

Green building factor in machine learning based condominium price prediction

Suraya Masrom¹, Thuraiya Mohd², Abdullah Sani Abd Rahman³

¹Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Malaysia

²Faculty of Architecture, Planning and Surveying, Universiti Teknologi MARA, Perak Branch, Malaysia

³Faculty of Sciences and Information Technology, Universiti Teknologi PETRONAS, Perak, Malaysia

Article Info

Article history:

Received Feb 26, 2021

Revised Dec 24, 2021

Accepted Jan 5, 2022

Keywords:

Decision tree
Deep learning
Green building
Price prediction
Random forest

ABSTRACT

The negative impact of massive urban development promotes the inclusion of green building aspects in the real estate and property industries. Green building is generally defined as an environmentally friendly building, which rapidly emerged as a national priority in many countries. Acknowledging the benefits of green building, Green Certificate and Green Building Index (GBI) has been used as one of the factors in housing prices valuation. To predict a housing price, a robust approach is crucial, which can be effectively gained from the machine learning technique. As research on green building with machine learning techniques is rarely reported in the literature, this paper presents the fundamental design and the comparison results of three machine learning algorithms namely deep learning (DL), decision tree (DT), and random forest (RF). Besides the performance comparisons, this paper presents the specific weight correlation in each of the machine learning models to describe the importance of the green building to the model. The results indicated that RF has been outperformed others while Green Certificate and GBI have only been slightly important in the DL model.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Thuraiya Mohd

Faculty of Architecture, Planning and Surveying, Universiti Teknologi MARA

Perak Branch, Malaysia

Email: thura231@uitm.edu.my

1. INTRODUCTION

The increasing demand for condominiums in urban areas has created environmental pollution issues. As a result, green building has been given wide attention by properties stakeholders such as investors, developers, occupiers, and consumers in combating the issue of environmental impact. In Malaysia, the real estate market has reached a high level of maturity, where house buyers are becoming more selective and demanding [1]. Besides prime location, green building elements such as attractive landscape, indoor air quality, non-toxic and sustainable materials, are regarded as higher preferences by the house owners [2]. Therefore, predicting the condominium prices with green buildings is becoming important for the developer as well as to the potential buyers so that they can acquire information on the condominium price trends.

Recently, machine learning [3], [4] has become a vital predictive approach in a variety of domains [5]–[7], including in real estate valuation. Several studies have proven the capability of machine learning in generating higher accurate results in housing price prediction [8], [9]. Despite the widespread use of machine learning in real estate valuation as well as in the building price prediction, the inclusion of the green building aspect is difficult to be found in the literature. Recent studies on green building mostly relied on conventional

approaches such as multiple linear regression. Therefore, the objective of this study is to extend the state-of-the-art of green building by focusing on machine learning price prediction models.

The existence of advanced techniques of machine learning can be deployed and to start with a few of them is highly crucial. Besides identifying the machine learning performances through prediction accuracy and time efficiency, another important question that needs to be answered is how the green building aspect contributed to the performances of different machine learning models. This issue has not been broadly discussed in the existing research reports.

One of the important steps in machine learning deployment is to select the machine learning algorithms [10]. In the real estate industry, tree-based machine learning [11], [12] and deep learning (DL) [13] have been utilized wisely. Two common types of tree-based machine learning are random forest (RF) and decision tree (DT). These two algorithms have been reported as the best outperforming machine learning when tested on different cases of housing price prediction such as in [14], where the researchers have looked into the contribution of economic variables to RF modeling of the real estate market. By viewing different multi-featured aspects, researchers in [15], studied the performances of the RF algorithm in the housing price prediction. Different in [16] and [17], the researchers focused on the DT algorithm for implementing housing price estimation. As in [18], DL was used for the application of predicting housing prices of the real estate market in Taiwan while researchers in [19] reported the advantages of DL over the autoregressive integrated moving average (ARIMA) model in the forecasting of housing prices. Smart real estate assessment [20], DL with XGBoost [21] and DL based on textual information [22] are among the advancements of DL research in the housing industries.

