



HAL
open science

Predicting the opening state of a group of windows in an open-plan office by using machine learning models

Thi Hao Nguyen, Anda Ionescu, Evelyne Géhin, Olivier Ramalho

► To cite this version:

Thi Hao Nguyen, Anda Ionescu, Evelyne Géhin, Olivier Ramalho. Predicting the opening state of a group of windows in an open-plan office by using machine learning models. *Building and Environment*, 2022, 225, pp.109636. 10.1016/j.buildenv.2022.109636 . hal-04092101

HAL Id: hal-04092101

<https://hal.u-pec.fr/hal-04092101v1>

Submitted on 9 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Predicting the Opening State of a Group of Windows in an Open-Plan Office by using Machine Learning Models

Thi Hao Nguyen¹, Anda Ionescu¹, Evelyne Géhin¹, and Olivier Ramalho²

¹ Univ Paris-Est Creteil, CERTES, F-94010 Creteil, France

{thi-hao.nguyen, ionescu, gehin}@u-pec.fr

² Scientific and Technical Center for Building

Champs-sur-Marne 77447, France

{olivier.RAMALHO}@cstb.fr

Abstract. Window operation is among one of the most influencing factors on the indoor air quality (IAQ). The opening state of the windows can modify the air exchange rate and as such the pollutant transfer between indoor and outdoor environments. In this paper, we focus on the modeling of the windows opening state in a real open-plan office with five windows. For this purpose, three machine learning-based models were implemented: (i) Decision Tree, (ii) k-Nearest Neighbors and (iii) Kernel Approximation. IAQ, climatic parameters and the opening state of the windows have been monitored during an entire period of 18 months. The information about: (i) the environmental factors from the previous 24th hour and (ii) the current time (month, day of the week, hour of the day) was used to predict the current state of the windows. The predictor importance estimation and the calculated autocorrelation functions showed that the three most relevant factors were: the previous 24th hour of the windows status, the current time and the previous 24th hour of the prevailing mean outdoor air temperature. The three models perform well with the testing sets according to the different evaluation indicators. The developed methods can be helpful for understanding occupant behavior and also for controlling indoor air pollutants levels in buildings, either as a standalone model or a part of a real-time IAQ monitoring system.

Keywords: indoor air quality · windows opening state · machine learning model · time series · autocorrelation functions · open-plan office

1 Introduction

The outbreak of the COVID-19 virus towards the end of 2019 has left people all around the world with unforgettable memories. This virus rapidly spreads from one to another by interacting in a closed area, which serves as a warning for us to be more concerned about the environment in which we live. According to the World Health Organization (WHO), humans spend more than 90 percent of their time indoors [1]. As a consequence to the lockdown, restricted movement, working from home, and other factors, the time spent indoors increased and research on IAQ became of utmost importance.

The opening state of the windows has an important influence on IAQ, as it can modify the air exchange rate and as such the transfer between indoor and outdoor environments [16]. Opening a window may lead to a sudden increase in the air exchange rate and to both (i) a quick decrease of the concentration of indoors generated pollutant like CO₂ and (ii) a possible increase of the indoor concentration of pollutants coming from outdoors as PM. A research in a mock-up building revealed that the thermal comfort and indoor air quality can be improved by window opening/closing [26]. It is therefore necessary to understand and model the influence of this factor on IAQ.

Window-opening activity is affected by a variety of parameters, such as outdoor temperature, air quality, human presence and season [28, 31, 27]. Occupant's behavior is an important factor but it can vary among individuals [27], leading to different impacts on the indoor environment [28].

On the one hand, theoretical physics-based models (models based on physics rules) struggle to explain the changes in window-opening behavior [9], in the perspective of direct modeling. On the other hand, machine learning models develop computational algorithms designed to simulate human intelligence by learning from their surroundings [13], in the perspective of inverse modeling. Considering the complexity of the underlying relationships, a machine learning model could be a good alternative to a physics-based model and a powerful tool for predicting or forecasting window-opening behavior.

In the last decades, Machine learning (ML) models have been effectively used in the prediction of indoor air quality [37, 7, 23, 28] and energy consumption [2, 12], proving the potential of using

57 machine learning models in indoor environments. Regarding windows opening modeling, a recent
58 study [35] has used the Deep Learning technique for Neural Networks (a specific type of ML)
59 for the detection and recognition of the opening state of the windows by using a camera in
60 order to propose frameworks for energy saving. According to the review paper [9], the common
61 ML models for predicting window-opening behavior include: logistic regression, artificial neural
62 networks (ANN), the Markov chain model, and support vector machines (SVM).

63 A stochastic window status profile generator (WinProGen) using Markov chains method has
64 been introduced by Calì and colleagues [6]. The model used a database with transition probabil-
65 ity matrices from 300 windows in 60 apartments in southern Germany, monitored during 2012
66 with 1-minute time step. Reliable predictions of buildings' energy performance are obtained when
67 applying these generated window state profiles to the dynamic simulation of two demonstrator
68 buildings. This model has the advantage of appropriately accounting for the process's time de-
69 pendency. However, according to the authors, this model struggled to deal with a large number
70 of input variables in comparison with the logistic regression method. Therefore, they proposed,
71 as future work, to develop a hybrid model, combining both the Markov chain technique and the
72 logistic regression analysis [6].

73 Logistic regression [21] is a statistical approach that determines the likelihood of a given event
74 (e.g., opening a window) occurrence based on relevant factor elements (e.g., outdoor/indoor air
75 temperature or PM2.5 concentrations). Most of the research used logistic regression to compute
76 the correlation between the probability of a window opening and the variables of influence [3,
77 38]. In these two studies, the research was conducted on 19 dwellings in Beijing [38] and 15
78 residencies in Denmark [3]. Predictive models of the occupants' window opening behavior were
79 established based on multivariate linear logistic regression. This method has the advantage of
80 providing interpretative parameters and could be regularized to minimize over-fitting. However,
81 the model struggles to address the complicated relationships, due to its low flexibility [11].

82 Other researchers attempted to apply the data-mining approach to discover the effects of
83 the window opening and closing behavior in energy consumption in buildings [10]. This paper
84 proposes a framework for identifying valid window operating patterns, in measured data, by
85 combining logistic regression analysis with two data-mining approaches: (i) cluster analysis and
86 (ii) association rules mining. In this study, 8 non-numerical and 7 numerical variables were used

87 for calculating the probability of opening and closing of a window. In total, a huge quantity of
88 detailed data was used. The authors succeeded to obtain distinct behavioral patterns to serve as
89 a basis for 12 association rules, which classified two typical window opening office user profiles:
90 (i) physical environmental driven and (ii) contextual driven. Based on that, appropriate recom-
91 mendations for different natural ventilation strategies as well as robust building design could be
92 achieved.

93 A similar study [22] suggested a generic model that identifies window states using a fully
94 connected feed-forward neural network. For both training and testing processes, this model used
95 around 20 million data samples, which were measured in Germany and USA. The proposed model
96 was evaluated on an additional data set, which was divided into adaptation set and evaluation set.
97 During the adaptation process, the pre-trained weights were adapted by running several tuning
98 iterations, while no hyperparameter tuning or further calibration was required. Based on this
99 procedure, the only required step is the weight adaptation when applied to the other buildings,
100 otherwise, this model did not require any parameter search or calibration. The resulted model
101 could be used by the engineers and designers as a standalone, or as a part of a thermal building
102 simulation.

103 Six machine learning algorithms were trained in the research of Park et al. [28]. The authors
104 have used monitoring data of 23 sample homes located in Seoul and suburban areas for predicting
105 the occupant's behaviour in the manual control of windows. According to the analysed predictive
106 performance, the k-NN model shows the best fitness with the monitored data set. Regarding
107 the input parameters, the Gini importance score indicated that there are five main driving
108 parameters: (i) prevailing mean outdoor air temperature (PMA), (ii) mean daily temperature,
109 (ii) CO₂ indoor concentration, (iv) relative humidity indoors and (v) the difference between
110 outdoor temperature and the operative temperature indoors.

111 The Kernel Approximation method has been mainly applied in speech enhancement methods
112 [39]. Regarding the Decision Tree, this method has been used to classify the most important
113 parameters among a large range of variables such as: sociodemographic data, health and lifestyle
114 habits, ergonomic and psychological factors for the Sick Building Syndrome (SBS) [33].

115 For our study case, we decided to study the ability of different ML classifiers including:
116 Decision Tree, k-NN classification and Kernel Approximation (SVM kernel), to predict the state

117 of the window opening in an open-plan office, as presented hereafter. The reason why we chose
 118 Decision Tree is that this method offers the possibility to obtain the extracted rules, which
 119 can be applied then for other study cases. Regarding k-NN, this method is recommended as 'a
 120 theoretically optimal method of classification' [17]. Finally, we chose Kernel Approximation as
 121 it can take into account the non-linearity relationship among the variables. Some information
 122 about these three methods is presented in the section 3.

123 2 Study Case and Features Selection

124 2.1 The open-plan office

125 The studied open-plan office is located in the suburban town of Champs-sur-Marne, approx-
 126 imately 30 km East of Paris, France. The office has a total area of 132 m^2 and a volume of
 127 364 m^3 . This office is situated on the 2nd floor of the building and occupied by 6 to 15 people,
 128 from 8:00 a.m. to 6:00 p.m., from Monday to Friday. The cleaning task is vacuuming, which
 129 generally takes place at the end of the week, on Friday, during the end of the day (around 8
 130 p.m.). Figure 1 represents the layout of this office.

131 The studied building is a relatively modern one, with walls that are around 20 cm thick. It
 132 has two floors and several offices, conference rooms, experimental laboratories, etc. Inside and
 133 outside the office, measurement devices were installed. The monitoring was performed during
 134 18 months, from January 1st, 2014 to June 30th, 2015. Temperature (T), relative humidity
 135 (RH) and carbon dioxide (CO_2) concentration indoors were measured by a Q-Track instrument
 136 (TSI Inc.). Particulate matter concentrations in number (PN) were monitored by an optical
 137 particle counter (Grimm Dust Monitor 1.108). Concerning the outdoor environment, a permanent
 138 weather station located on the roof of the target building automatically recorded the temperature,
 139 relative humidity, atmospheric pressure, speed and direction values of wind. It has also detected
 140 rainy events. All of the parameters were monitored with 1-minute time-step.

