ORIGINAL ARTICLE
DOI: 10.5433/1679-0375.2022v43n1p75

Application of a machine learning model in study of energy efficiency in buildings: focus on light construction systems

Aplicação de um modelo de aprendizado de máquina em estudo de eficiência energética de edificações: foco para sistemas construtivos leves

Guilherme Natal Moro[1]; Rodrigo dos Santos Veloso Martins[2];
Thalita Gorban Ferreira Giglio[3]

Abstract

The use of machine learning techniques in thermoenergetic performance studies of buildings emerges as an alternative to conventional methods which analysis require greater data complexity. This research aims to apply a machine learning technique in the study of energy efficiency of a building executed in a light construction system. Thus, an algorithm was implemented referring to an optimized model of classification and regression tree (CART) for application in a data set. This data set includes 2048 parametric simulations of a housing in a light construction system to the climate of the city of São Paulo, whose output indicators are the annual thermal load for heating and the annual thermal load for cooling. From the application of a tree pruning methodology and the use of Grid Search and k-fold Cross Validation techniques, the training and testing of the model was repeated 100 times, thus obtaining average results of 1.11% of error for heating loads and 1.52% of error for predicting cooling loads. Subsequently, a sensitivity analysis was performed, revealing the thermal transmittance property of the walls as the parameter with the greatest influence on the prediction of heating load and the condition of contact between the ground and the floor as the parameter with the greatest influence on the prediction of cooling load. Finally, decision trees were generated for visual analysis of strategies that can be adopted to obtain better levels of thermoenergetic performance. Thus, a more simplified diagnosis of energy efficiency was obtained, with low complexity in the interpretation of its results, favoring greater diffusion of the technology in light systems.

Keywords: Classification and regression tree; computer simulation; energy consumption.

Resumo

A utilização de técnicas de aprendizado de máquina em estudos de desempenho termoenergético de edificações surge como uma alternativa aos métodos convencionais, que demandam uma maior complexidade de dados. Esta pesquisa tem por objetivo aplicar uma técnica de aprendizado de máquina em estudo de eficiência energética de uma edificação executada em sistema construtivo leve. Sendo assim, foi feita a implementação de um algoritmo referente a um modelo otimizado de árvore de classificação e regressão (CART) para aplicação em um conjunto de dados. Tal conjunto contempla 2048 simulações paramétricas de uma habitação em sistema construtivo leve para o clima da cidade de São Paulo, cujos indicadores de saída são a carga térmica anual de aquecimento e a carga térmica anual de refrigeração. A partir da divisão dos dados em conjuntos de treinamento e teste e da aplicação de técnicas de Grid Search e k-fold Cross Validation para otimização de hiperparâmetros, foram obtidos resultados médios de 1,11% de erro para cargas de aquecimento e 1,52% de erro para a predição de cargas de refrigeração. Posteriormente, foi feita uma análise de sensibilidade revelando a propriedade de transmitância térmica das paredes como parâmetro de maior influência na predição de carga de aquecimento e a condição de contato do solo com o piso como parâmetro de maior influência na predição de carga de refrigeração. Por fim, foram geradas as árvores de decisão para análise visual de estratégias que podem ser adotadas para a obtenção de melhores níveis de desempenho termoenergético. Obteve-se assim, um diagnóstico de eficiência energética mais simplificado na interpretação de seus resultados, favorecendo maior difusão da tecnologia em sistemas leves.

Palavras-chave: Árvore de classificação e regressão; simulação computacional; consumo energético.

[1] Graduate student., Civil Contruction Dept., UEL, Londrina, PR, Brazil; E-mail: guilherme.natal@uel.br
[2] Prof. Dr., Mathematics Dept, UTFPR, Apucarana, PR, Brazil; E-mail: rodrigomartins@utfpr.edu.br
[3] Prof. Dr., Civil Contruction Dept., UEL, Londrina, PR, Brazil; E-mail: thalita@uel.br

Introduction

Light construction systems, among the one executed in *wood frame* and steel frame, are those which show superficial density, below 60 kg/m^2 (ABNT, 2013). Due to the low weight, in Brazil, the residential buildings executions in light construction systems are hampered, among other aspects, by the low thermal wrapping capacity, which does not assist to the minimum parameter stablished by the simplified method of the Brazilian Association of technical Standards (ABNT, 2013). Because of this, it is necessary to perform a more meticulous study by using computational simulation to prove the thermal performance (ZARA, 2019).

However, the methods by computational simulation for the performance analysis and energy efficiency of light systems are normally complex and involve high operational cost. According to Sun, Haghighat and Fung (2020) it is necessary the development of prediction methods that are clear in their diagnosis and in their necessary characteristics which a house in a light construction system, reaches bigger level of thermo-energetic performance.

According to Justino, Silva and Silva Rabelo (2020), one of the alternatives which have been highlighted in the energy efficiency studies is the employment of Artificial Intelligence (AI), as well as the derived classifications in this area, as the machine learning. According to other authors, the use of AI allows the time save, financial resources and offers precise results.

In this discussion, some studies carried out in Brazil have been directed to a new prediction method of the energy efficiency level in buildings with the insertion of learning techniques of the machine. In Melo (2012), for example, a neural network was created to estimate the energy consumption in commercial buildings. Yet in Olinger (2019) a method based in a neural network was developed, and it was able to estimate the thermal comfort in offices, to São Paulo city, with awfully close results to the one obtained by computational simulation.