2. RESEARCH METHOD

2.1. Data collection and datasets

In this research, the scope of the building area is within the Federal Territory of Kuala Lumpur district. This is because most development of the condominium green building is located in this area [23]. The related data on the green condominium building of the Kuala Lumpur district were collected from the valuation and property service department [24], which after data cleaning, 240 records with 14 columns were used for the models. The dependent variable is the transaction price per square feet in Ringgit Malaysia (RM). The indicator of green building status was named as Green Certificate, which consists of the certification and the Green Building Index (GBI) code developed by the Malaysian Institute of Architects and the Association of Consulting Engineers Malaysia (ACEM) [25]. The green building status that has been given to the condominium, can be either certified or index (silver, gold and platinum). In the collected dataset, 81.7% are green building certified while the rest 18.3% are given by GBI index value. From the GBI index building, 5.4% are silver indexed, 12.5% are gold indexed and 0.4% with platinum indexed. By using Pearson correlation, the correlation of 13 independent variables to the condominium Transaction Price is listed in Table 1.

The correlation weights in Table 1 are a global weight, which indicates how important the variables are to the Transaction Price, beyond the specific prediction model. However, one interesting feature in RapidMiner software is specific dependency weight that calculates the important contribution of independent variables in a particular machine learning model. Therefore, although the Green Certificate variable contributed a very low correlation to the transaction price in general as shown in Table 1, it would be interesting to look at how important it is on the specific machine learning models.

Table 1. Selected independent variables with the Pearson correlation weight

Independent variable	Correlation weight
Green Certificate	0.032
Level Property Unit	0.062
Building Floor	0.070
Date of Transaction	0.135
Distance	0.190
Age of Building	0.227
Type of Property	0.282
No of Bedroom	0.242
Security of Building	0.350
Mukim	0.371
Population Density	0.458
Lot Area	0.714
Main Floor Area	0.714

2.2. Machine learning

The dataset was divided with a split training and testing approach with ratio 60:40 percentages. Thus, from the 240 records, 144 of the datasets were used for the machine learning training and 90 records were used for testing. For the DL algorithm, the configuration is given in Table 2.

For tree-based machine learning, preliminary experiments have been conducted to identify the optimal parameters. As for DT, the relevant parameter is the tree maximal depth. As listed in Table 3, six values of maximal depth have been observed. It has been identified that the lowest error rate was generated with 7 maximal depths.

RF is an extension of the DT algorithm, which has one additional parameter besides the maximal depth. As shown in Table 4, the maximal depth is dedicated to the internal sub-trees. Based on 12 configurations of sub-trees and depths, the most optimal setting has been presented by 20 numbers of trees and 7 maximal depths. This configuration generated 13.6% error rate.

All the experiments were implemented with the RapidMiner software tool in a notebook computer with 16 GB RAM. Figure 1 shows some of the processes in RapidMiner, which is to set the training and testing to 60:40 percentages. Therefore, from the 240 records, 144 were used for training and the rest 96 records for testing.

Table 2. DL configuration

Parameters	Configuration
Number of epochs	10
Output function	Linear
Inner function	Rectifier
Number of layers	4
Number of neurons per layer	Layer 1: 13 neurons Layer 2: 50 neurons Layer 3: 50 neurons Layer 4: 1 neuron

Table 3. DT optimal parameter

Maximal depth	Error rate
2	35
4	18
7	16.7
10	16.8
15	16.9
25	16.9

Table 4. RF optimal parameter

Number of trees	Maximal depth	Error rate (%)
20	2	31.3
60	2	30.4
100	2	31.8
140	2	30.8
20	4	17.2
60	4	17.8
100	4	18.0
140	4	18.0
20	7	13.6
60	7	15.9
100	7	15.1
140	7	15.5

3. RESULTS AND DISCUSSION

In this section, the results of the machine learning models are presented in two divisions. Firstly, the performance results of each machine learning algorithm are given regarding the prediction accuracy and processing time. Secondly, the weight of contributions of each independent variable to the building transaction price is analyzed to identify the importance of the GBI in each of the machine learning models.