141 It is possible to calculate the specific humidity (H_s) by calculating first the absolute humidity
 142 (H_{abs}) which is based on the relative humidity (RH), the air temperature (T) and the molar
 143 mass of the water (M_{water}) and of the air (M_{air}) by using Rankine's formula to approximate the

144 saturated vapor pressure required for the calculation (see Equations (1) and (2)).

$$H_{abs}\left(\frac{g}{kg} humidAir\right) = \frac{RH}{100} \times \frac{M_{water}}{M_{air}} \times e^{(13.7 - \frac{5120}{T+273.15})} \times 1000 \quad (1)$$

145

$$H_s\left(\frac{g}{kg} dryAir\right) = \frac{H_{abs}}{(1000 - H_{abs})} \times 1000 \quad (2)$$

146 As it is much easier to obtain the PM (particulate matter in mass concentration) value than
 147 the PN one for a real-time model, from the PN concentrations, we calculated the mass fractions
 148 of PM2.5 and PM10 according to the method of [8]. The equations (3) and (4) explain how to
 149 convert the particle concentrations obtained into mass concentration ($\mu\text{g}\cdot\text{m}^{-3}$) and then calculate
 150 the PM2.5 and PM10 fractions. According to [8], we first transform the concentration in number
 151 into mass concentration:

$$m(d_{pi}) = C_f \frac{\pi}{6} d_{pi}^3 n(d_{pi}) \quad (3)$$

152 where i corresponds to the channel number of the optical particle counter, d_{pi} corresponds to
 153 the average diameter between the lower and upper limit of the channel, $m(d_{pi})$ is the mass
 154 concentration, C_f is the correction factor which corresponds to the particle density and it is
 155 fixed at $1 \mu\text{g}\cdot\text{cm}^{-3}$ by default [8] and $n(d_{pi})$ corresponds to the concentration in number. Then,
 156 the equation (4) allows the calculation of PM2.5 and PM10 fractions.

$$PM = \sum_{i=1}^{15} m(d_{pi}) f(d_{pi}) \quad (4)$$

157 where PM corresponds to PM2.5 or PM10 and $f(d_{pi})$ is the fraction of d_{pi} taking into account
 158 the collection efficiency of the reference instruments [19]. These contributions can be estimated
 159 for each fraction of particles by the equations (5-8) below.

$$f_{PM10}(d_{pi}) = 1 \quad \text{for } d_{pi} < 1.5\mu\text{m} \quad (5)$$

$$f_{PM10}(d_{pi}) = 0.9585 - 0.00408d_{pi}^2 \quad \text{for } 1.5 < d_{pi} < 15\mu\text{m} \quad (6)$$

$$f_{PM10}(d_{pi}) = 0 \quad \text{for } d_{pi} > 15\mu\text{m} \quad (7)$$

$$f_{PM2.5}(d_{pi}) = [1 + \exp(3.233d_{pi} - 9.495)]^{-3.368} \quad (8)$$

160 The mean daily temperature and prevailing mean outdoor air temperature (PMA) were cal-
 161 culated using the seven-day weighted running mean outdoor air temperature. According to
 162 ASHRAE, equation (9) gives the preferred expression for PMA with "an exponentially weighted,
 163 running mean of a sequence of mean daily outdoor temperatures prior to the day in question"
 164 [18].

$$\text{PMA} = (1 - \alpha)[t_{e(d-1)} + \alpha t_{e(d-2)} + \dots + \alpha^6 t_{e(d-7)}] \quad (9)$$

165 For midlatitude climates, where people are more familiar with synoptic-scale weather variability,
 166 a lower value of α could be more appropriate so we chose $\alpha = 0.6$. In Equation (9), $t_{e(d-1)}$
 167 represents the mean daily outdoor temperature for the previous day, $t_{e(d-2)}$ is the mean daily
 168 outdoor temperature for two days before, and so on.

169 The studied office has a permanent mechanical exhaust ventilation. There is no air condition-
 170 ing and the heating system of the building is a central one. The single flow ventilation system
 171 provides a constant air extraction rate of $228 \text{ m}^3 \cdot \text{h}^{-1}$ (measured in 2014 at $\pm 6\%$). Ten air inlets
 172 are attached to the joinery of the five sliding windows. These five windows were equipped with
 173 contact sensors that detected each opening or closing event and recorded to a local server unit
 174 through a wireless zigbee protocol. The main entrance door is equipped with a door contactor.
 175 A motion detector was also used to record the occupancy of the office. The collected data is
 176 transmitted to and stored on a central server. The monitored window opening states represent
 177 time series with irregular time steps. The detection modules send back information as soon as
 178 a change of state occurs according to occupants action. Therefore, a pre-processing stage was
 179 performed to synchronize all the time series at the same time step (1 minute) [32].

180 2.2 Features Selection

181 The data quality and quantity have an influence on the majority of data-driven techniques,
 182 including data mining and machine learning. Furthermore, it is important to determine which
 183 factors impact the target value (the model output) and how many features (model inputs) can
 184 be used to build predictive models. In practice, several environmental factors may influence the
 185 accuracy of window opening prediction. However, due to realistic limits, it is impossible to search

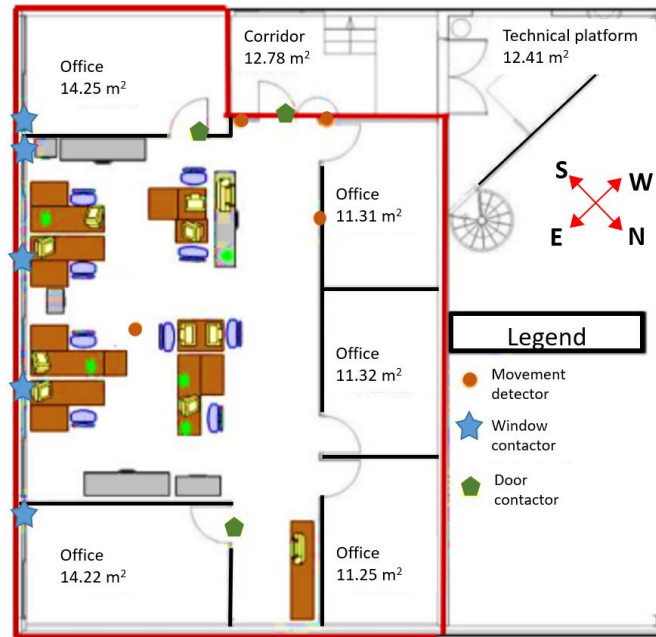


Fig. 1. The studied open-plan office layout.

186 for all of these features. According to some previous studies, the outdoor temperature, indoor
 187 CO₂ concentration and the prevailing mean air temperature were the most important variables
 188 in determining the probability of opening/closing windows, followed by indoor air temperature,
 189 outdoor and indoor humidity [4, 14, 38, 28].

190 In addition, non-environmental factors, such as: seasonal change, time of the day and personal
 191 preference, also affect the window-opening probability [25]. Thus, in our model, the following
 192 variables (features) were used as the initial input selection:

- 193 – temperature (T) and specific humidity (Hs) of both indoor and outdoor environments and
 194 the prevailing mean outdoor air temperature (PMA);
- 195 – indoor CO₂ and indoor particulate matter concentrations (PM2.5 and PM10);
- 196 – wind direction, raining condition, door status, occupancy status;
- 197 – month, day of the week, hour of the day.

198 The main statistics of the monitored environmental parameters for the years 2014 and 2015
 199 are displayed in Table 1 and Table 2, respectively. It should be noted that the comparison of these
 200 two years is not very representative as 2014 data covered the whole year, and the 2015 monitoring

201 set covered only the first 6 months. However, there are no significant differences between the
 202 averaged values of these two years. One can notice that the maximum values of PM2.5 and
 203 PM10 concentrations in 2014 are quite higher than those monitored during 2015 ($91.87 \mu\text{g.m}^{-3}$
 204 and $106.78 \mu\text{g.m}^{-3}$ in 2014 in comparison with $21.3 \mu\text{g.m}^{-3}$ and $43.71 \mu\text{g.m}^{-3}$ in 2015). This can be
 205 explained by the outdoor pollution episode of particulate matter that happened in March 2014,
 206 a quite remarkable event. In addition, higher specific humidity is observed in 2014 compared to
 207 2015, but the monitored data of 2015 does not include July to December.

Table 1. The statistics for environmental parameters of 2014

Features	Indoor CO ₂ (ppm)	Indoor PM _{2.5} ($\mu\text{g.m}^{-3}$)	Indoor PM ₁₀ ($\mu\text{g.m}^{-3}$)	Indoor T (°C)	Outdoor T (°C)	Indoor Hs (g/kg)	Outdoor Hs (g/kg)
Max value	1144.00	91.87	106.78	31.30	35.60	15.11	17.30
Min value	416.80	0.26	0.31	15.00	-4.30	4.28	3.98
Mean value	501.10	2.47	4.32	23.00	13.50	8.88	9.65
Median value	480.50	1.76	3.15	22.40	13.50	8.95	9.66
Std value	64.30	2.87	4.18	2.30	6.00	1.91	2.47

Table 2. The statistics for environmental parameters of 2015

Features	Indoor CO ₂ (ppm)	Indoor PM _{2.5} ($\mu\text{g.m}^{-3}$)	Indoor PM ₁₀ ($\mu\text{g.m}^{-3}$)	Indoor T (°C)	Outdoor T (°C)	Indoor Hs (g/kg)	Outdoor Hs (g/kg)
Max value	1038.82	21.30	43.71	33.33	39.22	13.33	14.94
Min value	421.48	0.13	0.16	18.24	-1.80	3.55	3.48
Mean value	498.45	2.50	4.45	23.10	11.28	6.44	7.11
Median value	477.02	1.93	3.40	22.30	10.30	6.22	6.69
Std value	61.38	2.11	3.70	2.43	7.01	1.55	2.09

208 In reality, the windows opening status does not change much within a given hour, hence using
 209 such a detailed database with a 1-minute time step is not necessary. In addition, some monitored
 210 data were missing, therefore we decided to use the hourly average data in this study. Based on
 211 the 1-minute time step data, the hourly average values of the selected parameters were calculated
 212 as in equation (10). A linear interpolation was applied in order to replace missing values.

$$x_{hourly} = \frac{1}{60} \sum_{i=1}^{60} x_{minute_i} \quad (10)$$

213 The window opening status for a specific hour was calculated as the mode value (most frequent)
 214 of the number of opened windows, according to the equation (11).

$$x_{hourly} = mode(x_{minute_i}) \quad (0 < i \leq 60) \quad (11)$$

215 In order to obtain more information about the monitored time series, the autocorrelation func-
 216 tions (ACF) were calculated. The ACF of a time series $Y(t)$ provides a measure of the correlation
 217 between y_t and y_{t+k} , where $k = 0, \dots, K$ ($k \in \mathbb{Z}$, K is not larger than $T/4$, where T is the total
 218 number of observations) and y_t is assumed to be the realization of a stochastic process. According
 219 to [5], the autocorrelation r_k for lag k is:

$$r_k = c_k / c_0 \quad (12)$$

220 where:

$$c_k = \frac{1}{T} \sum_{t=1}^{T-k} (y_t - \bar{y})(y_{t+k} - \bar{y}) \quad (13)$$

221 and c_0 is the sample variance, \bar{y} is the sample mean of the time series.

222 The ACF results of the environmental data monitored during 2014 are represented in the
 223 Figure 2. Very similar results were obtained for data of the year 2015 so they are not presented
 224 here. One can notice the persistence of the temperature (T) and specific humidity (Hs) indoors
 225 and outdoors, which means that a value at time t of the temperature or specific humidity is
 226 correlated to a value one day later ($t+24$), two days later ($t+48$), or even three days later
 227 ($t+72$). In addition, the ACF of the CO₂ concentrations becomes negative and remains at low
 228 levels, and then switches back to positive values after a lag of 17 hours. While for outdoor T and
 229 Hs (indoors and outdoors), the autocorrelations persist in the positive domain for long delays. In
 230 general, temperatures depict the same structures of spectral variability as CO₂: the fundamental
 231 frequency peaks at every 24 hours. The ACF of CO₂ alternates sign every 8 hours on a lag of 24
 232 hours. This implies that, instead of using the information of the 'previous hour', in the real-time
 233 system, we could use the value of 'the previous 24th hour' ($t-24$) environmental data as input for
 234 this model, which is easier to access.