Still in this discussion, new Brazilian labelling methods of energy efficiency levels of buildings involving learning techniques of the machine in the energy consumption prediction and comfort. As current normative instructions represented by the INI-C (Normative Inmetro Instruction to the Energy Efficiency Classification of Commercial, Services and Public Buildings) (BRASIL, 2021a) and INI-R (Normative Inmetro to the Energy Efficiency Classification Residential Buildings) (BRASIL, 2012b) incorporates in their simplified methods, one meta-model of analysis that uses a learning technique in different climatic realities, generating some estimations of the annual thermal load of refrigeration and heating without the need of the computation simulation employment by the designer.

In the international sphere, there is a consolidated scenery related to the studies development with the use of machine learning in predictions of energy consumption in buildings. In Tsanas and Xifara (2012), a set of simulated data by *Ecotec software* referred to the parameters that had a profound influence of an edification in the acquisition of the heating and cooling, are submitted to two learning machine methods, known as, linear regression and *random forest* . In Seyedzadeh *et al*. (2019), the researchers compare through the investigation of methods of artificial neural network, support vector machine, gaussian process, *random forest* , and regression trees with a propelled gradient, in applications to two sets of data, being one of them, the same which was present in Tsanas and Xifara (2012).

The same way, Pham *et al*. (2020) propose a model based on *random forest* to predict the energy consumption and time of five different buildings throughout a year, examining, thus, the generalization capacity and efficiency of the model. Besides this, in Walker *et al*. (2020), corresponding data were used in commercial buildings, monitored for two years for the training with models as propelled trees, *random forest* and support vector machine, which were analyzed and tested to foresee the energetic demand of these buildings.

The most popular approaches which involve learning techniques of machine use models that are pointed by Sun, Haghighat and Fung (2020). The authors carried out a literature revision involving an analysis of the evolution of the use of learning methods in the machine in the predictions of energy consumption. The increase of the publications in the area from 2015 on, observing a focus in the methods like linear regression, regression trees, support vector machine, artificial neural networks, besides *random forest*.

Thus, a model that has being highlighted in the literature revisions of the area is the Classification and Regression Tree, represented by the CART, a decision tree with high potential in which the entrance variables show some support both to the dataset and classification, which present answers with a discreet class label and set of regression data, which show answers with continuous values.

Moreover, the model builds a tree from a root-node, represented by the parameters that had biggest influence in making decision, ranging a way through different internal nodes, until it reaches the leaves, in which a wanted prediction occurs (ZEKIĆ-SUŠAC; MITROVIĆ; HAS, 2021). This method has a significant advantage, when compared to the others, is easy to understand and has implementation and operation with reasonable complexities, contributing to the spread of technology (BOURDEAU *et al*., 2019).

To analyze the efficiency of the implementation of machine learning model is important to apply the error metrics. In Seyedzadeh *et al*. (2019), for example, the adopted metrics were a quadratic average error of the root, absolute average error, and determination coefficient. But in the literature revision shown in Sun, Haghighat and Fung (2020) are emphasized, besides these ones, the absolute average error percentage and the quadratic average error. It is important to point out that although there are several statistics metrics to error analysis, each one allows the reader to different interpretation possibilities for the same result, according to the measurement unit provided by the mentioned metric.

Besides, in order to obtain better results with the implementation, there are optimizations that can be done through some techniques as *Grid Search* (SEYEDZADEH *et al*., 2019), based on the search for values and contributions of hyperparameters that can impact in a positive way in the training, which can be combined with the cross validation technique (ZEKIĆ-SUŠAC; MITROVIĆ; HAS, 2021).

In general, the studies, based on machine learning techniques are observed. They aim to simplify the evaluation methods of the thermal and energy performance and have been discussed and improved in recent studies. In this way, they are suitable for application in light construction systems, improvement and diffusion of the technology, through thermo-energetic adaptations that are more interpretable by designers. According to Nunes *et al*. (2020), the development of the studies is important and should help in the spread of technology of light systems, baring in mind its significant potential in reaching a high level of performance and energetic efficiency, in different weather conditions.

In this context, the research aims to apply a machine learning technique in a study of energy efficiency in buildings in order to obtain, in a simplified and accurate way, the annual thermal load of heating and cooling of a dwelling in light system.

Material and method

A set of energy efficiency data of light construction edifications generated in Zara (2019) was investigated in this study from 2048 parametric simulations, having as output data, values of annual thermal cooling and heating load of a social interest housing in *wood frame*, for the weather of São Paulo.

To reach the proposed aim, the adopted methodologic procedures went through a characterization phase of the set of data, followed of the selection and experimental applications, metric definition of the error analysis, application of machine learning model to the set of data, and finally, a phase in which the learning model of the machine was analyzed.

A selection of the computational learning model of the machine as well as the tests and applications developed in this research were carried out through the implementation in *Python* of the corresponding algorithms to each of the defined phases. Below, the accurate details of the research phases.