3.1. The machine learning performances

Table 5 lists the performance results of each machine learning model. The R^2 indicates the correlation between predicted values and actual values [26] which, the more it nearer to 1, the fitter it is. Root

mean square error (RMSE) [27] explains the average distances/differences between the prediction and the actual values. Additionally, relative error describes how large the error is relative to the actual values in percentage ratio. It can be seen that all machine learning models performed very well to predict the transaction price with a lower error value.

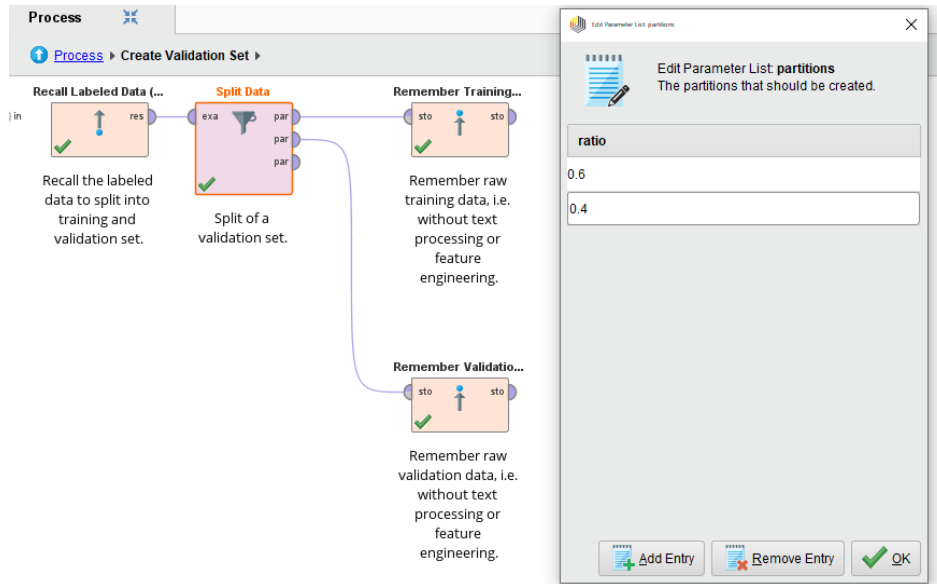


Figure 1. The ratio of training and testing for all the machine learning models

Table 5. The R[^] and root mean square error (RMSE) results

Machine learning algorithm	R [^] (+/-Std.Dev)	RMSE (+/-Std.Dev)	Relative Error (+/-Std.Dev)
Deep learning (DL)	0.932 (0.069)	684036.166(288080.644)	13.8% (2.8%)
Decision tree (DT)	0.848(0.278)	826146.506(295228.403)	14.3% (1.4%)
Random forest (RF)	0.936(0.116)	636316.217(371553.444)	12.6% (1.2%)

The most outperformed model from the three machine learning algorithms was RF with 94% fitness and 12.6% relative error. A very small difference has been generated by the DL algorithm with 0.93% fitness and 13.8% relative error. Although DT has presented the lowest performances compared to the two algorithms, the results were still within a good range, which is above 70% of R[^] and below 20% of relative error. Additionally, in terms of efficiency in completing the prediction, DT was the fastest algorithm as presented in Figure 2. It only took 106 seconds of total time to complete the training and prediction testing. Otherwise, both DL and RF have taken up to above 400 seconds of total times.