235 Furthermore, the ‘weekly periodicity’ (at the lag of 168 hours) in the ACF values of CO₂ and
 236 PM10 concentration is noteworthy. The information of the ‘previous 168th hour’ data could be
 237 then used as input for the model when the ‘previous 24th hour’ data is not available. Besides, it
 238 can be also noticed that the ACFs of PM concentrations and number of opened windows present
 239 high values at a lag of 24 hours (see Figure 2d). We decided to use also the PM concentrations
 240 and the number of opened windows, corresponding to the 24 hours lag, as inputs of the prediction
 241 model.

242 In conclusion, non-environmental, environmental features and window status of **the previous**
 243 **24th hour** moment, were selected as initial inputs of a model built in order to predict the opening
 244 status of windows at the **current hour** as presented in the next section.

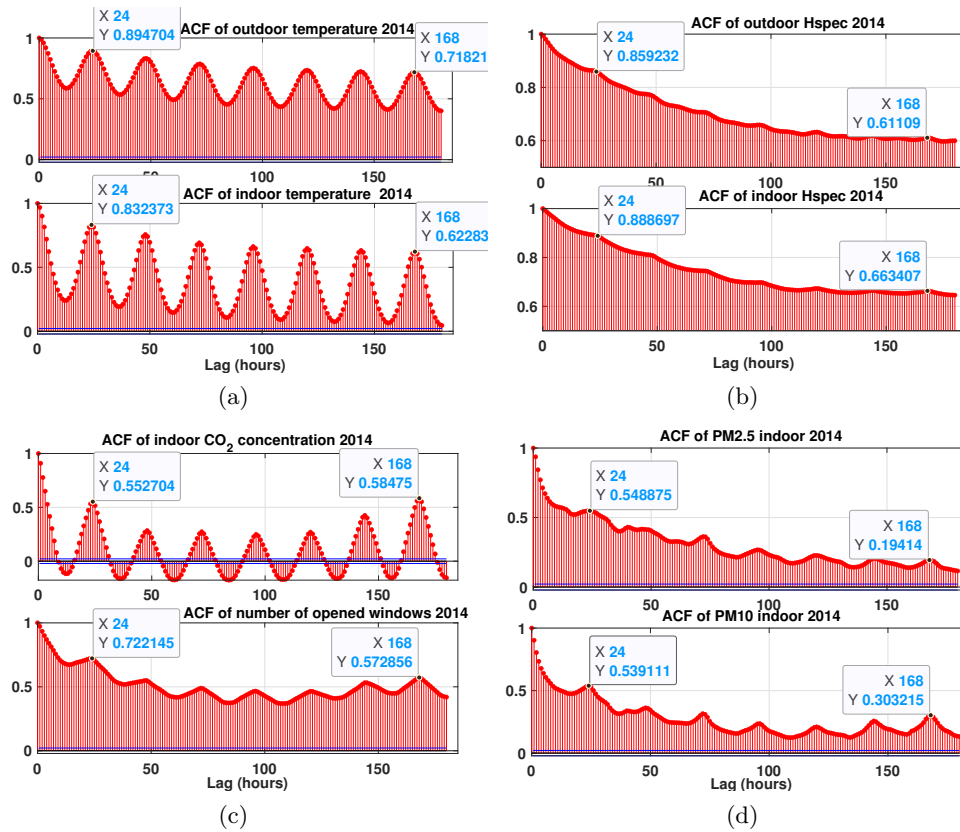


Fig. 2. Autocorrelation values of environmental variables in 2014: (a) Indoor and outdoor temperature, (b) indoor and outdoor humidity, (c) indoor CO₂ and number of opened windows, and (d) indoor PM2.5 and PM10. The 24-hour and 7-day peaks are indicated on the plot of each ACF (X represents the lag and Y represents the ACF value).

245 3 Modeling implementation

246 In this section, the different ML models (Decision Tree, kNN classification, Kernel Approxima-
247 tion) are briefly introduced, followed by the data pre-processing and finally the models parame-
248 terization.

249 3.1 Models Description

250 **Decision Tree [29]:** Decision Tree is a supervised ML Algorithm that employs a set of rules to
251 make decisions in the same way that people do. Some classification methods, such as Naïve Bayes,
252 are probabilistic, although a rule-based technique is also available. The idea behind Decision Tree
253 is to use dataset attributes to create binary yes/no questions, and then segment the dataset until
254 all the data points from each class become isolated. With this strategy, one can organize the
255 data in a tree structure. A node is added to the tree when a question is asked. Furthermore, the
256 first node is known as the root node. The answer to a question separates the dataset and creates
257 new nodes based on the value of a characteristic. If the process is stopped after a split by some
258 conditions (for example: stop splitting if more than 95% belong to a single class, stop splitting
259 if less than 5 individuals, do not split if the new node has less than 5 individuals, . . .), the final
260 nodes are known as leaf nodes.

261 The algorithm attempts to partition the dataset into the lowest subset feasible at each split.
262 The aim, like with any other Machine Learning method, is to minimize the loss function as
263 much as feasible [34]. Stochastic Gradient Descent is a popular loss function for classification
264 algorithms. Given that the loss function should be differentiable, it is not possible to use in this
265 circumstance. However, because data points from distinct classes have to be separated, the loss
266 function should assess a split based on the proportion of data points from each class before and
267 after the split. In other words, a loss function that assesses the split based on the cleanliness of
268 the resultant nodes is desirable. Examples of loss functions that compare the class distribution
269 before and after a split are Gini Impurity and Entropy [34].

270 To summarize, Decision Tree is a rule-based method for solving classification and regression
271 tasks. There is an obvious trade-off between interpretability and performance. A small tree is
272 simple to perceive and comprehend, but it contains a lot of variation. A little modification in

273 the training set can result in an entirely different tree and predictions. A large tree with several
 274 splits, on the other hand, produces better classifications. However, it is most likely to remember
 275 the training dataset (overfitting).

276 **k-Nearest Neighbor classification [15]:** k-Nearest Neighbors models are a type of instance-
 277 based model that is used mainly for classification in the Machine Learning field. Its fundamental
 278 is as follows: similar objects exist in close proximity. The basic steps of the k-NN algorithm for
 279 classification are described below:

- 280 1. Load the data
- 281 2. Initialize k to your chosen number of neighbors
- 282 3. For each sample in the data, calculate the distance between the query sample and the current
 283 sample from the data by using distance calculation algorithms (such as Euclidean, Chebyshev,
 284 City Block, etc).
- 285 4. Return the mode (the value that appears the most often) of k nearest (smallest distance)
 286 neighbors.

287 The k-NN classification is recommended as 'a theoretically optimal method of classification' [17].
 288 However, this method is not easy to interpret and it does not offer the possibility to extract a
 289 rules set in order to apply it to another dataset. In addition, the k-NN classification cannot deal
 290 with both numerical and categorical data at the same time. It is required to convert numerical
 291 data to categorical data.

292 **Kernel Approximation [30]:** Kernel approximation is an effective technique for overcoming
 293 the low scalability of kernel-based techniques by establishing an explicit mapping $\psi: R^d \rightarrow R^s$
 294 such that $K(x, y) \approx \psi(x)^T \psi(y)$. By doing so, an efficient linear model can be well learned in the
 295 transformed space with $O(ns^2)$ time and $O(ns)$ memory while retaining the expressive power of
 296 nonlinear methods, where n is the number of samples in the original d -dimensional space and s
 297 is the number of features, which is normally a very high number.

298 The Random Features is one of the most popular techniques to speed up kernel methods
 299 in large-scale problems. The Random Kitchen Sinks [30] and Fastfood [36] are two examples of
 300 random feature expansions, these schemes tried to approximate Gaussian kernels of the kernel
 301 classification algorithm to use for big data in a computationally efficient way. Firstly, they find

302 a random transformation so that its dot product approximates the Gaussian kernel. That is:

$$K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle \approx T(x_1)T(x_2)' \quad (14)$$

303 where $T(x)$ maps x in R^p (p is the number of input features) to a high-dimensional space (R^m).

304 The Random Kitchen Sinks scheme uses the random transformation

$$T(x) = m^{-1/2} \exp(iZx)' \quad (15)$$

305 where $Z \in R^{m \times p}$ is a sample drawn from $N(0, \sigma^{-2})$ and σ^2 is a kernel scale. This scheme requires

306 $O(mp)$ computation and storage.

307 The Fastfood scheme introduces another random basis V instead of Z using Hadamard ma-

308 trices combined with diagonal Gaussian scaling matrices.

$$V = \frac{1}{\sigma\sqrt{d}} SHGIIHB \quad (16)$$

309 where $\Pi \in \{0, 1\}^{d \times d}$ is a permutation matrix and H is the Walsh-Hadamard matrix. S, G and B

310 are all diagonal random matrices. When the implemented function uses the Fastfood scheme for

311 random feature expansion and uses linear classification to train a Gaussian kernel classification,

312 the model only needs to form a matrix of size $n \times m$, with m typically much less than n for big

313 data, in comparison with support vector machine that requires computation of the $n \times n$ Gram

314 matrix. This random basis reduces the computation cost to $O(m \log p)$ and reduces storage to

315 $O(m)$.

316 3.2 Data pre-processing

317 After recalculating the number of opened windows for a specific hour using the mode value

318 (equation (11)), these values were then categorized into four different groups, labeled as follows:

- 319 – ALL CLOSED: all of the windows are closed ($x_{hourly} = 0$)
- 320 – MOSTLY CLOSED: 1 window is opened ($x_{hourly} = 1$)
- 321 – MOSTLY OPENED: 2 or 3 windows are opened ($2 \leq x_{hourly} < 4$)

322 – ALL OPENED: 4 windows or more are opened ($x_{hourly} \geq 4$)

323 The office is equipped with five windows. In 2015, one window sensor was out of order, thus the
 324 respective window remained closed all the time. Therefore, the maximum number of opened win-
 325 dows is five in 2014 and four in 2015. The distribution profiles according to the non-environmental
 326 parameters (month, day of the week and hour of the day) and the initial statistics of these four
 327 groups during the years 2014 and 2015 are displayed in Figure 3 and Figure 4, respectively.

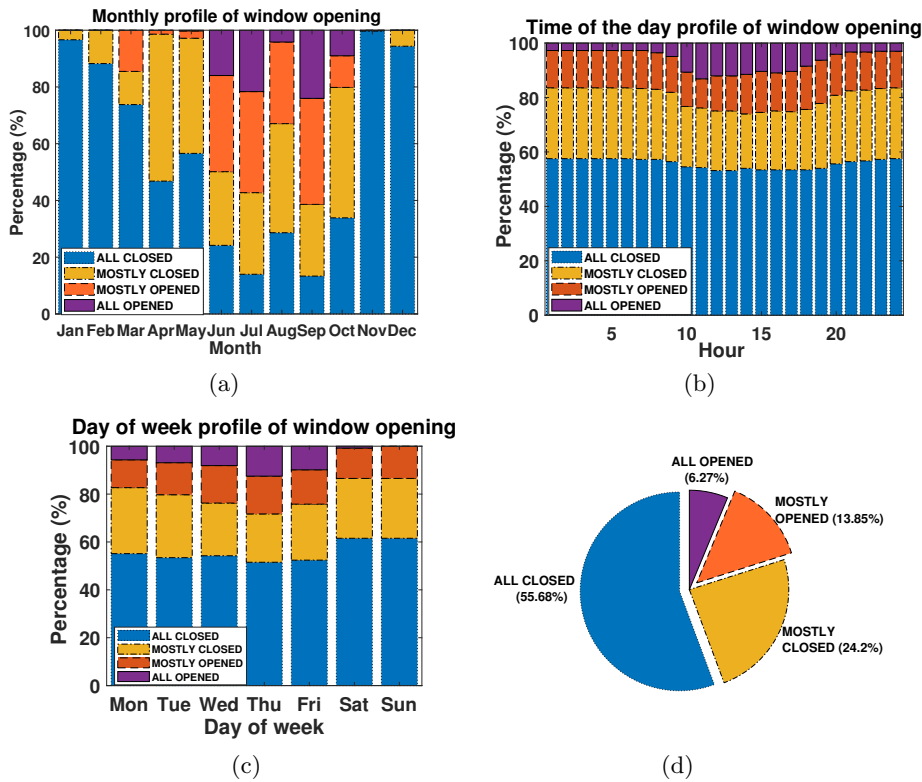


Fig. 3. Distribution profile of window opening during 2014 according to the (a) Month, (b) Hour of the day and (c) Day of the week and (d) Statistics for the window opening categories.