Characterization of the set of data

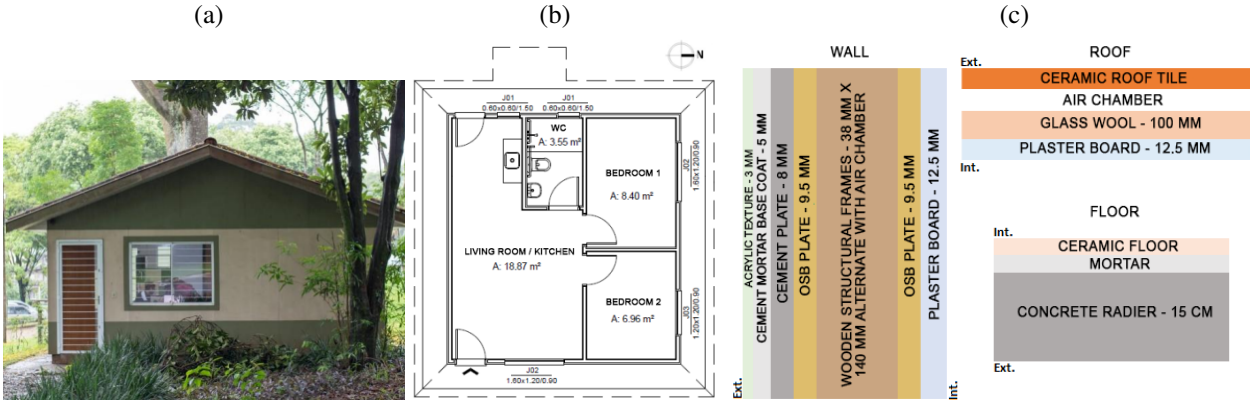
The set of interest data of this study was developed by Zara (2019), in which computational simulations with the intention to obtain the thermal energy performance of a social interest housing were done, developed in *wood frame*, and considering the weather of São Paulo city (23° 37' S, 46° 39' O, 802 m). The project follows the demand of the extinct housing program "My House, my life" and the construction system, considered innovative in the country, has DATec (Technical Evaluation Document) registered in the National Technical Evaluation System (SINAT). As shown in Figure 1(a)-(c), it is possible to observe the main façade of the studied house, its ground plan and the structure of the light construction system adopted, as well.

The house has a useful total area of 41.3 m^2, divided in two bedrooms, interconnected living room and kitchen, and bathroom.

The carried-out simulations through the *EnergyPlus* software considered 2048 different combinations resulting of the variation of 10 thermal-physical parameters, as it is shown in Table 1.

Wall Thermal Transmittance (Uwall) represented by the equal value 0.63 $W/m^2.K$ is referred to a wall composition in a light system with the inclusion of a thermal insulator (glass wool), while the value of 1.88 $W/m^2.K$ represents a composition without the inclusion of glass wood, keeping itself only with air chamber among sealing plates in OSB (*oriented strand board*).

Figure 1 – Studied House: (a) main façade; (b) ground plan; (v) compositions of the constructions systems.



Source: Zara (2019).

Table 1 – Variable Parameters, respective indexes (ID) and variation level.

Variable parameters	ID	Levels	Values
Wall thermal transmittance	Uwall	2	1.88/0.63
Wall solar absorbance	αwall	2	0.2/0.8
Roof thermal capacity	Ctroof	2	45/17
Roof thermal transmittance	Uroof	2	0.37/0.63
Roof solar absorbance	αroof	2	0.2/0.8
Roof emissivity	εroof	2	0.1/0.9
Roof contact with the outside	roof	2	0/1
Floor contact	floor	2	0/1
Existence of shutters	shutt.	2	0/1
Solar orientation	orient.	4	0/90/180/270

Source: Zara (2019).

The wall solar absorbance (αpar) and roof (αroof) vary in two extreme levels, representing a light color (0.2) and a dark color (0.8). In the case of the roof, the thermal capacity (CTroof) varied as the type of tile, being 45 kJ/m^2.K referred to the composition with ceramic tiles and 17 kJ/m^2.K referred to the composition with fiber cement tiles. The roof thermal transmittance (Uroof), in turn, considers two thickness of glass wool, where the value of 0.37 W/m^2.K corresponds to the composition with insertion of glass wool in the 100 mm thickness while 0.63 W/m^2.K represents a composition of glass wool of 50 mm.

Yet to the roof, the emissivity of the surfaces was considered in two levels of variation, being these ones 0.10 representing the insertion of a radiant barrier (aluminum blanket) and 0.90 for the other types of surfaces. The floor contact varied in 0, when there is no contact with the outside and 1, when there is direct contact with the outside. For the floor contact, it was considered 0, when there was no floor contact, and 1 when there is direct floor contact. Moreover, the window shading varied in two levels, considering the existence of shutters, being 0 in case there isn't, and 1 in case there is.

Finally, the only varied parameter in more two levels was the solar variation, evaluation the edification of every each 90° in total of 4 cases: (0°) north, (90°) east, (180°) south and (270°) west.

For each combination of the analyzed parameters, two performance indexes were generated: the annual thermal heating load (kWh/m^2. year) and the annual thermal cooling load (kWh/m^2. year).

Adopted machine learning model

The CART model was selected having in mind the use of decision trees in the area literature and by the advantage of being easier to understand and interpret the results due to the possibility of being shown graphically. Based on a training done with input vectors and their respective answers, CART builds a decision tree that begins with a root-knot, in which the vectors are divided in different internal knots or leaf-knots. To the internal knots, the inputs are continuously divided in subsets through a criterium, while the leaf-knots represent the output, that is, the answer to a specific input vector. The division criterium for a knot can be determined with the use of entropy or an impurity Gini index for classification problems and absolute medium error or quadratic medium error for regression set (MARSLAND, 2011).