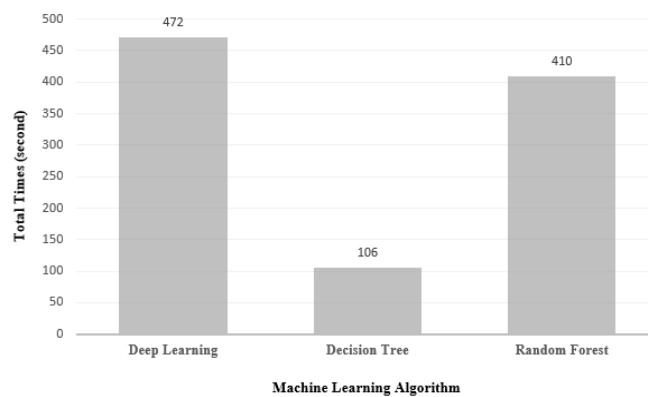


Figure 2. The efficiency of each machine learning model measured from the total times

Furthermore, the following Figures 3 to 5 can illustrate the accuracy of each model. The prediction charts show the prediction versus the actual values of the transaction price. The more predicted values (gray dot) that closer to the diagonal dotted line means the better is the machine learning.

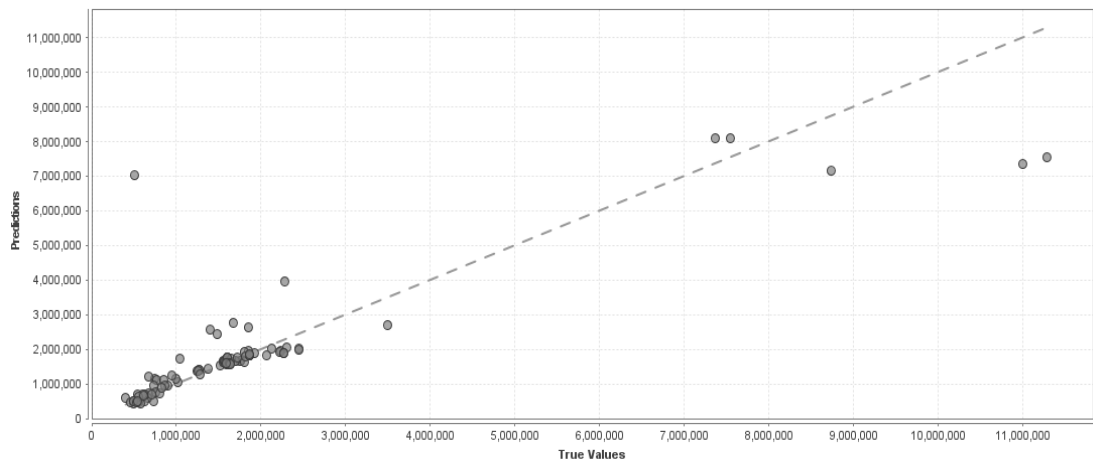


Figure 3. The prediction chart of DL

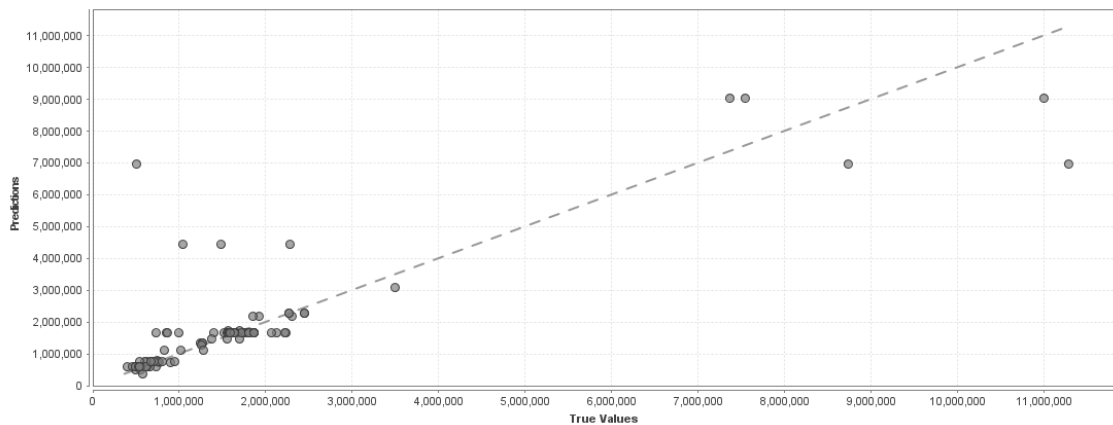


Figure 4. The prediction chart of DT

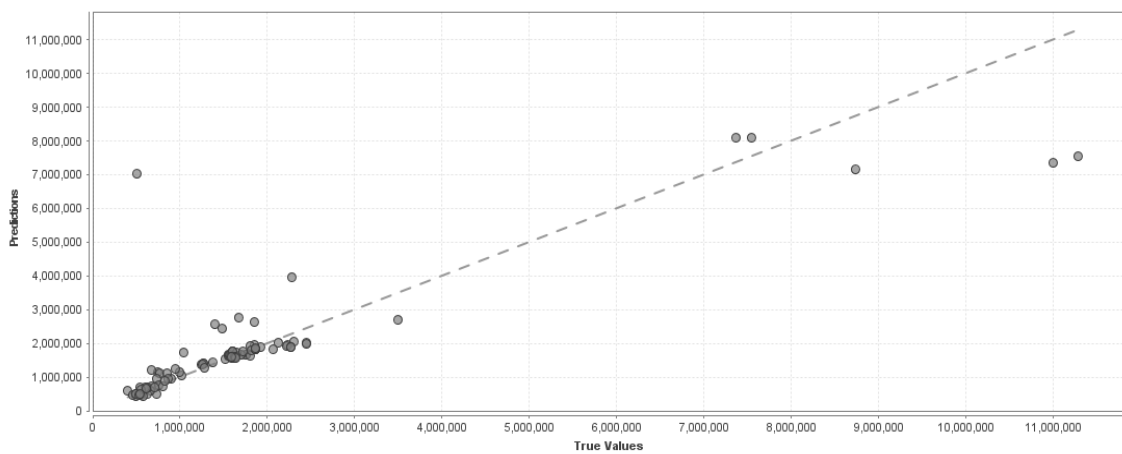


Figure 5. The prediction chart of RF

DT in Figure 4 has more gray dots that are far from the dotted line compared to DL and RF. Therefore, it shows that DT produced lower accuracy results compared to the other two algorithms. The results plotted in these prediction charts were consistent with the results listed in Table 5.