328 Figure 3d shows that in 2014, for more than half of the time (55.68%), the status of this group
 329 of windows is ‘ALL CLOSED’. This label is dominant during the winter period (November –
 330 March). ‘MOSTLY CLOSED’ and ‘MOSTLY OPENED’ labels are quite equally distributed with
 331 24% and 14%, respectively. The fourth label ‘ALL OPENED’ accounts for just 6.3% of the total
 332 time and it appears only in summer and the beginning of autumn (June – October) and during

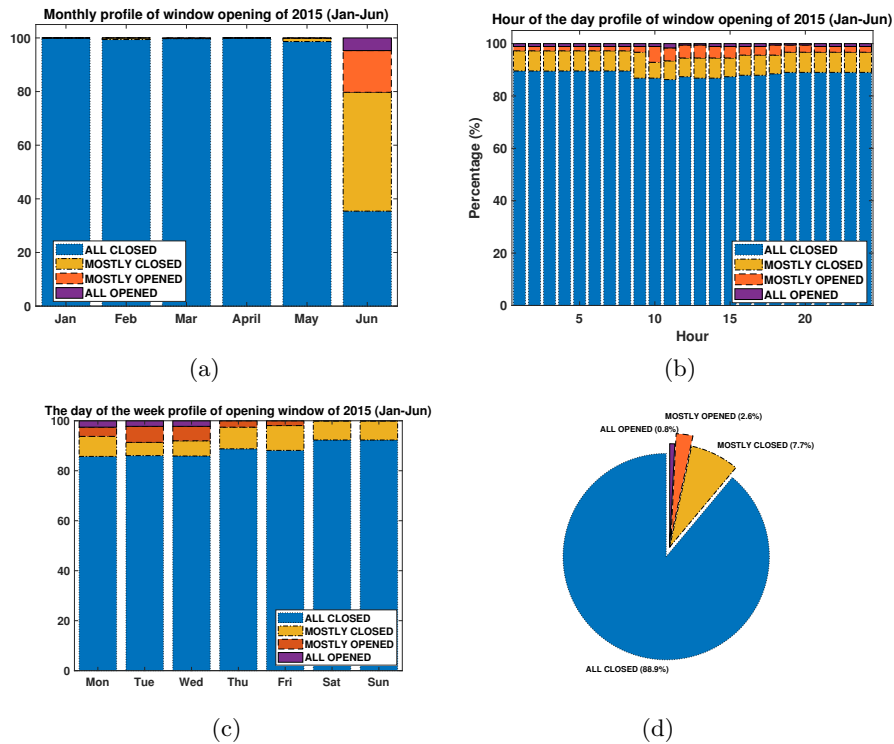


Fig. 4. Distribution profile of window opening of 2015 according to the (a) Month, (b) Hour of the day and (c) Day of the week and (d) Statistics for the window opening categories.

333 the working time (9 a.m. – 6 p.m.). This is expected because “during the working time, the
 334 occupants tend to open at least one window, and rarely open the full five windows at the same
 335 time” [32].

336 The statistics for the window opening state according to categories show in 2015 even a higher
 337 percentage (88.9%) of the “ALL CLOSED” label. The “ALL OPENED” label is obtained only
 338 in June with 0.8% for the 6-month period. The “ALL CLOSED” profile can be observed almost
 339 all the time from January to April (Figure 4a). This is quite different in comparison with the
 340 distribution profile of the year 2014 without an obvious reason.

341 Regarding the environmental parameters, Figure 5 represents the mean values and standard
 342 deviations of these variables according to the groups. Differences in the mean values of outdoor
 343 temperature, specific humidity (indoors and outdoors) and PM10 indoors can be observed for
 344 the four windows categories (Figure 5a,b and d). For these parameters, the higher the value,
 345 the greater number of windows are opened. For indoor temperature and PM2.5 the differences

346 between groups are small. The indoor mean CO₂ concentration keeps a stable value among these
 347 four groups (Figure 5c). Given that the measurement uncertainty is 50 ppm ± 3% for reading,
 348 the range of variation 480-520 ppm is less than the uncertainty. So, we can consider that the
 349 CO₂ value does not vary significantly, which means that the office is “well ventilated”.

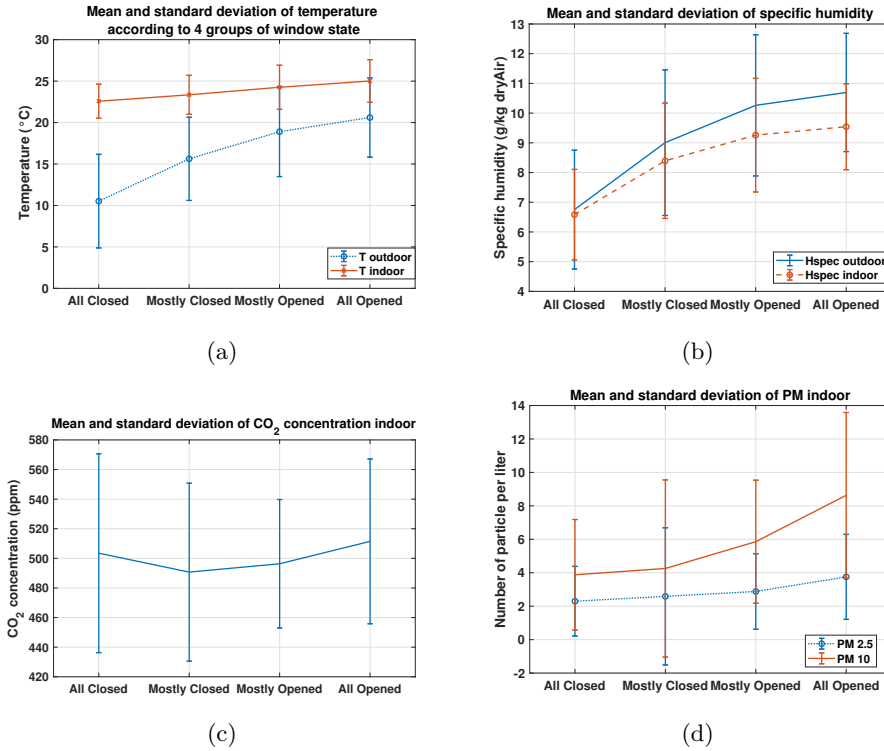


Fig. 5. Statistic profile of 4 groups of window opening during 2014 according to (a) Temperature, (b) Specific humidity (c) CO₂ concentration and (d) PM concentration.

350 For the model implementation, we need different data sets: training, validation, testing, etc.
 351 We decided to divide the time series data into sets of 25 hours and use the 20 first hours for
 352 training and validation, and the remaining 5 hours for testing (ratio 80:20 – see Figure 6). The
 353 reason why we did not use the day 365th for training is that we need the windows status of this
 354 day to evaluate the testing set of the 364th day (‘previous 24th hour’). In total, 6980 hours were
 355 used for training.

356 As k-NN method can not deal with numerical and categorical data at the same time, quan-
 357 titative data had to be recoded to generate qualitative (categorical) data. Numerical data were

358 obtained from environmental parameters monitoring; in order to be transformed into categori-
 359 cal data, the values of each variable were divided into 10 groups (or categories) based on their
 360 percentiles in order to equally represent the groups. The first 10 percentiles belong to the first
 361 group, the data of percentiles from 11 to 20 belong to the second group, and so on.

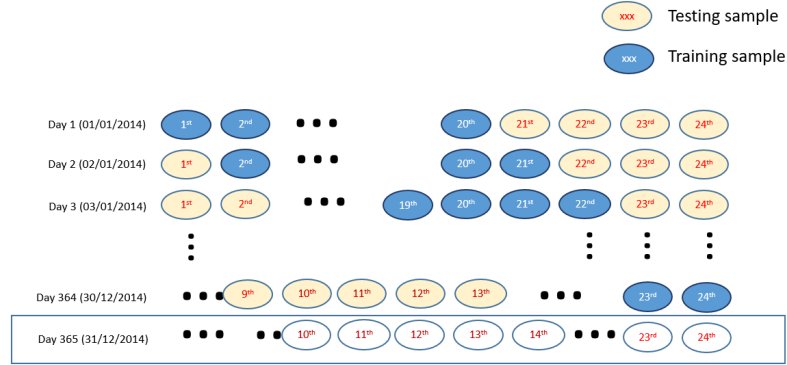


Fig. 6. Figure explaining how we split the data into training and testing sets (sets of every 25 hours).

362 3.3 Models parameterizations

363 The Classification Learner application of Matlab[®] was used for the model development. The
 364 'OptimizeHyperparameters' option for 'all' the input parameters was used to obtain the best
 365 values for the hyperparameters of the models and to avoid overfitting. This optimization attempts
 366 to minimize the cross-validation loss (error) by varying the parameters. The summary of the
 367 obtained values of the different hyperparameters for the three models are presented in Table 3.

368 The other general parameters of the models are listed below:

- 369 – Number of data – training set: 6980 samples (80% data of 2014)
- 370 – Number of data – testing set:
 - 371 • Testing set of 2014 (which will be called 'test set 2014'): 1745 samples (the rest of 20%
 372 data of 2014)
 - 373 • Testing set of 2015 (which will be called 'test set 2015'): 4345 samples (data from January
 374 to June 2015)

Table 3. Summary of the different hyperparameters for the three models.

Algorithm	Hyperparameter	Value
Decision Tree	Maximum number of Splits	4454
	Split Criterion	deviance
	Minimum leaf size	1
	Tree Depth	16
k Nearest Neighbor	Number of neighbor (k)	3
	Distance metric function	hamming
	Standardize	true
Kernel Approximation	Kernel function	polynomial
	Polynomial Order	3
	Standardize	true

- 375 – Data type: hourly averaged data
- 376 – Validation method: 10-fold cross validation
- Initial number of input variables: 16 variables as in Table 4:

Table 4. Summary about the input variables for the predicting model.

