encit 2020

Associação Brasileira de Engenharia e Ciências Mecânicas

18th Brazilian Congress of Thermal Sciences and Engineering
November 16–20, 2020 (Online)

ENC-2020-0397
PREDICTION OF RESIDENTIAL BUILDINGS EFFICIENCY BASED ON DIFFERENTIAL EVOLUTION OPTIMIZATION AND RANDOM FOREST MODEL

Matheus Henrique Dal Molin Ribeiro[1,2]
mribeiro@utfpr.edu.br
Ramon Gomes da Silva[1]
gomes.ramon@pucpr.edu.br
Viviana Cocco Mariani[3,4]
viviana.mariani@pucpr.br
Leandro dos Santos Coelho[1,4]
leandro.coelho@pucpr.br

[1] Industrial & Systems Engineering Graduate Program (PPGEPS), Pontifical Catholic University of Parana (PUCPR). 1155, Rua Imaculada Conceição, Curitiba, PR, Brazil. 80215-901
[2] Department of Mathematics, Federal Technological University of Parana (UTFPR). Via do Conhecimento, KM 01 - Fraron, Pato Branco, PR, Brazil. 85503–390
[3] Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Parana (PUCPR). 1155, Rua Imaculada Conceição, Curitiba, PR, Brazil. 80215-901
[4] Department of Electrical Engineering, Federal University of Parana (UFPR). 100, Avenida Coronel Francisco Heraclito dos Santos, Curitiba, PR, Brazil. 81530-000

Abstract. *The objective of this paper is to develop an efficient ensemble learning model to predict the heating load (HL) and the cooling load (CL) of residential buildings, considering eight input variables (relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution). Feature engineering is an important step in predictive modeling, once the design of correct features can improve the models' predictive accuracy. For the eight input variables, thirty-three statistical features are obtained. Therefore, it is investigated the predictive performance in terms of mean squared error (MSE) for 10-fold cross-validation procedure with random forest (RF) model, when principal component analysis (PCA) and differential evolution optimization algorithm (DE) are employed during the feature selection process. The PCA is employed to reduce the feature space into the principal components to explain 95% of the data variability, and DE to select the most suitable set of inputs to predict HL and CL. Empirical results show that errors of DE-RF are lower than PCA-RF and RF to predict both outputs. The improvement on MSE achieved by DE-RF ranges between 11.58% - 11.63% regarding RF, and 9.73% - 61.41% regarding PCA-RF, for HL and CL, respectively.*

Keywords: *Building Energy, Differential Evolution, Feature Engineering, Principal Component Analysis, Random Forest*

1. INTRODUCTION

The buildings are the most consumers of energy worldwide. In the Brazilian context, they demand 51% of the produced energy. Especially, according to Brazilian Energetic Research Company (EPE) (2017), the residential sector is the greatest consumer, with 25.5% of representative. In this respect, due to this demand, they also can produce high levels of gases that cause the greenhouse effect. Therefore, designing an energy-efficient building is important. However, it is a challenge due to factors that affect this characteristic. In this context, several kinds of research have been conducted efforts to study the sensitivity of design variables to achieve the energy-efficient buildings (Silva and Ghisi, 2020), as well as the prediction of heating and cooling loads of residential buildings (Zhou *et al.*, 2020).

Especially, the prediction of heating (HL) and cooling (CL) loads of residential buildings plays a key role, once the heating, ventilation, and air conditioning system are drivers of energy demand (Wang *et al.*, 2018). In this respect, with the purpose to perform accurate predictions of HL and CL, according to simulated scenarios, the development of efficient models is important. For this purpose, ensemble learning models (Ribeiro *et al.*, 2020, 2019) can be adopted, which are the kernel of the predictive modeling on the last years. This class of models is characterized by the use of a set of learners to solve the same problem and achieve high accuracy. The widely adopted ensemble learning models are the bagging, boosting and stacking models (Ribeiro and Coelho, 2020; Moreno *et al.*, 2019). Through the use of these approaches, it is possible to obtain answers by varying some building design parameters once a model has been adequately trained (Tsanas and Xifara, 2012).