The present study used an algorithm by the Scikit-Learn library (PEDREGOSA, 2011) version 0.23.1 through the programming language *Python* 3.8.3, that presents itself as a decision tree resulting of an optimization of the CART model.

The algorithm was prepared to divide the dataset into 70% of those selected at random so that training could be performed, while the remaining 30% were used for testing. The training was performed with hyperparameter optimization using *Grid Search* and cross validation.

Considering the effect that a specific division into training and testing sets can have on the results, this training and testing process was repeated 100 times and the results of the iterations were averaged, as suggested by Refaeilzadeh, Tang and Liu (2009). Besides, metric analysis of error was stablished to be able to perform the proper analysis of the obtained results. In this way, with the training proportions and established tests, as well as the metric analysis of error, were defined which are hyperparameters of the algorithm CART that can be determined to optimize the results obtained by the implementation.

Hyperparameters test

The hyperparameters of a determined machine Learning model are variables that can define or adjust several internal processes involved in the execution of the algorithm. (SEYEDZADEH *et al*., 2019). An example of such process is the pruning held in a CART tree after its maximum growth, avoiding overfitting problems. The overfitting occurs when the model adapts itself too much to the dataset which was trained but does not present a good generalization capacity when submitted a one new dataset (MARSLAND, 2011). To avoid such overfitting the data, the model training was conducted considering different values of the pruning calibration hyperparameter α of what is considered to be in excess in the tree. More precisely, a new error metric $R_\alpha(T)$ is considered with the introduction of a punitive term to the number of terminal nodes of the tree, according to equation (1):

$$R_\alpha(T) = R(T) + \alpha\left|T'\right|, \tag{1}$$

in which $|T'|$ is the number of terminal knots in the tree T and $R(T)$ is the value corresponding to the result obtained by a metric of error applied to tree T.

However, as bigger the size of the tree, less residue it is going to show, but, bigger is the overfitting risk to the data and bigger the risk of overfitting will be to the data and bigger the punitive term that results of the product between the effective value of the α variable with the quantity of terminal knots of the tree. With this, the pruning finds a T tree that minimizes $R_\alpha(T)$, obtaining a balance between the training error and the size of the tree with the objective of improving its capacity of generalization. The model also provided a method which implements the pruning process and returns the effective values of variable α, that is, the significant values to be tested by this hyperparameter.

Grid search and cross validation

The *grid search* technique was applied to the training set, composed of 70% of the data, to obtain the best combination of values for the available hyperparameters, thus optimizing the model's performance. In this work the hyperparameter space was defined by as a set of possible values for the pruning hyperparameter α described in the previous section. For each α value, training was performed using Cross Validation, whose performance was evaluated according to a pre-established error analysis criterion.

The k-*fold* Cross Validation technique was used in this work, where the training data set was fractioned into a certain number of subsets called *folds* (PEDREGOSA, 2011). When 10 *folds* were stablished, the set of training was divided in 9 *folds* , which were submitted to a training, and 1*fold* was used for the validation of this training. With this, all pruning hyperparameter values were tested, and then, a next *fold* was defined among the 10 to be used in a new training validation carried out by other 9 *folds*. The process was repeated until the other 10 *folds* were used in the validations. In this way, the application of the technique allowed to register some values for analysis, as the training performance done with each one of the validation *folds*. Therefore, 10 *folds* resulted in the register of 10 different vectors with the performance of each one of the considered pruning hyperparameter values, in which the first vector showed the training performance done with the first *fold* withdrawn for the training of the other 9, and the second vector showed the training performance done with the second *fold withdrawn* for the training of the other 9 *folds* , and so on. Then, to choose the optimal value of the pruning hyperparameter α, the average of the results obtained for each α value over the 10 trainings performed in the Cross Validation was recorded.

Metric of error analysis

To make a precise analysis of the efficiency obtained with the implementation of the CART model, metrics of error frequently used in the area were adopted, as Sun, Haghighat and Fung (2020). Among them, the Medium Absolute Percentual Error (MAPE), described in equation (2), the Medium Quadratic Error (MSE), described in equation (3), the Medium Absolute Error (MAE), described in equation (4) and the determination coefficient (R^2), described in equation (5), given by:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y_i} - y_i}{y_i}\right| * 100, \tag{2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2, \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y_i} - y_i|, \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (5)$$

in which, n is the number of samples, y_i is the real value of the concerned parameter, $\widehat{y_i}$ is the predicted value for the same parameter and $\bar{y}$ is the average of the y_i values.

In addition, the Error Histogram for both loads, illustrating the error distribution during the 100 defined iterations for the training and test, described in equation (6):

$$Error = \widehat{y_i} - y_i, \quad (6)$$

in which y_i is the real value of the concerned parameter and valor $\hat{y}_i$ is the value predicted in training for the same parameter. The criterion adopted as error, used for the elaboration of the Histogram, equation (6), is used to express the value of the differences between the values predicted in training and the real value of the variables presented by the set of data.

Sensibility analysis

In each iteration of training and test, the input parameters of the machine learning model were submitted to a sensibility analysis. This analysis permitted to understand which most influent parameters are and help in the understanding of the decision tree generated in this study.