3.2. The correlations of variables in the machine learning models

This section explains the importance of each independent variable in each of the machine learning models. Figure 6 presents the weights of correlation of each independent variable/attribute in the DL model. It can be seen that only 11 out of the 13 attributes were contributed to the price prediction in DL.

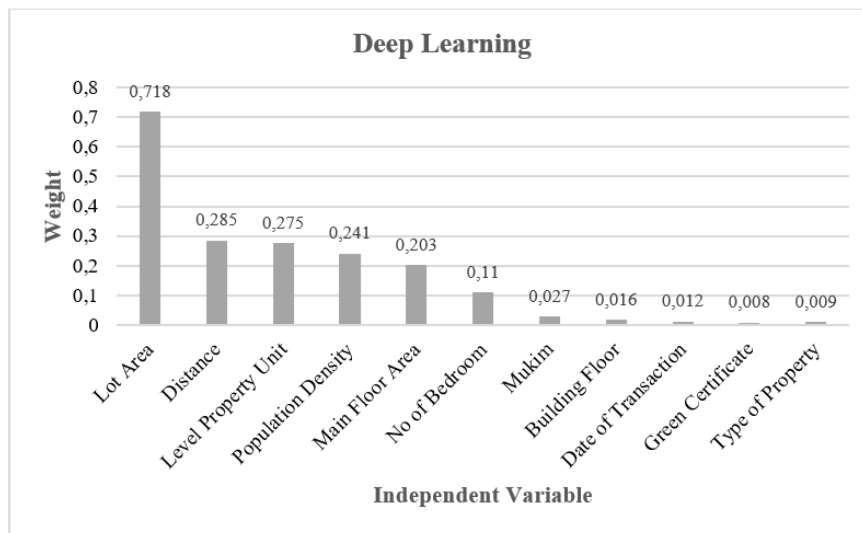


Figure 6. Weight of each independent variable in DL

It can be seen in Figure 6 that the Green Certificate has a very low weight (0.0078) to the prediction of transaction price in the DL model. Main floor area, which generally has a strong correlation with transaction price as given in Table 1 before, inversely has lower weight in DL (0.2). The only variable that has a very strong weight in DL is lot area (0.7 weight value). In contrast, main floor area has become the most important variable in the DT prediction model as presented in Figure 7.