Aiming to generate an efficient ensemble learning model to predict HL and CL, the use of feature engineering (FE) can generate additional features to give some insights for the developed model. However, in the same way that additional features can help to improve the models' accuracy, redundant features can hamper the learning process (Thom de Souza *et al.*, 2020). Faced with this, it is necessary to develop feature selection (FS) approach, which is a tool usually employed to obtain a set of non-redundant features which can improve the accuracy metric of the predictive model.

Therefore, the objective of this paper is to evaluate the predictive performance of the model composed by FE, FS, and ensemble learning models. In this respect, the bagging approach named random forest (RF) is used to predict HL and CL of residential buildings. The adopted dataset is composed of two output variables (HL and CL), and eight input variables (features). To achieve high accuracy, before training RF model, FE is performed, specifically, statistical features such as average, standard deviation, skewness, minimum, and maximum are computed for each output over the eight features. Moreover, for each one of eight features, exponential of order two and three, hyperbolic tangent, and logarithm are computed. With the purpose to obtain the most suitable set of inputs two strategies are adopted. The first strategy uses an evolutionary algorithm called differential evolution (DE) optimization for FS, and the second uses principal component analysis (PCA) for dimensional reduction, both are coupled with the RF model generating two different models, named as DE-RF and PCA-RF, respectively. Also, the performance of DE-RF, PCA-RF, and RF are computed in terms of mean squared error (MSE). The 10-fold cross-validation and 30 independent runs are adopted, as proposed by Tsanas and Xifara (2012). At the end of the process, the importance of selected features for the predictive model is computed.

The main contributions of this paper are

- Comparison of different approaches to perform FS (evolutionary approach using DE optimization algorithm) and dimensional reduction (PCA) coupled with ensemble learning model (RF)

- Development of an integrated and accurate framework which employs FE, FS and ensemble learning models to predict HL and CL of residential buildings.

The remainder of this paper is organized as follows: In Section 2.1 a brief description of the dataset adopted in this paper is presented. The methods applied in this study are described in Section 2.2. Results are mentioned in Section 3. Finally, Section 4 concludes this study with considerations and some directions for future research proposals.

2. MATERIAL AND METHODS

This section summarizes the concepts and steps which are used to perform the predictive modeling.

2.1 Dataset

The dataset adopted in this paper was proposed by Tsanas and Xifara (2012). The energy efficiency analysis is performed using 12 different building shapes (taking the elementary cube - 3.5 × 3.5 × 3.5) simulated in Ecotect. The buildings differ concerning the glazing area, the glazing area distribution, and the orientation, amongst other parameters. Various settings as functions of the mentioned characteristics to obtain 768 building shapes were simulated. The dataset comprises 768 samples and 8 features (X_1 to X_8), aiming to predict two real-valued responses (CL - Y_1 and HL - Y_2), and can be obtained in the UCI Machine Learning Repository at `https://archive.ics.uci.edu/ml/datasets/energy+efficiency`. Table 1 presents the variables and statistical indicators minimum, median, mean, maximum, and standard deviation (Std). Additional details are provided in Tsanas and Xifara (2012).

2.2 Methods

This subsection presents the concepts employed in this paper to perform predictive modeling.

2.2.1 Random Forest

The RF is an ensemble learning method which uses several decision trees to predict or classify some target variable. It is a combination of bootstrap aggregation with predictors selected randomly to compose each node of the decision tree. In general, the main objective of this approach is to improve the performance of regression trees through the reduction of their variance (Breiman, 2001). Also, the RF once uses decision trees in the training process, is robust to deal with multicollinearity issue. In the RF model, the number of predictors of each node should be tuned. In this paper, they are defined by grid-search during the training process.

Table 1: Description and statistical indicators of the input (**X**) and output (**Y**) variables

Variable	Description	Statistical Indicator				
		Minimum	Median	Average	Maximum	Std
X_1	Relative Compactness	0.62	0.75	0.76	0.98	0.10578
X_2	Surface Area	514.5	673.75	671.71	808.50	88.08612
X_3	Wall Area	245	318.5	318.5	416.50	43.62648
X_4	Roof Area	110.25	183.75	176.60	220.50	45.16595
X_5	Overall Height	3.5	5.25	5.25	7	1.75114
X_6	Orientation	2	3.5	3.5	5	1.11876
X_7	Glazing Area	0	0.25	0.2344	0.4	0.13322
X_8	Glazing Area Distribution	0	3	2.8125	5	1.55096
Y_1	Heating Load	6.01	18.95	22.3072	43.1	10.09020
Y_2	Cooling Load	10.9	22.08	24.5878	48.03	9.51331

2.2.2 Feature Engineering

The FE is the task of formulating the most appropriate features (numerical representation of the data) given the data, the model, and the task Zheng and Casari (2018). In most of the cases, statistical, and combination between several features are performed to obtain desirable features from the data.