The adopted calculation of the sensibility analysis considered a normalized reduction of the (MSE) occasioned by each input parameter in the training process of the tree. In this way, a proportional reduction through the ratio between the reduction occasioned by the parameter in the division of the concerned knot and the general reduction produced by the leaves of the tree. With this, the parameters that presented the biggest values are the one of bigger influence in making decisions.

Results

The showed results and the implementation analysis of CART model trained and tested with the dataset presented in Zara (2019).

Energetic performance prediction

To analyze the energetic performance prediction, initially the average and the standard deviation of the referred number of leaves and the obtained errors, making a wider vision of the presented values possible for each prediction load. As shown in Table 2, the obtained values through the test phase, referred to the heating load prediction related to each one of the 100 iterations (PHL), as well as the resulting values of the average calculation and the standard deviation. In the same way, Table 3, shows the obtained values through the training and test referred to the annual thermal cooling load prediction related to each 100 iterations (APTCL), as well as the resulting values of the average calculation and the standard deviation.

From these obtained values it is noticed that for the heating load results, the average number of leaves of the trees generated was 1116.70 $\pm$ 26.727, with 0.25 $\pm$ 0.015 kWh/m^2.year values for MAE, 0.17 $\pm$ 0.025 (kWh/m^2.year) ^{2} for MSE, 1.11 $\pm$ 0.063% for MAPE and 0.9983 $\pm$ 0.0003 for R^2. For the cooling load, the average number of leaves of trees generated was 1085.52 $\pm$ 38.567, with the 0.16 $\pm$ 0.017 kWh/m^2. year for MAE, 0.09 $\pm$ 0.040 (kWh/m^2.year) ^{2} for MSE, 1.52 $\pm$ 0.145% for MAPE and 0.9980 $\pm$ 0.0009 for R^2. Such results are considered satisfactory when comparing to studies that also used the predictions of heating and cooling loads, as in Tsana and Xifara (2012) in which the MAPE values obtained through the training with a random forest model for this type of prediction was 2.18 $\pm$ 0.64% and 4.62 $\pm$ 0.70%, respectively. Figure 2(a)-(b) illustrates the error histogram obtained for the results of annual thermal cooling load and annual thermal heating load, respectively.

According to Figure 2, it is observed in both predictions load, a bigger concentration of error, calculated by the difference of predicted values with the real values represented in the test phase, next to zero and a smaller occurrence of error which were bigger than 2 and smaller than -2 in kWh/m^2. year.

Results of sensibility analysis

In Zara (2019), the sensibility analysis of the input parameters of the dataset was obtained for each environment of the house, through a variance analysis (ANOVA), which has as its goal, to analyze the average of the simulations results and determine which parameters show a bigger influence in the system response. When comparing the obtained results through this variance analysis with the sensibility analysis generated by the implementation on CART algorithm, it is noticed that there is a similarity referred to the behavior of the variables, with slight differences that are natural due to the application of different methods. To compare it, an average of the values generated by the environment was obtained, Zara (2019),

Table 2 – Results for annual thermal heating load.

	Iteration	Number of leaves	Depth	α	MAE	MSE	MAPE	R^2
	0	1134	12	1.98×10^{-17}	0.25	0.17	1.1	0.9982
	1	1140	12	1.98×10^{-17}	0.23	0.14	1.02	0.9986
	2	1098	12	1.98×10^{-16}	0.25	0.17	1.09	0.9982
HLi	3	1097	12	1.98×10^{-16}	0.26	0.19	1.15	0.9981
	4	1120	12	1.98×10^{-17}	0.26	0.17	1.15	0.9983
				⋮				
	98	1107	12	1.98×10^{-16}	0.22	0.12	1.07	0.9988
	99	1136	12	9.92×10^{-18}	0.23	0.14	1	0.9986
$\bar{HL}$	-	1116.70 ± 26.727	-	-	0.25 ± 0.015	0.17 ± 0.025	1.11 ± 0.063	0.9983 ± 0.0003

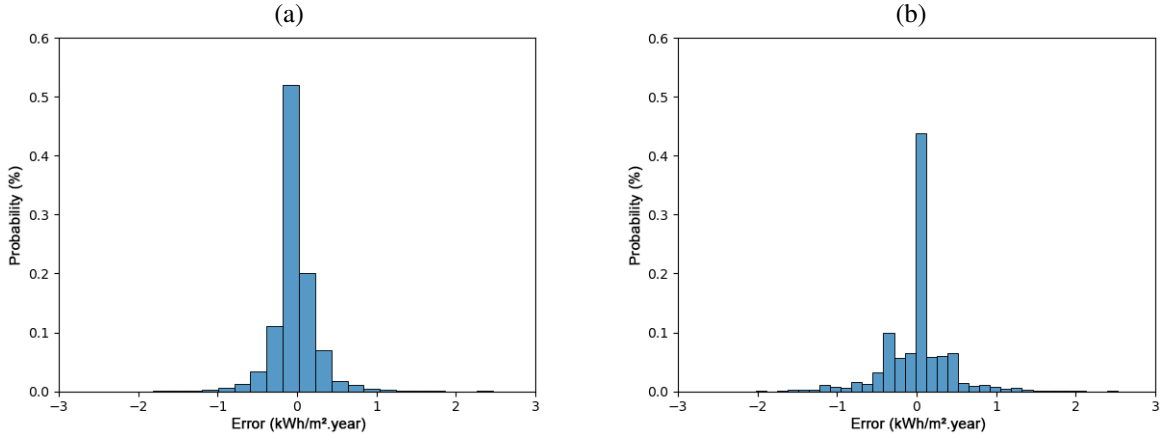
Source: The authors.