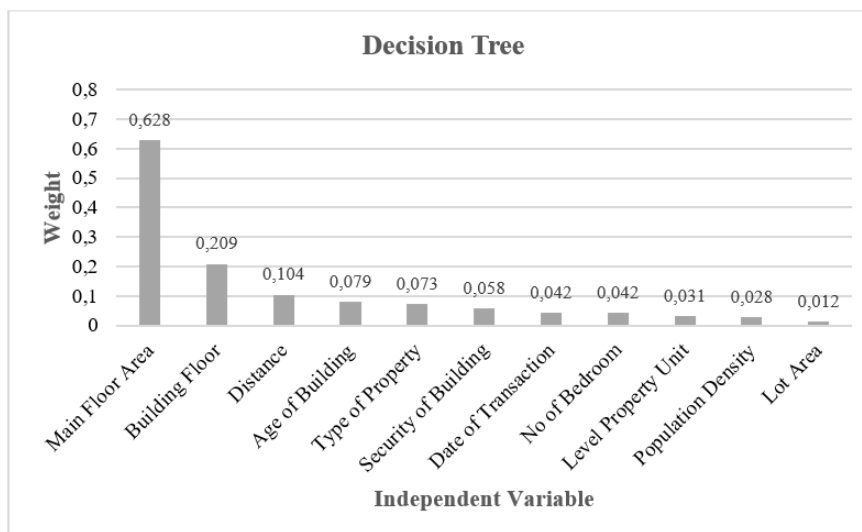


Figure 7. Weight of each independent variable in DT

In Figure 7, the green certificate seems to be no more significant in DT. Different from DL, main floor area is the most important to DT (0.6) but the weight of the lot area is only 0.012, which is very low compared to 0.7 in DL. Lastly, Figure 8 presents the correlation weights of the independent variables in RF.

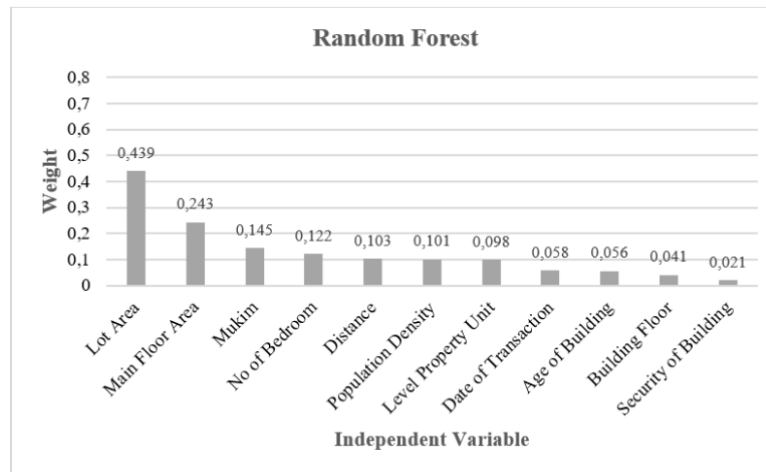


Figure 8. Weight of each independent variable in RF

Similar to DT, the Green Certificate was disappeared to have any weight in RF. Nevertheless, similar to DL, lot area played the most important role in RF. However, Lot Area in RF has a very low weight (0.439) compared to the weight in DL (0.718).

Regardless of each machine learning model, it can be generally viewed that Green Certificate and GBI have not yet been an important factor in the condominium price prediction for the area of the Federal Territory of Kuala Lumpur. This is maybe caused by the imbalanced GBI data distribution in the dataset. Additionally, the insufficient green building dataset can indicate that the green building prospect in Malaysia is still in its infancy.