Idx	Name	Value	Type of data	Moment
1	Month	month	categorical	Current moment
2	DoW	day of the week	categorical	Current moment
3	HoD	hour of the day	categorical	Current moment
4	T_out	outdoor temperature	numerical/categorical ^a	previous 24 th hour
5	T_in	indoor temperature	numerical/categorical ^a	previous 24 th hour
6	Hs_out	outdoor specific humidity	numerical/categorical ^a	previous 24 th hour
7	Hs_in	indoor specific humidity	numerical/categorical ^a	previous 24 th hour
8	CO2_in	indoor CO ₂ concentration	numerical/categorical ^a	previous 24 th hour
9	PM2.5in	indoor PM2.5 concentration	numerical/categorical ^a	previous 24 th hour
10	PM10in	indoor PM10 concentration	numerical/categorical ^a	previous 24 th hour
11	Prv_Wd	state of group of windows	categorical	previous 24 th hour
12	PMA	prevailing mean outdoor air temperature	numerical/categorical ^a	previous 24 th hour
13	WindD	wind direction	categorical	previous 24 th hour
14	Rain	raining status	categorical	previous 24 th hour
15	Occ	occupancy status	categorical	previous 24 th hour
16	Door	entrance door status	categorical	previous 24 th hour

^a This variable is coded in 10 categories for the k-NN classification model. For Kernel Approximation and Decision Tree, the monitored numerical data is kept as original.

378 4 Results and discussion

379 4.1 Rank of the important scores of predictors

380 Because input variables have a direct influence on the model predictive performance, it is es-
381 sential to determine which variables are the most important for the model development. The
382 input selection is based on the relevance of the different predictors by evaluating the relative
383 contribution of a given input to the performance of a particular model. This approach is called
384 model-dependent and the advantage of this method is that the input selection is strongly related
385 to the model performance, giving useful information for building predictive models.

386 Figure 7 shows the relative importance of the factors for window opening status prediction
387 by using the Decision Tree model. Similar results were obtained for the other two methods (k-
388 NN and Kernel Approximation) and will not be presented here. This figure shows the relative
389 significance of the categorical variables (month, day of the week, hour of the day, and the pre-
390 vious 24th hour windows state), as well as the previous 24th hour value of the prevailing mean
391 outdoor temperature outdoors (PMA). According to this observation, these parameters are the
392 most important ones for this modeling. Surprisingly, an important influencing factor - the out-
393 door temperature, has a small effect on the model's performance. This can be explained by the
394 substantial impact of the specific humidity and PMA, which are calculated using the outside
395 temperature value as in the equations (1) and (9). The rain condition and the status of occu-
396 pancy show very low importance. Based on this result, we decided to implement the models
397 without these two parameters (Rain and Occupancy). In conclusion, 14 parameters were selected
398 as inputs for our predicting models: Month, DoW, HoD, T_out, T_in, Hs_out, Hs_in, CO2_in,
399 PM2.5_in, PM10_in, Prv_Wd, PMA, WinD, Door.

400 4.2 Performance of the window opening state model

401 Data monitoring starts on the 1st of January 2014 and ends on the 30th of June 2015 (13104
402 samples-hours). We have decided to use 80% data of the year 2014 for the training and validation
403 set (6980 samples). The remained data was divided into 2 sets for testing: (i) the rest of 20%
404 of the data of the year 2014 (1745 samples) and (ii) data from January 2015 - June 2015 (4345
405 samples), because we want to observe the different behaviors of the built model when it has to

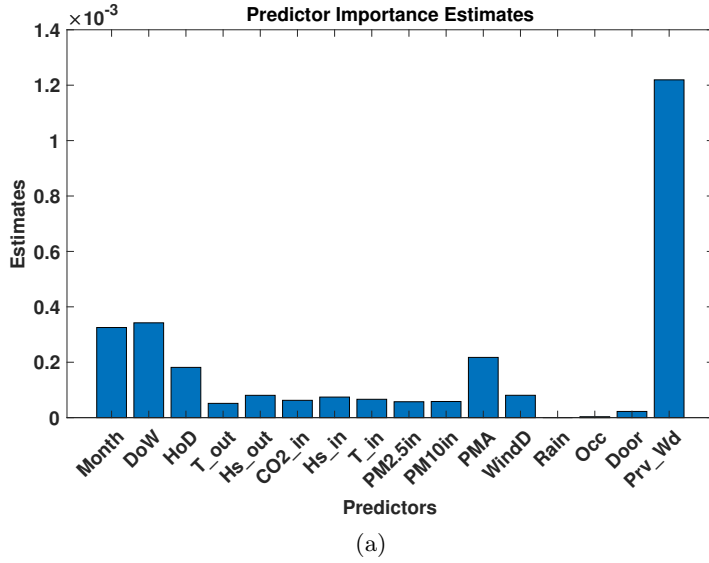


Fig. 7. Predictors importance for predicting window opening status for a DT with the input containing all the available parameters. The Month, DoW and HoD correspond to the current moment, all the other variables correspond to the previous 24th hour (see table 4).

406 deal with data of the same period (the same year 2014) and with data from a completely new
 407 period (data of 2015).

408 **Performance of the Decision Tree classifier**

409 Based on the results of the hyperparameters optimization presented in the table 3, a Decision
 410 Tree of 541 nodes (Tree Depth = 16) has been obtained after using 80% data of the year 2014 for
 411 training and validation, with accuracies of 98.09% and 89.81%, respectively. Using this trained
 412 decision tree, we predicted the testing set containing the rest of 20% of the data of 2014 and
 413 then we compared it to the monitored values. A value of 86.36% for accuracy (% of well-classified
 414 data) was achieved for this test. A confusion matrix of the Decision Tree method for this testing
 415 set is displayed in Figure 8a.

416 As we can see from the figure 8a, the model has a tendency of mislabeling one sample as
 417 a 'neighbor label'. The explanation for this could be that the environmental factors change
 418 gradually, the 'ALL OPENED' and 'ALL CLOSED' states are easily identifiable, but the 'ALL
 419 CLOSED' and 'MOSTLY CLOSED' ones can be ambiguous. The decision tree achieves 910
 420 correct predictions and misses 58 (31+24+3) when the true label is 'ALL CLOSED'; 31 samples

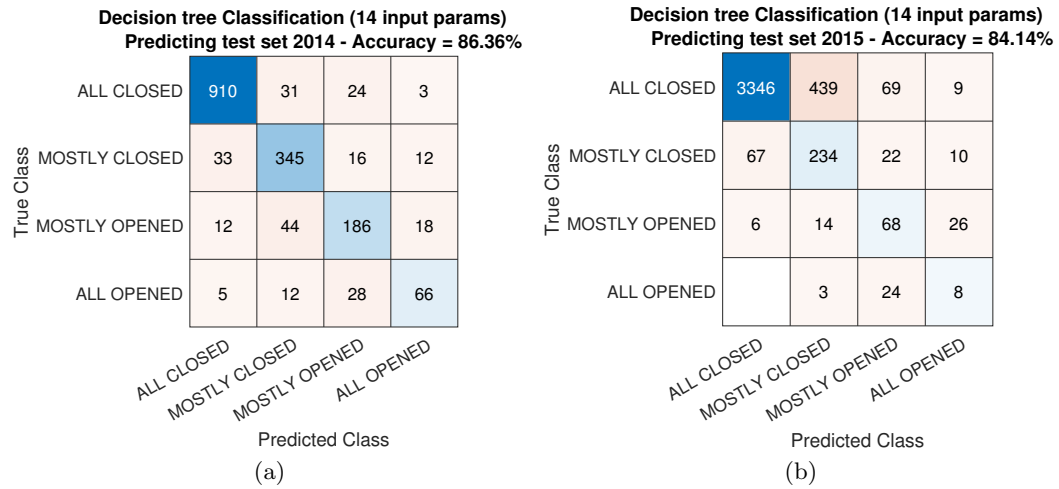


Fig. 8. Confusion matrix of Decision Tree classification (14 input parameters - including information about wind direction and door status) for (a) test set 2014 (1745 samples) and (b) test set 2015 (4345 samples).

421 were incorrectly predicted to be in the 'MOSTLY CLOSED' state, 24 samples were wrongly
 422 labeled as 'MOSTLY OPENED,' and 3 samples were misclassified as 'ALL OPENED.' Similarly,
 423 when the true label is 'MOSTLY CLOSED,' 345 samples are properly predicted whereas 61 are
 424 incorrectly classified (33+16+12). The labels 'MOSTLY OPENED' and 'ALL OPENED' are
 425 accurately predicted in 186 and 66 examples, respectively.

426 Using the same trained Decision Tree classifier, we predicted the window status of the first 6
 427 months from January to June, of 2015, and compared them to the monitored values. A value of
 428 84.14% for accuracy was achieved.

429 The confusion matrix for this testing set (data of 2015) is displayed in Figure 8b. Similar to
 430 the test set 2014, the true label 'ALL CLOSED' has the highest number of right predictions when
 431 the model successfully labeled 3346 samples and mislabeled 517 samples. The label 'MOSTLY
 432 CLOSED' also ranks second with 243 accurate samples, and 'MOSTLY OPENED' follows in the
 433 third position with 68 correctly classified samples. Specifically, the model can properly identify
 434 just 8 samples of the 'ALL OPENED' label while misclassifying up to 24 samples as 'MOSTLY
 435 OPENED'. The more detailed evaluation of these confusion matrices will be discussed in the
 436 next section.

437 Performance of the k-NN classifier

438 Regarding the k-NN classification model, k=3 was obtained after the hyperparameters opti-
 439 mization (see table 3). The achieved accuracies were 99% for training and 92.3% for validation.

440 The confusion matrix obtained on the test set 2014 is displayed in Figure 9a. This model
 441 obtained a value of overall accuracy of 86.53%. From the figure, the highest number of wrong
 442 classified belongs to the "MOSTLY OPENED" label, while 40 samples are wrongly predicted
 443 as "ALL OPENED". Similar to the Decision Tree model, 'ALL CLOSED' label achieved the
 444 highest performance, 96.2% sample of this label were correctly predicted (931 corrects from a
 445 total of 968 samples). The 'MOSTLY CLOSED' label got the second rank with 84.7% correctly
 446 predicted samples (344 correct from a total of 406 samples). Finally, the 'MOSTLY OPENED'
 447 and 'ALL OPENED' labels rank the last as they have only 66.2% and 56.8% correct predictions,
 448 respectively.

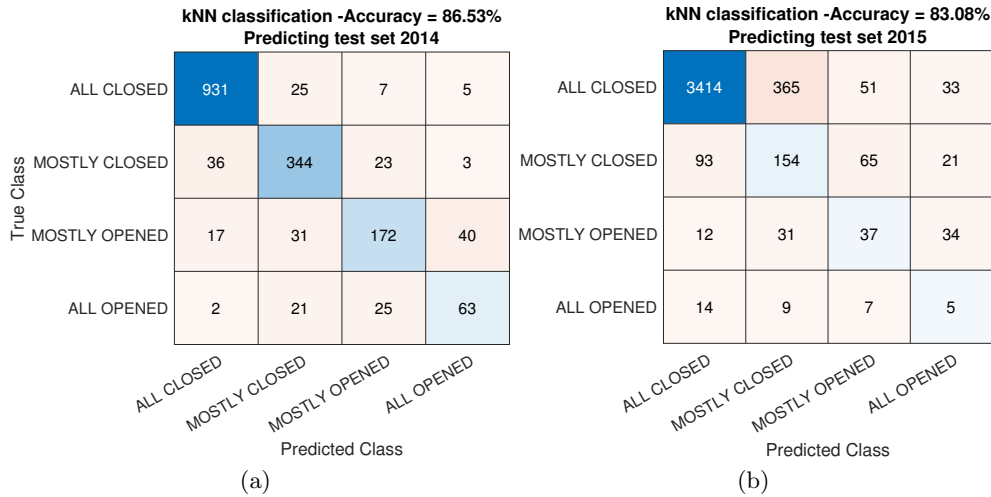


Fig. 9. Confusion matrix of k-NN classification (14 input parameters) for (a) test set 2014 (1745 samples) and (b) test set 2015 (4345 samples).