Table 3 – Results for annual thermal cooling load.

	Iteration	Number of leaves	Depth	α	MAE	MSE	MAPE	R^2
	0	1124	12	0	0.14	0.07	1.48	0.9985
	1	1093	12	9.92×10^{-18}	0.17	0.08	1.72	0.9983
	2	1091	12	9.92×10^{-18}	0.13	0.05	1.31	0.9989
CLi	3	1098	12	1.98×10^{-17}	0.18	0.15	1.56	0.9969
	4	1142	12	0	0.14	0.06	1.42	0.9987
				⋮				
	98	1117	12	0	0.17	0.09	1.72	0.9980
	99	1106	12	0	0.15	0.09	1.27	0.9982
$\bar{CL}$	-	1085.52 ± 38.567	-	-	0.16 ± 0.017	0.09 ± 0.040	1.52 ± 0.145	0.9980 ± 0.0009

Source: The authors.

Figure 2 – Typical Error Histogram: (a) cooling load; (b) heating load.



Source: The authors.

finding values, which refer to the loads heating predictions, of 60.13% of thermal transmittance influence of the walls, followed of 22.9% for the floor contact, 6.73% for the roof contact with the outside and 5.14% for the solar absorptance of the walls. For the sensibility analysis generated by the implementation of CART algorithm, was obtained in which refers to the heating load prediction, the average of values generated in each of the 100 iterations of training, obtaining, thus, values of 55.70% for the thermal transmittance of the walls, followed of 26.77% for floor contact, 7.63% for the roof contact with the outside and 6.84% for the solar absorptance of the walls.

On the other hand, in which refers to the cooling load prediction, in Zara (2019) values of 65.41% of floor contact influence were followed of 13.00% for solar absorptance and 9.43% for thermal transmittance of the walls were obtained. Since the implementation of the CART algorithm generated values of 61.32% of floor contact influence, followed of 15.39% for the solar absorptance of the wall and 15.15% for the thermal transmittance of the walls.

Decisions trees

To precisely analyze the energy efficiency point of view with trees generated, for each prediction among the 100 iterations, the one whose number of leaves was closer to the average of this parameter, was selected. According to Table 4, the obtained results for the selected iterations, being them, iteration 15, referred to the heating load prediction (HL_{15}) and iteration 86, referred to cooling load prediction (CL_{86}).

Only the first 3 levels of the trees referred to the selected iterations will be shown, making a better comprehension of the decision tree generated by the training of CART model with the dataset of interest possible.

In Figure 3 the selected decision tree is shown to predict the heating load. It is shown in the top part of the tree for parameters that have a bigger influence when taking decision, that is, the parameters that caused a bigger reduction of MSE in the training, being the thermal transmittance (Uwall) the parameter which is presented in the root of the tree (highest point), followed by the floor contact condition of the house (floor) and later, roof contact condition with the outside and the solar absorptance of the walls (αwall). These parameters are the ones of biggest influence when taking the decision of the tree.

In the same way, Figure 4 shows the 3 first levels of the decision tree referred to the cooling load. In this case, the parameter presented in the root of the tree is the condition of contact of the house to the floor, showing itself with the parameter with bigger influence in taking decision of the tree, followed by the solar absorptance (αwall) and thermal transmittance of the walls (Uwall).

Baring in mind the visual representation of these trees, the coincidence of the parameters that live in the top of the tree with the parameters of bigger influence obtained in the sensibility analysis is noticed, due to the applied concept of proportional reduction of MSE in the training in both cases.

Concerning to the annual heating thermal load prediction of a house in light construction system located in São Paulo city, it is observed through the tree generated, Figure 3, that an important strategy to obtain a better thermal-energetic performance would be a priority in the choice of insulating material in the composition of the external walls of light systems (Uwall=0.63W/m^2.K). The low thermal transmittance of the walls due to the incorporation of a thermal insulating material in its composition complicates the heat loss for the external environment and favors the most efficient use of the air conditioning system.

From the decision of the usage or not of thermal insulation of the walls, the attention for the thermal exchanges between the ground and the floor of the house, according to the most influenced parameter in the results of heating load results is directed. It is observed that when a housing unit does not have contact with the ground, upper pavement of a housing unit, the adoption of walls with thermal insulation can reduce around 50% the heating thermal load in comparison to a house without thermal insulation in the walls. The floor contact, condition of a grounded housing unit, favors the heat loss of the internal environment for the ground elevating the heating thermal load to keep the internal temperature conditions suitable. It is also observed that the absence of roof contact, housing unit with an intermediate pavement of a multi-familiar building or grounded pavement of a detached house, allows a reduction in the energy consumption due to the reduction of the area of contact to the internal environments with external air. Finally, the impacts associated to the parameters with solar absorptance of the walls, solar orientation of a house and the existence of shutters, is observed. Thus, in a general way, with a basis in the decision tree, the smaller value of heating thermal load would be obtained for a housing unit of intermediate pavement, without floor contact and with the roof, with walls with thermal insulation and painted in a light color.