In this paper, for the input features presented in Table 1, thirty-three features are obtained according to two strategies. The first, consists in compute the exponential of order 2 and 3, hyperbolic tangent (tanh), and logarithm (log) for the features X_1 to X_8. In the second strategy, for the respective output, for the set of features are computed the average, median, standard deviation, minimum, maximum, and skewness. After performing FE, for each variable response, there are forty one features.

2.2.3 Feature Selection

The FS is the process designed to solve the trade-off between finding the smallest subset of features with the best classification accuracy (Thom de Souza *et al.*, 2020). Generally, the common technique applied in this process is the wrapper method, which allows to try out subsets of features, which means that won't accidentally prune away features that are uninformative by themselves but useful when taken in combination (Zheng and Casari, 2018).

The use of optimization algorithms perform FS has been explored extensively in recent years as observed in de Rosa *et al.* (2020), Thom de Souza *et al.* (2020), Wei *et al.* (2020), and Aljarah *et al.* (2020). Therefore, in this paper, the wrapper approach which combines the RF model with the DE algorithm is employed. The DE algorithm was introduced by Storn and Price (1995) for optimization of nonlinear and non-differential continuous functions. It is a population optimizer, which starts the optimization process by search space sample in multiples initial points randomly chosen, i.e., a population of vectors representing the subjects or candidate solutions. The DE algorithm consists of the maintenance of a population of the candidate solutions subject to crossing, evaluation and selection operations (de Vasconcelos Segundo *et al.*, 2017). The step-by-step of the DE algorithm consists of four phases: initialization, mutation, crossover and selection. These operations are repeated generation after generation until that some stop criterion is satisfied. In this paper, the crossover and mutation operators are defined as equals 0.5 and 0.8, respectively.

By combining RF and DE, the optimization process starts generating an N-sized vector, where N is the total number of features in a dataset, and each position of the vector can only assume values such as 0 or 1, where 0 represents the features that were not selected and 1 represents the features that were selected. According to the select features, the RF is trained using 10-fold cross-validation, and the cost function is computed, the MSE. The process is conducted until the optimization process ends. Thus, the set of selected variables, those with higher values than 0.5 is obtained.

On the other hand, a second approach which can be employed, and in this case, to feature dimensionality reduction, is the PCA. The PCA is a technique widely employed for dimensional reduction or feature extraction coupled with machine learning models (Gárate-Escamila *et al.*, 2020; Gharsellaoui *et al.*, 2020). The objective of this technique is to extract and maintain only the most relevant and important features of the set. To do this, the PCA projects the original data into principal components (PC), which are combinations of original features (Jolliffe, 2002). In this respect, when PCA is combined with RF, the set of correlated features generates the PC which can explain 95% (Ribeiro *et al.*, 2019) of the data variability, are new features, and then the RF model is trained using 10-CV to compute the MSE. While a percentage greater than 95 can lead to keeping information redundant to the data, less than that can exclude important components. For both process, PCA-RF, and DE-RF, the input features are centered by its mean and scaled by its Std.

2.2.4 Cross-Validation and Performance Evaluation

As a way to ensure out-of-sample generalization, the k-fold cross-validation can be employed. In this process, the data is randomly split into k subsets approximately of the same size, and while k-1 folds are used for training the model, the remaining fold is separated and used to validate the consequent model, resulting in an accuracy. This process is repeated k times, until all folds are used k-1 times for training and once for validation. The arithmetic mean is obtained from these k validations accuracy. In this paper, the number of cross-validation partitions, k, was defined as 10. For statistical confidence, the training and testing process is repeated 30 independent runs with the dataset randomly permuted in each run before splitting in training and testing subsets.

In this paper, the accuracy of the trained model is accessed through the MSE, which is computed as follows,

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{1}$$

where n represents the number of observations, y_i and $\hat{y}_i$ are the i-th observed and predicted values, respectively. Figure 1 presents the roadmap of proposed data analysis.