449 Similarly, the confusion matrix for the same trained k-NN model applied on the test set 2015
 450 is represented in Figure 9b.

451 Same as the Decision Tree model results for the test set 2015, one can observe that a signifi-
 452 cant number of "ALL CLOSED" labels are misclassified as "MOSTLY CLOSED" (365 samples
 453 - 9.4%). Eventhough, "ALL CLOSED" label still achieved the highest number of correct classifi-
 454 cations (88.4% - 3414 correct predictions out of 3863 total samples). The "MOSTLY CLOSED"

455 and "MOSTLY OPENED" achieved their ranks as second and third with 46.2% and 32.5%, re-
 456 spectively. The 'ALL OPENED' label, again, got the last position with only 5 correct predictions
 457 (14.3%).

458 **Performance of the Kernel Approximation classifier**

459 The polynomial kernel function of order 3 has been obtained after the hyperparameter op-
 460 timization. In comparison with the two other classification models, when using the Kernel Ap-
 461 proximation classifier, the training accuracy results were even lower: only 81.7% for training and
 462 80.6% for validation.

463 The confusion matrices for Kernel Approximation classifications for the years 2014 and 2015
 464 are displayed in Figure 10. While the accuracy was only 79.3% for the test set 2014, this method
 465 achieved up to 92.9% for the test set 2015. Similar to the two other models, this model also has
 466 a tendency of mislabeling one sample as a 'neighbor label'. According to the Figure 10, Kernel
 467 Approximation misclassified the "MOSTLY CLOSED" as "MOSTLY OPENED" quite a lot (60
 468 samples) and vice versa (58 samples). For the testing set of 2015, the same mistake also was
 469 showed when 75 samples were mislabeled as 'MOSTLY CLOSED' and up to 82 samples were
 470 wrongly classified as 'MOSTLY CLOSED' instead of 'ALL CLOSED'.

471 It is interesting to note that the Kernel Approximation method has a different rank of correct
 472 predictions among labels in comparison with the two other models for the test set 2015. For
 473 test set 2015, the true label 'ALL CLOSED' still has the highest number of right predictions
 474 (97.3%), however, the 'MOSTLY OPENED' (42.9%) and 'MOSTLY CLOSED' (36%) labels
 475 switched their ranks as second and third, respectively. 'ALL OPENED' label, again, has the
 476 last position. Specifically, this method has the highest correct predictions for the label "ALL
 477 OPENED" of test set 2015 with up to 15 samples on a total of 33.

478 **4.3 Accuracy statistics for the Decision Tree model**

479 For a deeper analysis of the results, it is necessary to analyse the detailed statistics of the accuracy
 480 according to the day of the week, the hour of the day, and the month. We decided to present in
 481 this subsection only the results obtained for the Decision Tree model because for the other two
 482 models, they keep the same global trend.

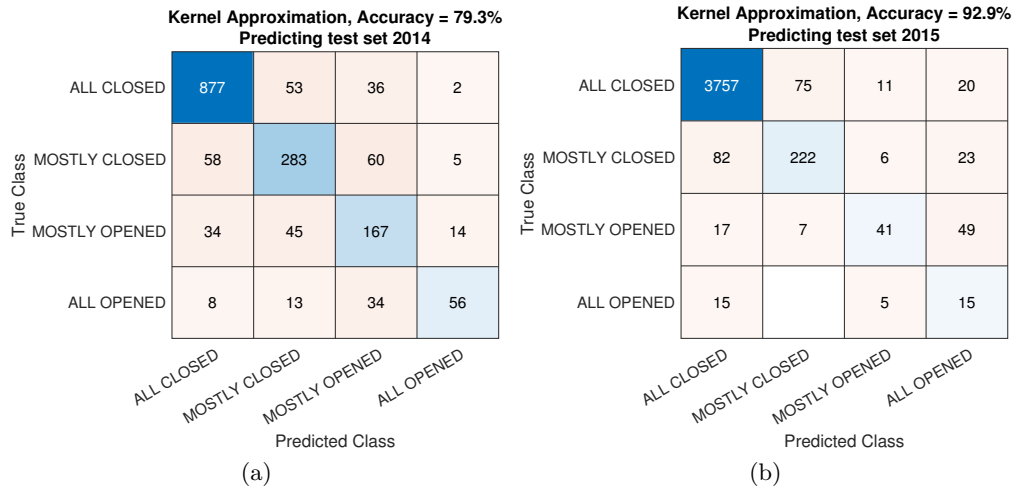


Fig. 10. Confusion matrix of Kernel Approximation classification (14 input parameters) for (a) test set 2014 (1745 samples) and (b) test set 2015 (4345 samples).

483 The statistics for the test set 2014 are showed in the Figure 11. The highest accuracies were
 484 obtained when predicting the windows state for Saturday (100%), winter season (October –
 485 February, more than 90%) and night-time periods (8 p.m. – 7 a.m., more than 88% except for
 486 the 11 p.m. when maybe the guard round took place). This is expectable because the windows
 487 are mainly closed during this time.

488 The lower accuracy values correspond to the months of the summer season (June – September,
 489 around 70%, except for August 79% - the month of vacation), lunch-time periods (12 a.m.– 2
 490 p.m. around 76%) and the ‘office leaving’ hour (5 p.m. - 73%). In all these periods, there are
 491 more changes in the status of the windows and they mostly contain the labels ‘ALL OPENED’
 492 and ‘MOSTLY OPENED’. Interestingly, Tuesday and Sunday have the lowest values of accuracy
 493 (around 81%).

494 Similarly, the statistics for the Decision Tree accuracy for the test set 2015 are presented
 495 in Figure 12. The results show that: Saturday (97%), January (98%) and night-time periods
 496 (10 p.m. – 7 a.m., around 88%) obtained the highest values of accuracy. In contrast, the lowest
 497 accuracy values correspond to the month of June (the only month that has ”ALL OPENED”
 498 status in 2015), day-time periods (9 a.m.–6 p.m.) and the working days (Monday to Friday).
 499 Tuesday, again, has the lowest value of accuracy (only 79%). According to the hour of the day,
 500 the prediction accuracy at 5 p.m. is still the lowest (75%), probably because it corresponds to

501 the “office leaving” hours. Some people tend to close the windows before leaving while others
 502 leave them opened.

503 4.4 Evaluation and Discussion

504 While the Accuracy can be used to evaluate the model’s percentage of well-classified data, Recall
 505 and Precision are two other important indicators to evaluate the performance of classification. In
 506 addition, the F1 coefficient has been used for evaluating the model’s predictive performance by
 507 combining the results from both Recall and Precision. The quality of a classifier can be evaluated
 508 by these indicators, which are calculated using the true positive (TP), the true negative (TN),
 509 the false positive (FP) and the false negative (FN), based on the equations (17 - 20).

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (17)$$

510

$$Sensitivity(Recall) = TP/(TP + FN) \quad (18)$$

511

$$Precision(F_{rate}) = TP/(TP + FP) \quad (19)$$

512

$$F1 = 2(Recall)(Precision)/(Recall + Precision) \quad (20)$$

Table 5. Summary about the overall accuracy of the three models

Algorithm	Test set 2014	Test set 2015
Decision Tree	86.36	84.14
k Nearest Neighbor	86.53	83.08
Kernel Approximation	79.30	92.90

513

514 Table 5 summarises again the general accuracy values of the three methods: Decision Tree, k-
 515 NN and Kernel approximation, when predicting the test set 2014 and the test set 2015. Decision
 516 Tree and k-NN obtained quite the same performance achieving similar results for the two testing
 517 sets (around 84%). Meanwhile, the Kernel Approximation achieved a significant higher accuracy
 518 when predicting the data of 2015. The fact that Kernel Approximation model’s accuracy when
 519 predicting the test set 2014 is lower than predicting the test set 2015 can be explained by the

520 particular distribution of labels in 2015 and by the high performance of this method for separation
 521 in the case of nonlinear problems.

522 Figure 13 represents the calculated Recall (Sensitivity) values for each state of window opening.
 523 For the test set 2014, one can notice that the three models give quite similar results, slightly
 524 lower for the Kernel Approximation method. While the highest Recall value is obtained when
 525 predicting the 'ALL CLOSED' state of the group of windows ($\approx 90\%$), the lowest value cor-
 526 responds to the 'ALL OPENED' label ($\approx 60\%$). Similarly for test set 2015, the highest Recall
 527 value is still obtained when predicting the 'ALL CLOSED' label (90%) while the lowest belongs
 528 to the 'ALL OPENED' label (excepting the Recall value obtained by the Kernel Approximation
 529 method for test set 2015, where the lowest value belongs to 'MOSTLY OPENED' label).

530 Figure 14 and Figure 15 represent the Precision values and F1 scores, respectively. The same
 531 situation is obtained for both testing sets. While the highest values are obtained when predicting
 532 the 'ALL CLOSED' state, the lowest values correspond to the 'ALL OPENED' label (excepting
 533 the Precision value obtained by the Kernel Approximation method for the test set 2014, where
 534 the lowest value belongs to 'MOSTLY OPENED' label). For the test set 2015, regarding the
 535 Precision values, an even lower value of 5.4% is observed for the 'ALL OPENED' label, by the
 536 k-NN model. The reason for which the model's accuracy when predicting the 'ALL OPENED'
 537 label was much lower than for the 'ALL CLOSED' label is the particular distribution of labels
 538 during the two years. The windows are mainly "ALL CLOSED" and this label is "well learned"
 539 by the model. Window opening models are often biased towards the over-represented class where
 540 windows remained closed [22].

541 In general, the Accuracy gave us an overall result without the information about a specific
 542 label. Meanwhile, in the case of Recall and Precision indicators we got a detailed accuracy for
 543 each label in different perspectives: Precision - How many predicted samples of this label are
 544 correct? Recall - How many samples of this label are correctly predicted? From the Figure 13,
 545 one can observe the significant differences in Recall values of Kernel Approximation for 'ALL
 546 OPENED' label and Decision Tree for 'MOSTLY OPENED' label of test set 2015. Similarly,
 547 figure 14 reveals the high differences in Precision values of 'MOSTLY OPENED' and 'MOSTLY
 548 CLOSED' label for the test set 2015 when using the Kernel Approximation. However, when we
 549 calculated the F1 values, these differences were smaller.

550 Overall, the Decision Tree method appears to be the best classification model, with the best
551 balance of Recall, Precision and F1 values regarding the four labels. Kernel Approximation oc-
552 casionally achieved the highest evaluation values (particularly for the test set 2015 for 'ALL
553 CLOSED' and 'ALL OPENED' labels). This can be explained by its high performance in separa-
554 tion in the case of nonlinear problems. However, the overall accuracy for the test set 2014 of
555 this method is slightly lower in comparison with the two other methods. In addition, Decision
556 Tree also provides the list of classification rules (export in .txt file), which can easily be used
557 to apply for new data. Regarding the kNN model, the low values of these evaluation indicators
558 could be explained by the fact that this method has been applied on categorical data for all the
559 parameters, by contrast to the other methods, which allow the both types of inputs (numerical
560 and categorical). This decoding operation probably leads to a loss of information.