4. CONCLUSION

This paper presents the report of research that seeks to address the role of green building through the Green Certificate and GBI in predicting the condominium price. Focused on DL, DT and RF algorithms, interpretation of the finding was described with some limitations on the methodology as well as on the tested dataset. The main challenge of this study is the limited size of green building datasets from the metropolitan area of Kuala Lumpur. Therefore, more extensive works are needed shortly for the green building as well as in the price prediction model either with the established conventional methods or with the robust machine learning approaches.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Teknologi MARA for the full support of this research.




REFERENCES

- [1] M. Solla, L. H. Ismail, A. sharainon M. Shaarani, and A. Milad, "Measuring the feasibility of using of BIM application to facilitate GBI assessment process," *Journal of Building Engineering*, vol. 25, Art. no. 100821, Sep. 2019, doi: 10.1016/j.job.2019.100821.
- [2] P. A. M. Khan, A. Azmi, N. H. Juhari, N. Khair, and S. Z. Daud, "Housing preference for first time home buyer in Malaysia," *International Journal of Real Estate Studies*, vol. 11, no. 2, pp. 1–6, 2017.
- [3] V. S. Padala, K. Gandhi, and P. Dasari, "Machine learning: the new language for applications," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 4, pp. 411–421, Dec. 2019, doi: 10.11591/ijai.v8.i4.pp411-421.
- [4] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, 2020.
- [5] N. Razali, S. Ismail, and A. Mustapha, "Machine learning approach for flood risks prediction," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 1, pp. 73–80, Mar. 2020, doi: 10.11591/ijai.v9.i1.pp73-80.
- [6] M. E. Amran *et al.*, "Optimal distributed generation in green building assessment towards line loss reduction for Malaysian public hospital," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1180–1188, 2019.