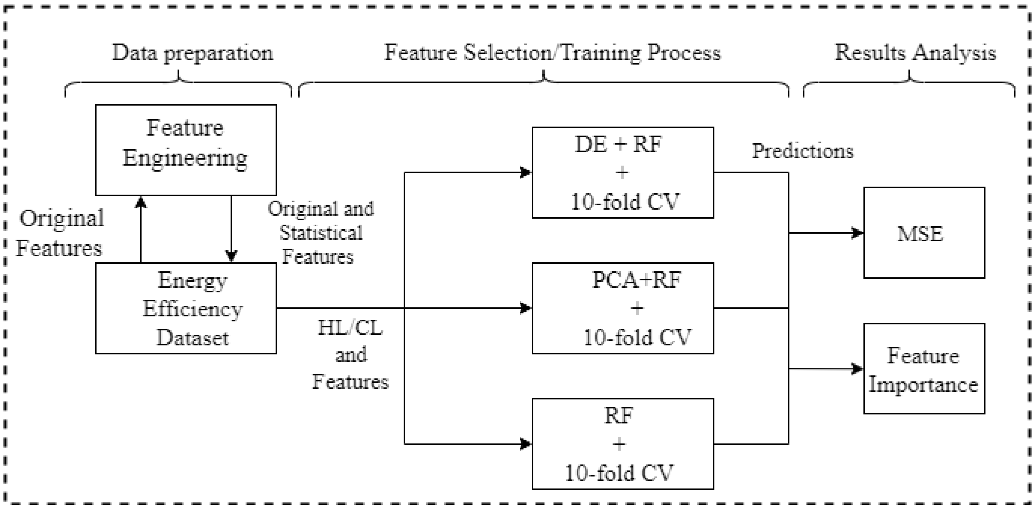


Figure 1: Roadmap of proposed framework

3. RESULTS

This section describes the main results obtained by proposed predictive modeling. Table 2 presents the results for the performance evaluation of each model, where best accuracy is highlighted in bold.

Table 2: Results of the ensemble learning models in terms of MSE (30 runs) in the prediction of the HL and CL.

Indicator	Output					
	HL			CL		
	PCA–RF	DE–RF	RF	PCA–RF	DE–RF	RF
Minimum	0.5016	**0.2201**	0.2510	2.8294	**2.4948**	2.9671
Median	0.5911	**0.2352**	0.2675	3.1833	**2.8521**	3.2089
Average	0.6128	**0.2365**	0.2675	3.1490	**2.8426**	3.2166
Maximum	0.8146	**0.2566**	0.2945	3.5681	**3.1132**	3.5126

Considering the HL, the best predictive model is the DE–RF, which uses DE for feature selection and RF to predict the output variable. In terms of average accuracy over 30 independent runs, this approach is able to enhance the accuracy, where the improvement in the performance is 61.41% and 11.58% regarding PCA–RF, and RF, respectively. On the other hand, for CL variable, the same results are achieved, where again, DE–RF provide better accuracy in terms of MSE, regarding to compared models. In fact, the improvement in the performance is 9.73% and 11.63% regarding to PCA–RF, and RF, respectively.

Figures 2 and 3 illustrate the scaled feature importance of each feature selected by the DE algorithm and used by RF ensemble learning model to predict the HL, and CL, respectively. Considering the computed features, for both cases, the feature base on Glazing area (X_7) has high importance compared with remain features, especially, exponential or order two and three for HL and CL. Also, the hyperbolic tangent and median features are important to predict HL and CL, respectively. Considering the original dataset features (X_1 to X_8), for HL the most important features are relative compactness (X_1), overall height (X_5), and glazing area distribution (X_8). Moreover, to predict CL, roof area (X_4) and glazing area distributions are the most influential.