Although when it is referred to the cooling loads prediction, Figure 4, the main strategy in the obtention of the best thermal-energetic performances for houses in light construction systems would keep in contact with the ground. According to the strategies of light colors in the decision tree, the best combination for the reduction of consumption for cooling would be, besides the adoption of light colors (solar absorptance $= 0.2$) and thermal insulation in the external walls (Uwall $= 0.63$W/m^2.K)

It is noticed that there are still many challenges to the designers, because some solutions of energy efficiency favor the reduction of the heating consumption but increase the cooling consumption. Even though, the usage of glass wool as well as the light colors in the walls of the houses without roof contact is an efficient combination both to the winter and summer in São Paulo city.

Conclusions

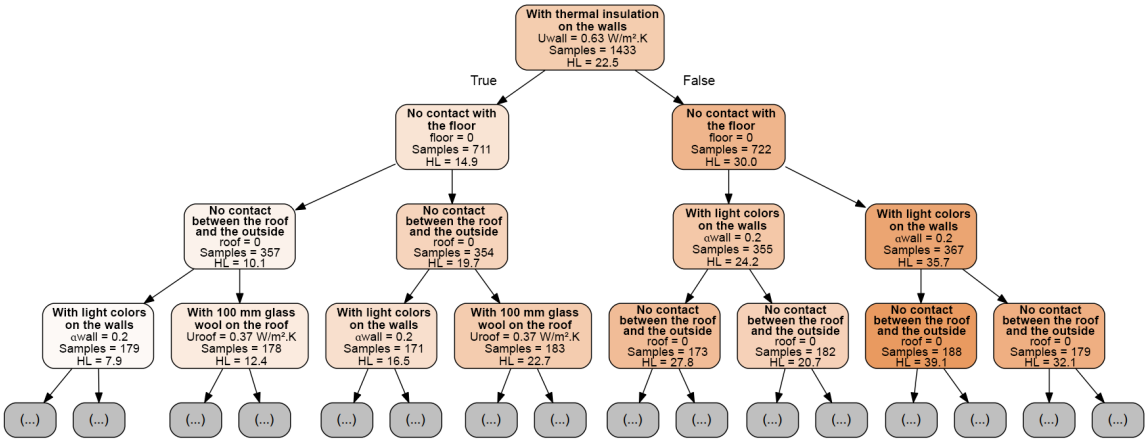
This study showed the implementation of CART algorithm in a dataset used in the prediction of thermal-energetic performance of a dwelling in light construction systems for the weather of São Paulo city.

Table 4 – Results for annual thermal heating load.

	Iteration	Number of leaves	Depth	α	MAE	MSE	MAPE	R^2
HL_{15}	15	1116	12	1.98×10^{-16}	0.24	0.15	1.08	0.9984
CL_{86}	86	1085	12	1.98×10^{-17}	0.16	0.07	1.44	0.9985

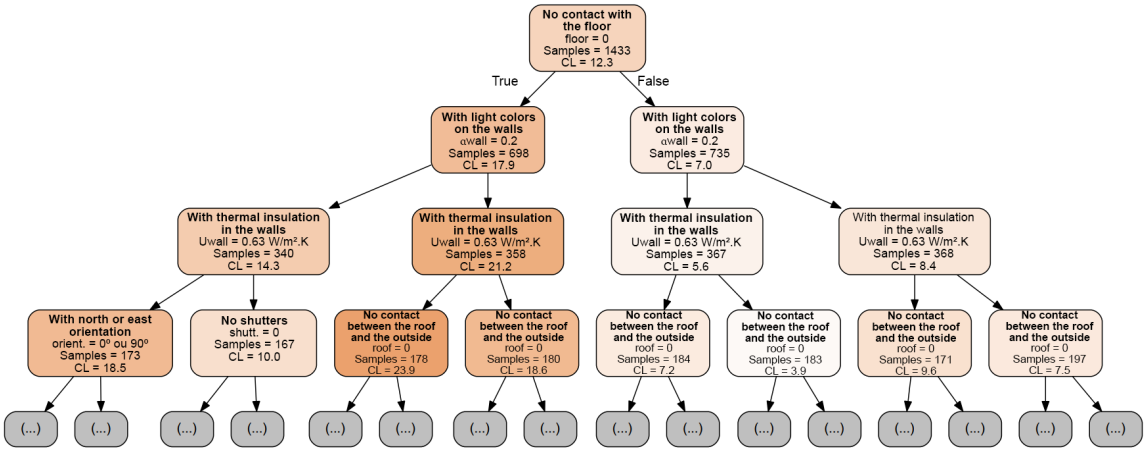
Source: The authors.

Figure 3 – Three first level parameter of the selected tree referred to the heating load prediction.



Source: The authors.

Figure 4 – Three first parameters of the selected tree referred to the cooling load prediction.



Source: The authors.

