in the forest. The strength of the Random Forest lies in selecting features when sorting each cover. Such random feature selection is capable of producing a low error rate.

Decision Trees

The decision tree algorithm is a machine learning technique that builds a representation of classification rules with a sequential hierarchical structure by recursively partitioning the training data set. This learning produces a decision tree as an n-ary branching tree.

Naive Bayes

The Naive Bayes algorithm is a simple probabilistic classification method for calculating probabilities by summing the frequencies and combinations of values from a given dataset [13].

K-Nearest Neighbors

The K-Nearest Neighbors algorithm is an algorithm with the principle that each item in the dataset generally has the closest distance to other things that have the same property [14].

Confusion Matrix

At this stage, the calculation of the results of the classification performance of each testing method is carried out using the Confusion Matrix to obtain the results of accuracy, precision, f1 score and recall. A confusion Matrix is used to analyze how well the classifier recognizes data of different classes [15]. True Positive is when the predicted result is correct, and the actual value is positive. Moreover, True Negative is when the expected result is right, and the real value is negative. False Positive is when the predicted result is wrong, and the actual value is positive. Meanwhile, False Negative is when the prediction result is wrong, and the real value is negative.

Accuracy is the level of identification that produces the percentage of the total test data whose Classification is correct.

$$\text{ccuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

Recall can be used to measure completeness, namely the percentage of positive tuples labelled positive, formulated by the equation.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (2)$$

Precision can be used to measure certainty, that is, what percentage of tuples labelled positive are true in reality and are formulated by equations.

$$\text{Precision} = \frac{TP+TN}{TP+FP} \qquad (3)$$

F1 score is the average number between precision and recall, formulated by the equation.

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision+Recall} \qquad (4)$$

Evaluation

Cross-validation is a statistical method used to evaluate the performance results of the model that has been formed. The model will be trained using training data and validated using k-fold testing data. Therefore, the arithmetic means of the k-obtained performance measures is taken to determine the success of the cross-validation. This study uses tenfold. In 10 Fold Cross Validation, the dataset is divided into ten folds of the same size. For each of the ten sub-data cross-validations, use nine folds for training and one for testing.

RESULTS AND DISCUSSION

Obtaining the algorithm that has the best performance is undoubtedly arduous. One way is to analyze the confusion matrix results for each algorithm. The confusion matrix can display how many predictions are correct and how many predictions are wrong. Figure 4 depicts the confusion matrix of the Random Forest algorithm. Figure 4 shows that the Random Forest can correctly predict 4691 of the 4743 total class 0 data (unfilled). Class 1 (filled) can predict correctly as many as 1418 out of 1425 total data. Furthermore, Figure 5 displays the Confusion Matrix of SVM, and this algorithm correctly predicts as many as 4690 out of 4743 class 0 data and 1425 of the total class 1 data.

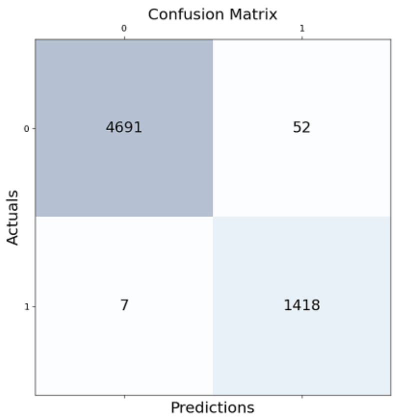

FIGURE 4. Confusion Matrix of Random Forest

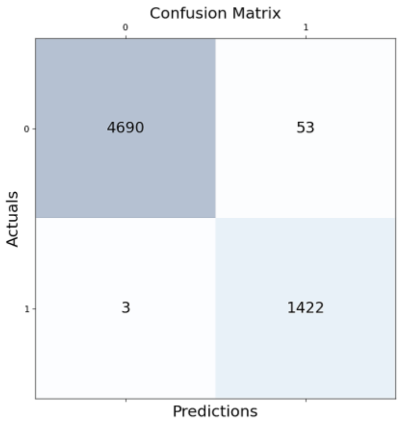

FIGURE 5. Confusion Matrix of SVM

Figure 6 shows the confusion matrix results from the Naive Bayes algorithm. Naïve Bayes can correctly predict 4581 out of 4743 total class 0 data and 1424 out of 1425 entire class 1 data. From Figure 7 it can be seen that the Decision Tree can correctly predict 4714 out of 4743 total class 0 data and can correctly predict 1395 out of 1425 whole data class 1.