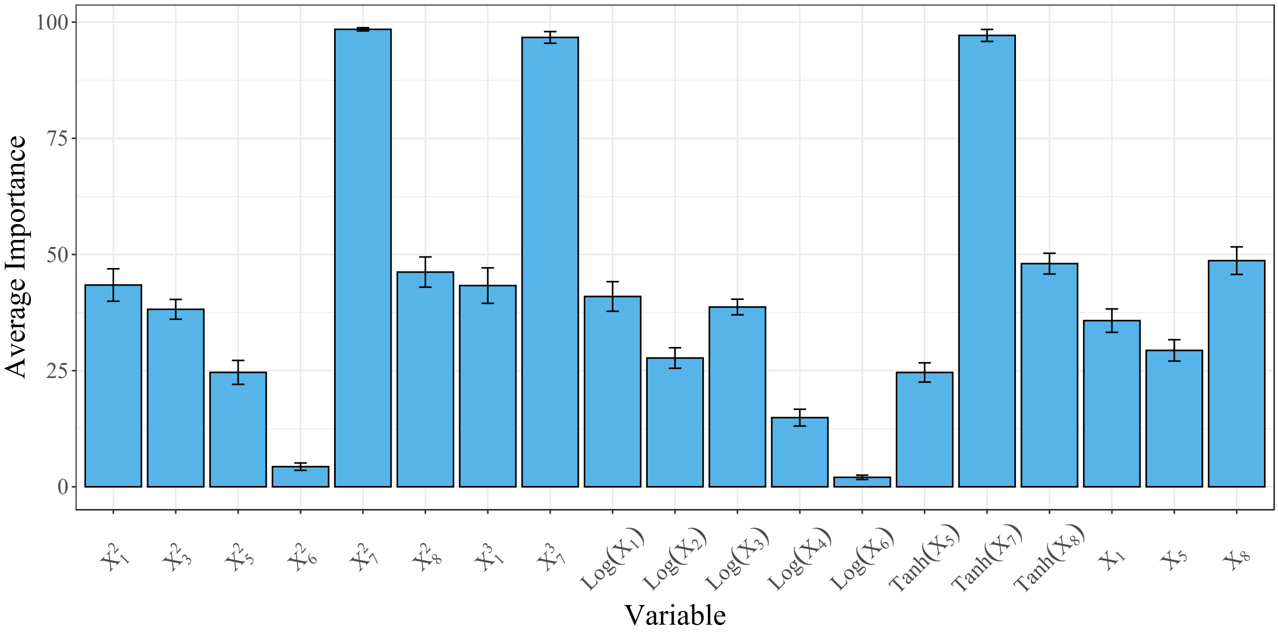


Figure 2: Feature Importance for HL

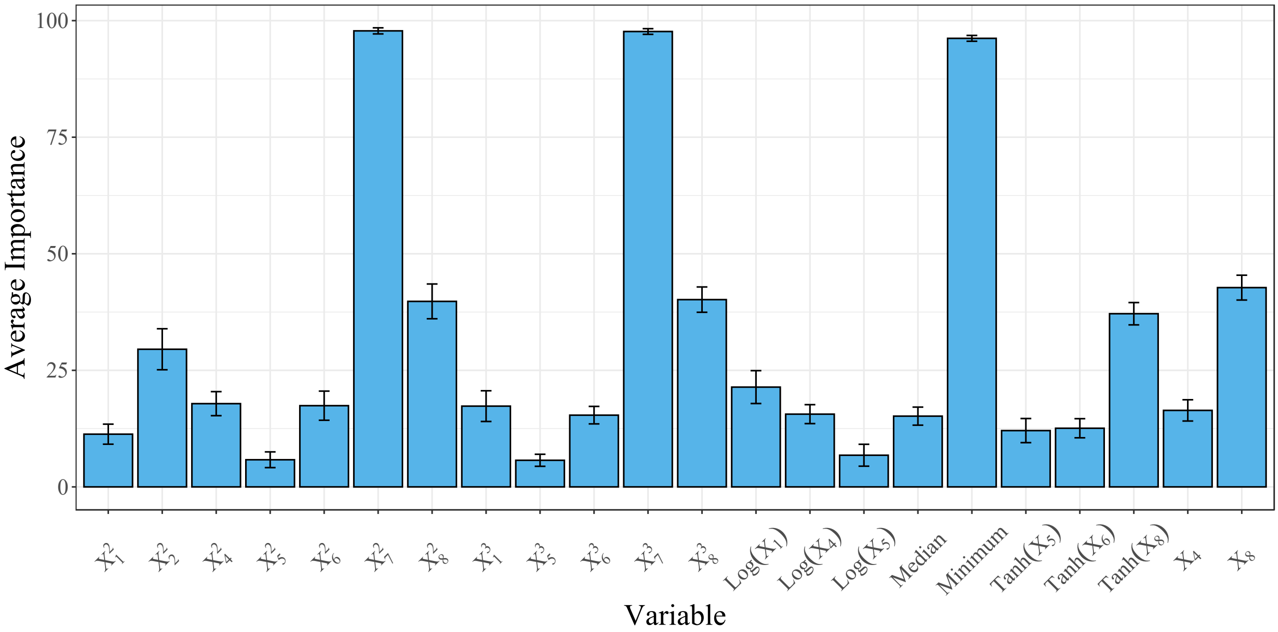
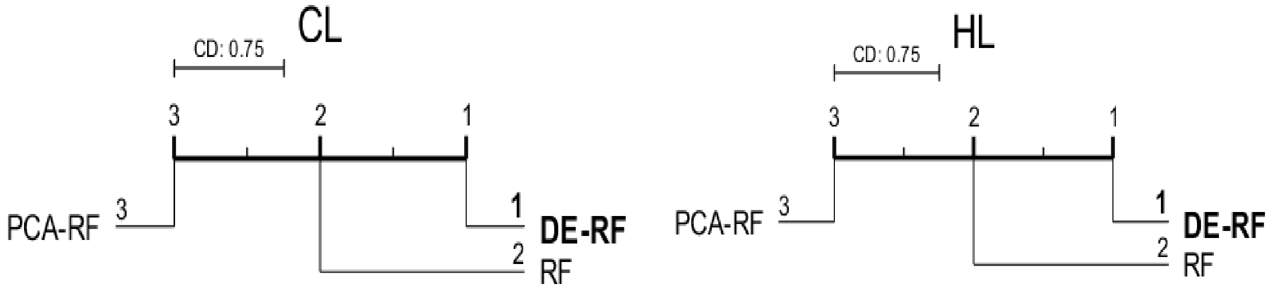


Figure 3: Feature Importance for CL

According to Friedman test, there is statistically difference between the MSE for three adopted approaches for the CL and HL variables ($\chi_2^2 = 60$, p-value < 0.05). Figure 4 illustrates the comparisons between three approaches obtained through Nemenyi post-hoc test. In this representation, those models that are not joined by a line can be regarded as different (Demšar, 2006). Moreover this information corroborates with results point out in Table 2.



(a) CL variable

(b) HL variable

Figure 4: Critical differences (CD) plots for different evaluated models.

The observed versus predicted HL (right) and CL (left) are depicted in Figure 5. While for HL output, the proposed framework which combines FE, FS, and ensemble learning model can capture the data variability, for the CL, this model had difficulty.