Therefore, a methodology was defined for the application of a pruning hyperparameter to the implementation, as well as the application of *Grid Search* and k-*fold* Cross Validation techniques for hyperparameter optimization. By doing that, metrics of error analysis were defined to obtain the prediction heating loads, the average number of leaves of tree 1116.70 $\pm$ 26.727, the values of 0.25 $\pm$ 0.015 kWh/m^2.year for MAE, 0.17 $\pm$ 0.025 (kWh/m^2.year)^2 for MSE, 1.11 $\pm$ 0.063% for MAPE and 0.9983 $\pm$ 0.0003 for R^2. For the prediction of cooling loads, the average number of leaves of the trees generated was 1085.52 $\pm$ 38.567, with values of 0.16 $\pm$ 0.017 kWh/m^2.year for MAE, 0.09 $\pm$ 0.040 (kWh/m^2.year) ^{2} for MSE, 1.52 $\pm$ 0.145% for MAPE and 0.9980 $\pm$ 0.0009 for R^2. The sensibility analysis allowed to understand the heating load predictions, the parameters of biggest influence are the thermal transmittance of the walls, followed by the floor contact, roof contact with the outside and solar absorptance of the walls. For the cooling load predictions, the results of the sensibility analysis revealed a bigger influence of floor contact followed by the solar absorptance and thermal transmittance of the walls.

Finally, the greatest contribution of this study, is the possibility of performing a visual analysis of the generated tree. This allowed to identify which combinations of energy efficiency measures that resulted in lower values of thermal load in dwellings executed in light construction systems.

Acknowledgments

To the CNPq (National Research Council) by the scholarship provided.

References

ABNT - ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICA. *NBR 15575*: edificações habitacionais Desempenho. Rio de Janeiro: ABNT, 2013.

BOURDEAU, M. *et al*. Modeling and forecasting building energy consumption: a review of data-driven techniques. *Sustainable Cities and Society*, Amsterdam, v. 48, p. 101533, 2019.

BRASIL. Ministério do Desenvolvimento, Indústria e Comércio Exterior. Instituto Nacional de Metrologia, Normalização e Qualidade Industrial. *Portaria no 42, de 24 de fevereiro de 2021*. Instrução Normativa Inmetro para a Classificação de Eficiência Energética de Edificações Comerciais, de Serviços e Públicas (INI-C). Brasília: INMETRO, 2021a. 139 p.

BRASIL. Ministério do Desenvolvimento, Indústria e Comércio Exterior. Instituto Nacional de Metrologia, Normalização e Qualidade Industrial (INMETRO). *Consulta pública no18, 12 de julho de 2021*. Instrução Normativa Inmetro para a Classificação de Eficiência Energética de Edificações Residenciais (INI-R). Brasília: INMETRO, 2021b. 78 p.

JUSTINO, M. P.; SILVA, F. S.; SILVA RABELO, O. Perspectiva de uso da inteligência artificial (IA) para a eficiência energética em prédios públicos. *Cadernos de Prospecção*, Salvador: v. 13, n. 3, p. 769, 2020.

MARSLAND, S. *Machine learning*: an algorithmic perspective. London: Chapman and Hall: CRC, 2011.

MELO, A. P. *Desenvolvimento de um método para estimar o consumo de energia de edificações comerciais através da aplicação de redes neurais*. 2012. Tese (Doutorado) - Universidade Federal de Santa Catarina, 2012.

NUNES, G. H.; MOURA, J. D. M.; GÜTHS, S.; ATEMA, C.; GIGLIOA, T. Thermo-energetic performance of wooden dwellings: Benefits of cross-laminated timber in Brazilian climates. *Journal of Building Engineering*, [London], v. 32, p. 101468, 2020. DOI: https://doi.org/10.1016/j.jobe.2020.101468.

OLINGER, M. S. *Predição de conforto térmico em escritórios ventilados naturalmente por meio de redes neurais artificiais*. 2019. Dissertation (Master's) – University Federal of Santa Catarina, Florianópolis, 2019.

PEDREGOSA, F.; PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, P.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, M. Scikit-learn: machine learning in Python. *The Journal of machine Learning research*, Cambridge, v. 12, p. 2825-2830, 2011

PHAM, A. D. *et al*. Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production*, Oxford, v. 260, p. 121082, 2020. DOI: https://doi.org/10.1016/j.jclepro.2020.121082.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. *In*: ENCYCLOPEDIA of database systems. New York: Springer-Verlag, Springer, 2009. p. 532–538.

SEYEDZADEH, S.; RAHIMIAN, F. P.; RASTOGIC, P.; GLESKA, I. Tuning machine learning models for prediction of building energy loads. *Sustainable Cities and Society*, Amsterdam, v. 47, p. 101484, 2019.

SUN, Y.; HAGHIGHAT, F.; FUNG, B. C. A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and Buildings*, Lausanne, v. 221, p. 110022, 2020.

TSANAS, A.; XIFARA, A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, Lausanne, v. 49, p. 560-567, 2012.

WALKER, S.; KHAN, W; KATIC, K. MAASSENA, W.; ZEILER, W. Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. *eEnergy and Buildings*, Lausanne, v. 209, p. 109705, 2020.

ZARA, R. B. *Influência dos parâmetros termofísicos no desempenho térmico de edificações residenciais em sistemas construtivos leves*. 2019. Master's (Dissertation in Civil Engineering - Londrina State University, Londrina, 2019.

ZEKIĆ-SUŠAC, M.; MITROVIĆ, S.; HAS, A. Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. *International Journal of Information Management*, Amsterdam, v. 58, p. 102074, 2021.

Received: Feb. 8, 2022
Accepted: May 19, 2022
Published: June 5, 2022