561 **5 Conclusion and Future work**

562 In conclusion, in this study, we have obtained three ML classification models to predict the
563 opening state for a group of windows in an open-plan office. To select the appropriate set of
564 features, the ACF values and predictor importance estimates were calculated. In our case, the
565 most pertinent inputs were: the previous 24th hour state of the windows (which can be related to
566 the personal preferences of the occupants), the day of the week, the month, the hour of the day
567 (which can be related to the occupancy and the personal preferences) and the previous 24th hour
568 of the prevailing mean outdoor air temperature (outdoor environment condition). The models
569 were then established by using these important parameters completed with the 'previous 24th
570 hour' of the following variables: the wind direction, entrance door status, indoor CO₂ and particle
571 matter (PM2.5 and PM10) concentrations, as well as both indoor and outdoor temperatures and
572 specific humidity. Validation tests have been used to compare the outputs of the models and the
573 measured windows states obtained in the years 2014 and 2015 in the open-plan office. According
574 to the different evaluation indicators, the results show that all the three models perform well
575 with the testing sets.

576 In the future, we can improve the over-represented 'ALL CLOSED' label by resampling in
577 order to have an unbiased data set or by providing different weights for each label to penalize

578 misclassification. In addition, with an algorithm that combines multiple trees and control for bias
 579 or variance, like Random Forests [20] or Gradient boosted trees [24], the Decision Tree model
 580 could have a better performance. For the k-NN model, an efficient method to deal with both
 581 the numerical and categorical data in order to avoid the loss of information needs to be further
 582 investigated. Furthermore, the high performance of Kernel Approximation approach - a good
 583 nonlinear separator, is also noteworthy.

584 We could then use one of the three developed models as a standalone, or as a part of a real-
 585 time IAQ monitoring system, in order to optimize the action to be taken to reduce the exposure
 586 of the occupants.

587 References

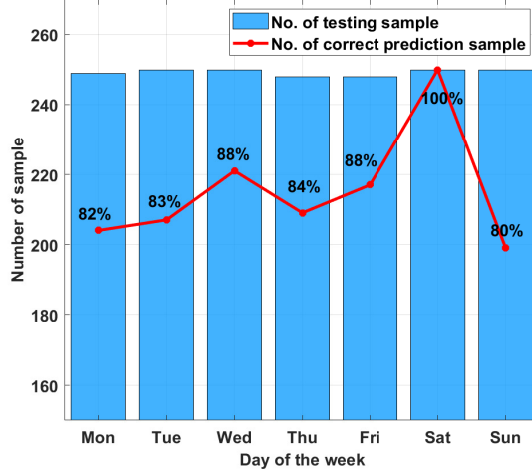
- 588 1. Report to congress on indoor air quality: Assessment and control of indoor air pollution. Tech. rep.,
 589 U.S. Environmental Protection Agency (1989)
- 590 2. Amasyali, K., El-Gohary, N.M.: A review of data-driven building energy consumption
 591 prediction studies. *Renewable and Sustainable Energy Reviews* **81**, 1192–1205 (2018).
 592 <https://doi.org/10.1016/j.rser.2017.04.095>
- 593 3. Andersen, C., Bro, R.: Practical aspects of parafac modeling of fluorescence excitation-emission data.
 594 *Journal of Chemometrics* **17**, 200 – 215 (04 2003). <https://doi.org/10.1002/cem.790>
- 595 4. Andersen, R., Fabi, V., Toftum, J., Corgnati, S.P., Olesen, B.W.: Window opening behaviour mod-
 596 elled from measurements in danish dwellings. *Building and Environment* **69**, 101–113 (2013)
- 597 5. Box, G., Jenkins, G.M., Reinsel, G.: *Time Series Analysis: Forecasting and Control*. 3rd ed. Engle-
 598 wood Cliffs, NJ: Prentice Hall (1994)
- 599 6. Cali, D., Wesseling, M.T., Müller, D.: Winprogen: A markov-chain-based stochastic window sta-
 600 tus profile generator for the simulation of realistic energy performance in buildings. *Building and*
 601 *Environment* **136**, 240–258 (2018)
- 602 7. Chen, S., Mihara, K., Wen, J.: Time series prediction of co2, tvoc and hcho based on ma-
 603 chine learning at different sampling points. *Building and Environment* **146**, 238–246 (2018).
 604 <https://doi.org/10.1016/j.buildenv.2018.09.054>
- 605 8. Cheng, Y.H., Lin, Y.L.: Measurement of particle mass concentrations and size distribu-
 606 tions in an underground station. *Aerosol and Air Quality Research* **10**, 22–29 (02 2010).
 607 <https://doi.org/10.4209/aaqr.2009.05.0037>

- 608 9. Dai, X., Liu, J., Zhang, X.: A review of studies applying machine learning models to predict oc-
609 cupancy and window-opening behaviours in smart buildings. *Energy and Buildings* **223**, 110–159
610 (2020)
- 611 10. D’Oca, S., Hong, T.: A data-mining approach to discover patterns of window opening and closing
612 behavior in offices. *Building and Environment* **82**, 726–739 (2014)
- 613 11. Dreiseitl, S., Ohno-Machado, L.: Logistic regression and artificial neural network classifica-
614 tion models: a methodology review. *Journal of Biomedical Informatics* **35**(5), 352–359 (2002).
615 [https://doi.org/10.1016/S1532-0464\(03\)00034-0](https://doi.org/10.1016/S1532-0464(03)00034-0)
- 616 12. Edwards, R.E., New, J., Parker, L.E.: Predicting future hourly residential electrical con-
617 sumption: A machine learning case study. *Energy and Buildings* **49**, 591–603 (2012).
618 <https://doi.org/10.1016/j.enbuild.2012.03.010>
- 619 13. El Naqa, I., Murphy, M.J.: *What Is Machine Learning?* Springer International Publishing (2015)
- 620 14. Fabi, V., Andersen, R., Corgnati, S., Olesen, B.: Occupants’ window opening behaviour: A literature
621 review of factors influencing occupant behaviour and models. *Building and Environment* **58**, 188–198
622 (2012)
- 623 15. Fix, E., Hodges, J.L.: *Discriminatory analysis : nonparametric discrimination, consistency properties.*
624 *USAF School of Aviation Medicine* (1951)
- 625 16. Godish, T., Spengler, J.D.: Relationships between ventilation and indoor air quality: A review.
626 *Indoor Air* **6**(2), 135–145 (1996). <https://doi.org/10.1111/j.1600-0668.1996.00010.x>
- 627 17. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning.* Springer Series in
628 *Statistics*, Springer New York Inc., New York, NY, USA (2001)
- 629 18. of Heating Refrigerating, A.S., Engineers, A.C.: *Thermal environmental conditions for human occu-*
630 *pancy.* Tech. rep., American Society of Heating Refrigerating and Air-Conditioning Engineers (2017)
- 631 19. Hinds, W.C.: *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles.*
632 *Wiley* (1999)
- 633 20. Ho, T.K.: Random decision forests. p. 278–282 (07 2016)
- 634 21. Hosmer, D., Lemeshow, S.: *Applied Logistic Regression.* Hoboken, vol. 354 (01 2000).
635 <https://doi.org/10.1002/0471722146>
- 636 22. Markovic, R., Grintal, E., Wölki, D., Frisch, J., van Treeck, C.: Window opening model using deep
637 learning methods. *Building and Environment* **145** (2018), 10.1016/j.buildenv.2018.09.024
- 638 23. Martínez-Comesaña, M., Eguía-Oller, P., Martínez-Torres, J., Febrero-Garrido, L., Granada-
639 Álvarez, E.: Optimisation of thermal comfort and indoor air quality estimations applied to in-

- 640 use buildings combining nsga-iii and xgboost. *Sustainable Cities and Society* **80**, 103723 (2022).
641 <https://doi.org/10.1016/j.scs.2022.103723>
- 642 24. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. *Frontiers in neurorobotics* **7**, 21 (12
643 2013). <https://doi.org/10.3389/fnbot.2013.00021>
- 644 25. Pan, S., Xiong, Y., Han, Y., Zhang, X., Xia, L., Wei, S., Wu, J., Han, M.: A study on influential factors
645 of occupant window-opening behavior in an office building in china. *Building and Environment* **133**,
646 41–50 (2018)
- 647 26. Park, J.: Long-term field measurement on effects of wind speed and directional fluctuation on
648 wind-driven cross ventilation in a mock-up building. *Building and Environment* **62**, 1–8 (2013).
649 <https://doi.org/10.1016/j.buildenv.2012.12.013>
- 650 27. Park, J., Choi, C.: Modeling occupant behavior of the manual control of windows in residential
651 buildings. *Indoor Air* **29** (11 2018). <https://doi.org/10.1111/ina.12522>
- 652 28. Park, J., Jeong, B., Chae, Y.T., Jeong, J.W.: Machine learning algorithms for predicting occupants'
653 behaviour in the manual control of windows for cross-ventilation in homes. *Indoor and Built Envi-*
654 *ronment* **30**(8), 1106–1123 (2020). <https://doi.org/10.1177/1420326X20927070>
- 655 29. Quinlan, J.R.: Induction of decision trees. *Machine Learning* **1**, 81–106 (1986).
656 <https://doi.org/10.1007/BF00116251>
- 657 30. Rahimi, A., Recht, B.: Random features for large scale kernel machines. *Advances in Neural Infor-*
658 *mation Processing Systems* **20**, 1177–1184 (01 2008)
- 659 31. Raja, I.A., Nicol, J., McCartney, K.J., Humphreys, M.A.: Thermal comfort: use of con-
660 trols in naturally ventilated buildings. *Energy and Buildings* **33**(3), 235–244 (2001).
661 [https://doi.org/10.1016/S0378-7788\(00\)00087-6](https://doi.org/10.1016/S0378-7788(00)00087-6)
- 662 32. Ramalho, O., Ouaret, R., Ionescu, A., Le Ponner, E., Candau, Y.: Tribu - suivi dynamique en
663 temps réel de la qualité de l'air intérieur dans un environnement de bureaux. contributions des
664 sources et modèle prévisionnel rapport, primequal apr eiai / projet tribu. Tech. rep., CSTB (2016),
665 https://www.primequal.fr/sites/default/files/tribu_rf.pdf
- 666 33. Sarkhosh, M., Najafpoor, A., Alidadi, H., Shamsara, J., Amiri, H., Andrea, T., Kariminejad, F.: In-
667 door air quality associations with sick building syndrome: An application of decision tree technology.
668 *Building and Environment* **188** (2021)
- 669 34. Tan, Pang-Ning, Steinbach, M., Adeyeye Oshin, M., Kumar, V., Vipin: *Introduction to Data Mining*
670 (05 2005)
- 671 35. Tien, P.W., Wei, S., Liu, T., Calautit, J., Darkwa, J., Wood, C.: A deep learning approach to-
672 wards the detection and recognition of opening of windows for effective management of building