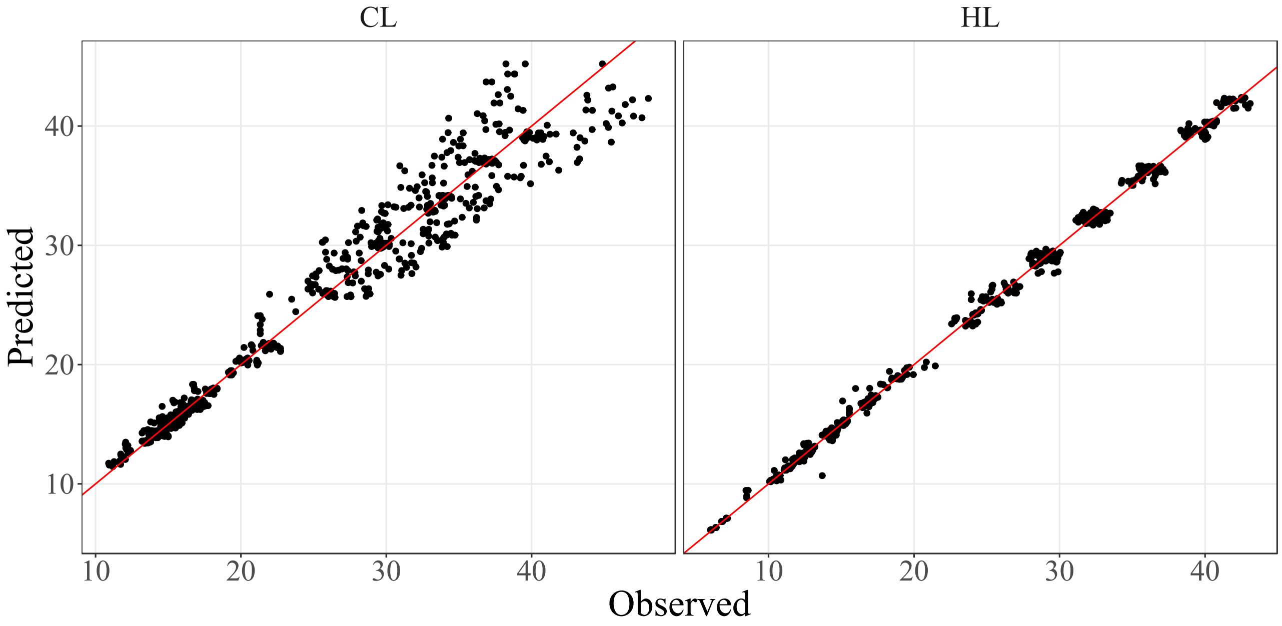


Figure 5: Observed versus Predicted values for CL (left) and HL (right) variables

4. CONCLUSION

In this paper was evaluated an ensemble learning model, specifically RF model to predict the energy efficiency of residential buildings in terms of HL and CL considering different inputs. To achieve the best accuracy, FE and FS were coupled with the RF model. The DE optimization algorithm was employed to perform feature selection, and PCA to perform the dimensional reduction. In this respect, the models DE-RF, PCA-RF, and RF were compared in terms of MSE to predict HL and CL.

The results illustrate that the proposed model can achieve good performance to predict the energy efficiency of residential buildings. In a broader perspective, the DE-RF outperforms PCA-RF and RF models in terms of MSE, which shows that perform feature selection through the optimization process is a promising approach. Moreover, through Friedman and Nemenyi post-hoc tests, it is possible to infer that DE-RF model has statistically smaller errors than PCA-RF and RF. As future research is desirable (i) to coupling stacking ensembles with DE for feature selection, (ii) to adopt different metaheuristics for FE such as the proposed by Thom de Souza *et al.* (2020).

5. ACKNOWLEDGEMENTS

The authors would like to thank the National Council of Scientific and Technologic Development of Brazil – CNPq (Grants number: 307958/2019-1-PQ, 307966/2019-4-PQ, 404659/2016-0-Univ, 405101/2016-3-Univ), PRONEX '*Fundação Araucária*' 042/2018, and *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001 for financial support of this work.

6. REFERENCES

Aljarah, I., Habib, M., Faris, H., Al-Madi, N., Heidari, A.A., Mafarja, M., Elaziz, M.A. and Mirjalili, S., 2020. "A dynamic locality multi-objective salp swarm algorithm for feature selection". *Computers & Industrial Engineering*, Vol. 147, No. 106628. ISSN 0360-8352. doi:10.1016/j.cie.2020.106628.

Breiman, L., 2001. "Random forests". *Machine Learning*, Vol. 45, No. 1, pp. 5–32. doi:10.1023/A:1010933404324.

de Rosa, G.H., Papa, J.P. and Yang, X.S., 2020. "A nature-inspired feature selection approach based on hypercomplex information". *Applied Soft Computing*, Vol. 94, No. 106453. ISSN 1568-4946. doi:10.1016/j.asoc.2020.106453.

de Vasconcelos Segundo, E.H., Amoroso, A.L., Mariani, V.C. and Coelho, L.d.S., 2017. "Economic optimization design for shell-and-tube heat exchangers by a tsallis differential evolution". *Applied Thermal Engineering*, Vol. 111, pp. 143–151. doi:10.1016/j.applthermaleng.2016.09.032.

Demšar, J., 2006. "Statistical comparisons of classifiers over multiple data sets". *Journal of Machine learning research*, Vol. 7, No. Jan, pp. 1–30.

Energetic Research Company (EPE), 2017. "Brazilian energy balance 2018-calendar year 2017".

Gharsellaoui, S., Mansouri, M., Trabelsi, M., Refaat, S.S. and Messaoud, H., 2020. "Fault diagnosis of heating systems using multivariate feature extraction based machine learning classifiers". *Journal of Building Engineering*, Vol. 30, p. 101221. ISSN 2352-7102. doi:10.1016/j.jobe.2020.101221.