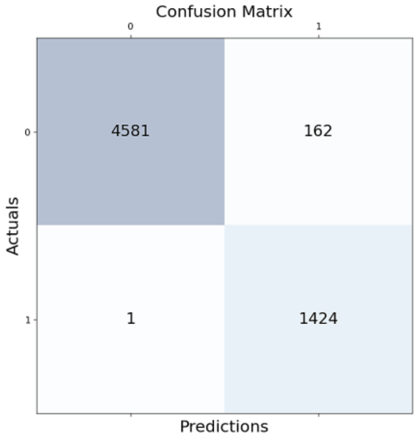

FIGURE 6. Confusion Matrix of Naive Bayes

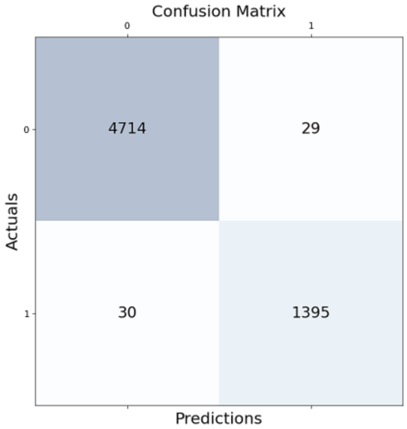

FIGURE 7. Confusion Matrix of Decision Tree

Figure 8 shows that K Nearest Neighbor can correctly predict 4713 data from 4743 total data and can correctly predict 1406 data from 1425 total class 1 data.

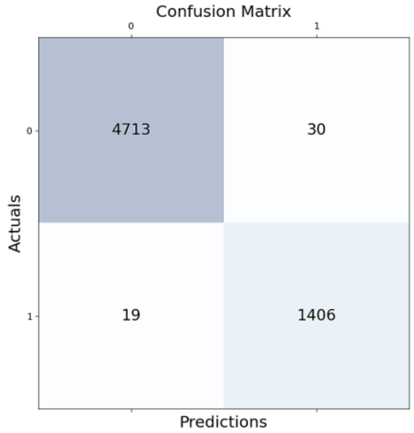

FIGURE 8. Confusion Matrix of K Nearest Neighbors

TABLE 2. Confusion Matrix Accuracy Value

Class	Support Vector Machine	Random Forest	Decision Tree	Naive Bayes	K Nearest Neighbor
0	98,88%	98,90%	99,39%	96,58%	99,37%
1	99,79%	99,51%	97,89%	99,93%	98,67%
AVG	99,34%	99,21%	98,64%	98,26%	99,02%

From Table 2 it can be seen that the performance of all the algorithms above in predicting filled and unoccupied occupancy is very good. In fact, all algorithms have an average accuracy value of more than 98%. The best score is obtained by the Support Vector Machine with an average accuracy value of 99.34%, followed by Random forest with a difference of only 0.13%, then K Nearest Neighbor with a value of 99.02% and Decision Tree with a value of 98.64%. . For the average accuracy value of the lowest confusion matrix obtained by the Naïve Bayes algorithm with a difference of only 1.05% below the Support Vector Machine which is worth 98.26%. This is because the Naïve Bayes method does not take into account the data order factor, whereas in SVM, it has a significant effect on the final prediction result.

This study also analyzes the results of precision, f1 score and recall. These results can be displayed in Table 3 below

.**TABLE 3.** Comparison Precision, Recall, F1 Score Value

Metrics	Support Vector Machine	Random Forest	Decision Tree	Naive Bayes	K Nearest Neighbor
Precision	94,6%	92,9%	83,1%	89,3%	90,9%
Recall	99,6%	87,5%	84,6%	99,8%	98,1%
F1 score	96,9%	87,8%	83,7%	94,0%	97,0%

The SVM precision results are 94.6%, Random Forest is 92.9%, Decision Tree is 83.1%, Naïve Bayes is 89.3%, and K Nearest Neighbor is 90.9%. For the highest recall results obtained by Naïve Bayes with a value of 99.8%, SVM is 99.6%, Random Forest 87.5%, Decision Tree 84.6%, and K Nearest Neighbor 90.9%. KNN obtained the highest F1 score results with a difference of only 0.1% over SVM, followed by Naïve Bayes with 94%, Random Forest with 87.8% and Decision Tree with 83.7%.

CONCLUSIONS

Based on the research results conducted to find the algorithm with the best performance in predicting occupancy detection, the Support Vector Machine method produces an accuracy value of 99.34%. Even though the difference is negligible with other comparison algorithms, the Support Vector Machine method has the best results in terms of accuracy. However, regarding Recall and F1 Score, the SVM value is below the Naïve Bayes and K Nearest Neighbor values. Even though the Recall value is not the highest, this is not a problem because there is a tradeoff between recall and precision values which means the precision will be very low when the recall is very high and vice versa. Therefore for model selection, the score between the two should not be used. The score used should be accuracy and F1 score. All values from the Support Vector Machine indicate the ideal value, which is above 93%. In conclusion, SVM is recommended for use in predicting occupancy detection.