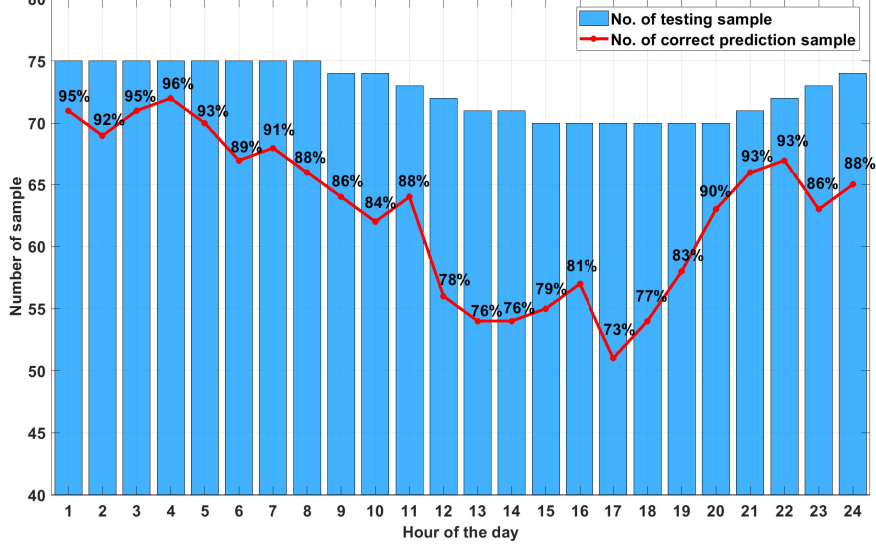
- 673 ventilation heat losses and reducing space heating demand. *Renewable Energy* **177**, 603–625 (2021).
674 <https://doi.org/10.1016/j.renene.2021.05.155>
- 675 36. Viet, L., Sarlos, T., Smola, A.: Fastfood: Approximate kernel expansions in loglinear time. 30th
676 International Conference on Machine Learning, ICML 2013 **28**, 244–252 (08 2013)
- 677 37. Wei, W., Ramalho, O., Malingre, L., Sivanantham, S., Little, J.C., Mandin, C.: Machine learn-
678 ing and statistical models for predicting indoor air quality. *Indoor Air* **29**(5), 704–726 (2019).
679 <https://doi.org/10.1111/ina.12580>
- 680 38. Yao, M., Zhao, B.: Window opening behavior of occupants in residential buildings in beijing. *Building*
681 *and Environment* **124**, 441–449 (2017)
- 682 39. Zhao, Y., Qiu, R.C., Zhao, X., Wang, B.: Speech enhancement method based on low-rank
683 approximation in a reproducing kernel hilbert space. *Applied Acoustics* **112**, 79–83 (2016).
684 <https://doi.org/10.1016/j.apacoust.2016.05.008>

The statistics for Decision Tree Models accuracy of each day of the week in the testing set including data of 2014



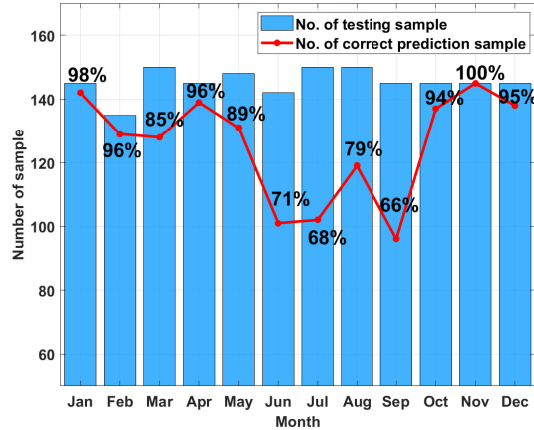
(a)

The statistics for Decision Tree Models accuracy of each hour of the day in the testing set including data of 2014



(b)

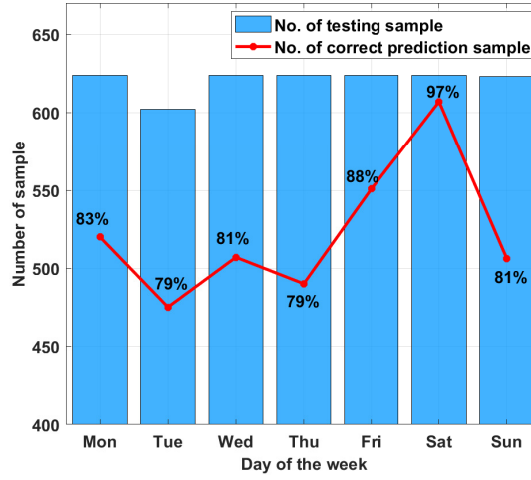
The statistics for Decision Tree Models accuracy of each month in the testing set including data of 2014



(c)

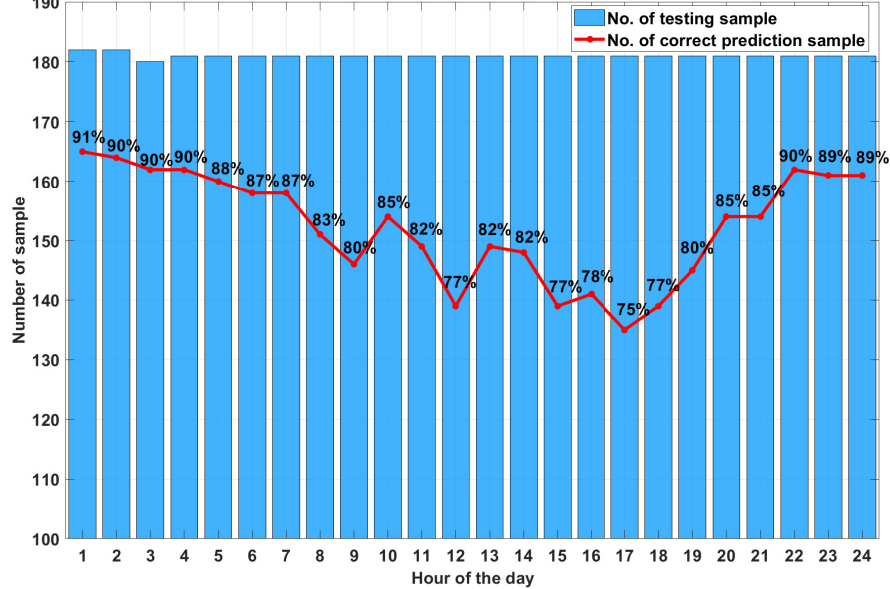
Fig. 11. The statistics for DT Models accuracy according to (a) each day of the week, (b) each hour of the day, and (c) each month for the test set 2014. The corresponding accuracy is displayed above the red curve for each time period.

The statistics for Decision Tree Models accuracy of each day of the week in the testing set including data from Jan-June of 2015



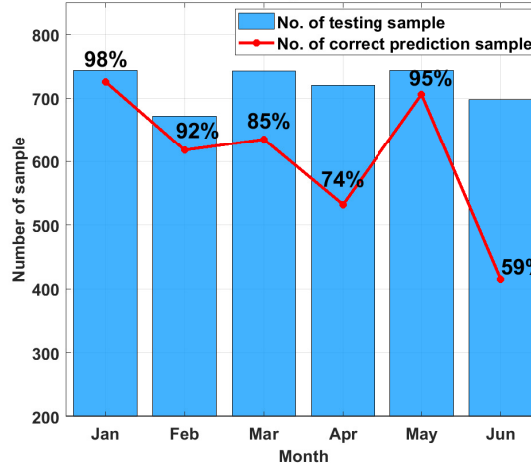
(a)

The statistics for Decision Tree Models accuracy of each hour of the day in the testing set including data from Jan-June of 2015



(b)

The statistics for Decision Tree Models accuracy of each month in the testing set including data from Jan-June of 2015



(c)

Fig. 12. The statistics for DT Models accuracy according to (a) each day of the week, (b) each hour of the day, and (c) each month for the test set 2015. The corresponding accuracy is displayed above the red curve for each time period.

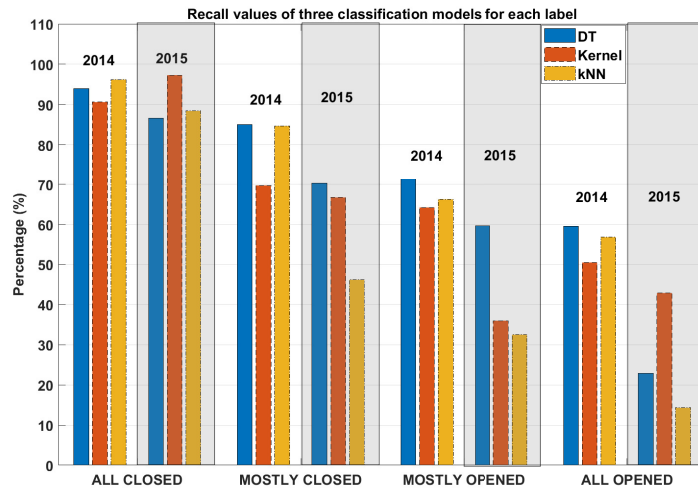


Fig. 13. Recall values of three classification models: Decision Tree, k-NN and Kernel approximation. The obtained values for testing data from January to June of 2015 are displayed in grey background.

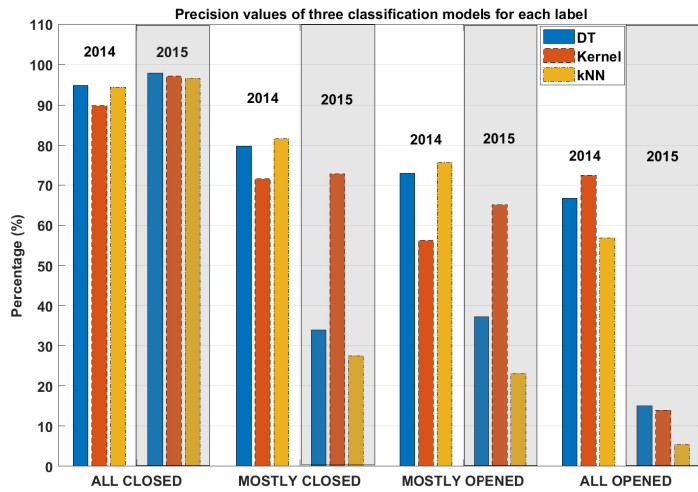


Fig. 14. Precision values of three classification models: Decision Tree, k-NN and Kernel approximation. The obtained values for testing data from January to June of 2015 are displayed in grey background.

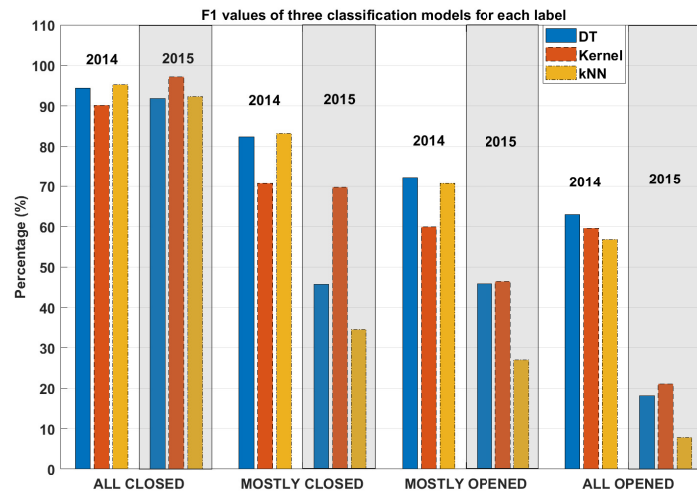


Fig. 15. F-1 values of three classification models: Decision Tree, k-NN and Kernel approximation. The obtained values for testing data from January to June of 2015 are displayed in grey background.