Gárate-Escamila, A.K., Hajjam El Hassani, A. and Andrès, E., 2020. "Classification models for heart disease prediction using feature selection and pca". *Informatics in Medicine Unlocked*, Vol. 19, p. 100330. ISSN 2352-9148. doi: 10.1016/j.imu.2020.100330.

Jolliffe, I., 2002. "Principal component analysis for time series and other non-independent data". *Principal Component Analysis*, pp. 299–337.

Moreno, S.R., da Silva, R.G., Ribeiro, M.H.D.M., Fraccanabbia, N., Mariani, V.C. and dos Santos Coelho, L., 2019. "Very short-term wind energy forecasting based on stacking ensemble". In *14th Brazilian Computational Intelligence Meeting (CBIC)*. Belem, Pará, pp. 1–8. doi:10.21528/CBIC2019-22.

Ribeiro, M.H.D.M. and Coelho, L.S., 2020. "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series". *Applied Soft Computing*, Vol. 86, No. 105837. ISSN 1568-4946. doi: 10.1016/j.asoc.2019.105837.

Ribeiro, M.H.D.M., Ribeiro, V.H.A., Reynoso-Meza, G. and Coelho, L.S., 2019. "Multi-objective ensemble model for short-term price forecasting in corn price time series". In *International Joint Conference on Neural Networks (IJCNN)*. Budapest, Hungary, pp. 1–8. doi:10.1109/IJCNN.2019.8851880.

Ribeiro, M.H.D.M., da Silva, R.G., Fraccanabbia, N., Mariani, V.C. and dos Santos Coelho, L., 2019. "Forecasting epidemiological time series based on decomposition and optimization approaches". In *14th Brazilian Computational Intelligence Meeting (CBIC)*. Belem, Pará, pp. 1–8. doi:10.21528/CBIC2019-18.

Ribeiro, M.H.D.M., da Silva, R.G., Mariani, V.C. and Coelho, L.d.S., 2020. "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil". *Chaos, Solitons & Fractals*, Vol. 135. ISSN 0960-0779. doi: 10.1016/j.chaos.2020.109853.

Silva, A.S. and Ghisi, E., 2020. "Estimating the sensitivity of design variables in the thermal and energy performance of buildings through a systematic procedure". *Journal of Cleaner Production*, Vol. 244, No. 118753. ISSN 0959-6526. doi:10.1016/j.jclepro.2019.118753.

Storn, R. and Price, K., 1995. "Differential evolution-a simple and efficient adaptive scheme for global optimization over continuous space". *Technical Report TR-95-012, International Computer Science Institute*.

Thom de Souza, R.C., de Macedo, C.A., dos Santos Coelho, L., Pierezan, J. and Mariani, V.C., 2020. "Binary coyote optimization algorithm for feature selection". *Pattern Recognition*, Vol. 107, No. 107470. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107470.

Tsanas, A. and Xifara, A., 2012. "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools". *Energy and Buildings*, Vol. 49, pp. 560 – 567. ISSN 0378-7788. doi:10.1016/j.enbuild.2012.03.003.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R.S. and Ahrentzen, S., 2018. "Random forest based hourly building energy prediction". *Energy and Buildings*, Vol. 171, pp. 11 – 25. ISSN 0378-7788. doi:10.1016/j.enbuild.2018.04.008.

Wei, W., Chen, S., Lin, Q., Ji, J. and Chen, J., 2020. "A multi-objective immune algorithm for intrusion feature selection". *Applied Soft Computing*, Vol. 95, No. 106522. ISSN 1568-4946. doi:10.1016/j.asoc.2020.106522.

Zheng, A. and Casari, A., 2018. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.".

Zhou, G., Moayedi, H., Bahiraei, M. and Lyu, Z., 2020. "Employing artificial bee colony and particle swarm techniques for optimizing a neural network in prediction of heating and cooling loads of residential buildings". *Journal of Cleaner Production*, Vol. 254, No. 120082. ISSN 0959-6526. doi:10.1016/j.jclepro.2020.120082.

7. RESPONSIBILITY NOTICE

The authors are solely responsible for the printed material included in this paper.