From the SVM algorithm, it is hoped that this algorithm can predict the condition of a room correctly. Even though the predicted results are not 100% same, at least the predicted results are close to the actual values achieved. The initial stage of implementing this algorithm is also expected to be able to define whether there are people in a room or not. But in the future, this algorithm can also be developed to predict the number of people who can occupy a room so that the room remains comfortable to live in.

REFERENCES

1. Shen W, Newsham G, Gunay B. Leveraging Existing Occupancy-Related Data for Optimal Control of Commercial Office Buildings: A Review. *Adv Eng Informatics*. 2017;33:230-242. doi:https://doi.org/10.1016/j.aei.2016.12.008
2. Singh A, Kansal V, Gaur M, Pandey MS. Predicting Smart Building Occupancy Using Machine Learning BT - Proceedings of Third Doctoral Symposium on Computational Intelligence. In: Khanna A, Gupta D, Kansal V, Fortino G, Hassanien AE, eds. Springer Nature Singapore; 2023:145-151.
3. Koklu M, Tutuncu K. Tree Based Classification Methods for Occupancy Detection. *IOP Conf Ser Mater Sci Eng*. 2019;675(1):12032. doi:10.1088/1757-899X/675/1/012032
4. Natarajan A, Krishnasamy V, Singh M. Occupancy Detection and Localization Strategies for Demand Modulated Appliance Control in Internet of Things Enabled Home Energy Management System. *Renew Sustain Energy Rev*. 2022;167:112731. doi:https://doi.org/10.1016/j.rser.2022.112731
5. Brooks J, Barooah P. Energy-Efficient Control of Under-Actuated HVAC Zones in Buildings. In: *2014 American Control Conference*. ; 2014:424-429. doi:10.1109/ACC.2014.6859151
6. Vigna I, Balest J, Pasut W, Pernetti R. Office Occupants' Perspective Dealing with Energy Flexibility: A Large-Scale Survey in the Province of Bolzano. *Energies*. 2020;13(17). doi:10.3390/en13174312
7. Wang W, Chen J, Hong T. Occupancy Prediction Through Machine Learning and Data Fusion of Environmental Sensing and Wi-fi Sensing in Buildings. *Autom Constr*. 2018;94:233-243. doi:https://doi.org/10.1016/j.autcon.2018.07.007
8. Mysen M, Berntsen S, Nafstad P, Schild PG. Occupancy Density and Benefits of Demand-Controlled Ventilation in Norwegian Primary Schools. *Energy Build*. 2005;37(12):1234-1240. doi:https://doi.org/10.1016/j.enbuild.2005.01.003
9. Ahmad J, Larijani H, Emmanuel R, Mannion M, Javed A. Occupancy Detection in Non-Residential Buildings – A Survey and Novel Privacy Preserved Occupancy Monitoring Solution TT - Occupancy Detection in Non-Residential Buildings. *Appl Comput Informatics*. 2021;17(2):279-295. doi:https://doi.org/10.1016/j.aci.2018.12.001
10. Holliday J, Sani N, Willett P. Ligand-Based Virtual Screening Using a Genetic Algorithm with Data Fusion. *Match Commun Math Comput Chem*. 2018;80(3).
11. Candanedo LM, Feldheim V. Accurate Occupancy Detection of An Office Room from Light, Temperature, Humidity and CO2 Measurements Using Statistical Learning Models. *Energy Build*. 2016;112:28-39. doi:https://doi.org/10.1016/j.enbuild.2015.11.071
12. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing*. 2020;408:189-215. doi:https://doi.org/10.1016/j.neucom.2019.10.118
13. Chen H, Hu S, Hua R, Zhao X. Improved Naive Bayes Classification Algorithm for Traffic Risk Management. *EURASIP J Adv Signal Process*. 2021;2021(1):30. doi:10.1186/s13634-021-00742-6
14. Mucherino A, Papajorgji PJ, Pardalos PM. k-Nearest Neighbor Classification BT - Data Mining in Agriculture. In: Mucherino A, Papajorgji PJ, Pardalos PM, eds. Springer New York; 2009:83-106. doi:10.1007/978-0-387-88615-2_4
15. Düntsch I, Gediga G. Confusion Matrices and Rough Set Data Analysis. *J Phys Conf Ser*. 2019;1229(1):12055. doi:10.1088/1742-6596/1229/1/012055