- [7] N. S. Ahmad Yasmin, N. A. Wahab, and A. N. Anuar, "Improved support vector machine using optimization techniques for an aerobic granular sludge," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 1835–1843, Oct. 2020, doi: 10.11591/eei.v9i5.2264.
- [8] T. D. Phan, "Housing price prediction using machine learning algorithms: the case of Melbourne City, Australia," in *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, Dec. 2018, pp. 35–42, doi: 10.1109/iCMLDE.2018.00017.
- [9] P.-Y. Wang, C.-T. Chen, J.-W. Su, T.-Y. Wang, and S.-H. Huang, "Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism," *IEEE Access*, vol. 9, pp. 55244–55259, 2021, doi: 10.1109/ACCESS.2021.3071306.
- [10] M. Praveena and V. Jaiganesh, "A literature review on supervised machine learning algorithms and boosting process," *International Journal of Computer Applications*, vol. 169, no. 8, pp. 32–35, Jul. 2017, doi: 10.5120/ijca2017914816.
- [11] R. S. Olson, N. Bartley, R. J. Urbanowicz, and J. H. Moore, "Evaluation of a tree-based pipeline optimization tool for automating data science," in *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, Jul. 2016, pp. 485–492, doi: 10.1145/2908812.2908918.
- [12] H. Shamsudin, M. Sabudin, and U. K. Yusof, "Hybridisation of RF(Xgb) to improve the tree-based algorithms in learning style prediction," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 4, pp. 422–428, Dec. 2019, doi: 10.11591/ijai.v8.i4.pp422-428.
- [13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [14] S. Levantesi and G. Piscopo, "The importance of economic variables on london real estate market: a random forest approach," *Risks*, vol. 8, no. 4, Art. no. 112, Oct. 2020, doi: 10.3390/risks8040112.
- [15] R. Sawant, Y. Jangid, T. Tiwari, S. Jain, and A. Gupta, "Comprehensive analysis of housing price prediction in pune using multi-featured random forest approach," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Aug. 2018, pp. 1–5, doi: 10.1109/ICCUBEA.2018.8697402.
- [16] H. Wu and C. Wang, "A new machine learning approach to house price estimation," *New Trends in Mathematical Science*, vol. 4, no. 6, pp. 165–171, Dec. 2018, doi: 10.20852/ntmsci.2018.327.
- [17] G.-Z. Fan, S. E. Ong, and H. C. Koh, "Determinants of house price: a decision tree approach," *Urban Studies*, vol. 43, no. 12, pp. 2301–2315, Nov. 2006, doi: 10.1080/00420980600990928.
- [18] C. Zhan, Z. Wu, Y. Liu, Z. Xie, and W. Chen, "Housing prices prediction with deep learning: an application for the real estate market in Taiwan," in *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, Jul. 2020, vol. 2020-July, pp. 719–724, doi: 10.1109/INDIN45582.2020.9442244.
- [19] F. Wang, Y. Zou, H. Zhang, and H. Shi, "House price prediction approach based on deep learning and ARIMA model," in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Oct. 2019, pp. 303–307, doi: 10.1109/ICCSNT47585.2019.8962443.
- [20] H. Xu and A. Gade, "Smart real estate assessments using structured deep neural networks," in *2017 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCCom/IOP/SCI)*, Aug. 2017, pp. 1–7, doi: 10.1109/UIC-ATC.2017.8397560.
- [21] Y. Zhao, G. Chetty, and D. Tran, "Deep learning with XGBoost for real estate appraisal," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec. 2019, pp. 1396–1401, doi: 10.1109/SSCI44817.2019.9002790.
- [22] X. Zhou, W. Tong, and D. Li, "Modeling housing rent in the atlanta metropolitan area using textual information and deep learning," *ISPRS International Journal of Geo-Information*, vol. 8, no. 8, Art. no. 349, Aug. 2019, doi: 10.3390/ijgi8080349.
- [23] S. M. Algburi, A. A. Faieza, and B. T. H. T. Baharudin, "Review of green building index in Malaysia; existing work and challenges," *International Journal of Applied Engineering Research*, vol. 11, no. 5, pp. 3160–3167, 2016.
- [24] K. J. Kam, S. Y. Chuah, T. S. Lim, and F. Lin Ang, "Modelling of property market: the structural and locational attributes towards Malaysian properties," *Pacific Rim Property Research Journal*, vol. 22, no. 3, pp. 203–216, Sep. 2016, doi: 10.1080/14445921.2016.1234361.
- [25] T. L. Mun, "The development of GBI Malaysia (GBI)," *Pam/Acem*, no. April 2008, pp. 1–8, 2009.
- [26] E. Kasuya, "On the use of r and r squared in correlation and regression," *Ecological Research*, vol. 34, no. 1, pp. 235–236, Jan. 2019, doi: 10.1111/1440-1703.1011.
- [27] W. Wang and Y. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," *IOP Conference Series: Materials Science and Engineering*, vol. 324, no. 1, Art. no. 012049, Mar. 2018, doi: 10.1088/1757-899X/324/1/012049.

BIOGRAPHIES OF AUTHORS






Associate Professor Ts. Dr. Suraya Masrom    is the head of Machine Learning and Interactive Visualization (MaLIV) Research Group at Universiti Teknologi MARA (UiTM) Perak Branch. She received her Ph.D. in Information Technology and Quantitative Science from UiTM in 2015. She started her career in the information technology industry as an Associate Network Engineer at Ramgate Systems Sdn. Bhd (a subsidiary of DRB-HICOM) in June 1996 after receiving her bachelor's degree in computer science from Universiti Teknologi Malaysia (UTM) in Mac 1996. She started her career as a lecturer at UTM after receiving her master's degree in computer science from Universiti Putra Malaysia in 2001. She transferred to the Universiti Teknologi MARA (UiTM), Seri Iskandar, Perak, Malaysia, in 2004. She is an active researcher in the meta-heuristics search approach, machine learning, and educational technology. She can be contacted through email at suray078@uitm.edu.my.



Thuraiya Mohd    is an Associate Professor in Universiti Teknologi MARA. Graduated with Ph.D. in Real Estate in 2012 and MSc in Land Development and Administration in 2003 from Universiti Teknologi Malaysia. Received Bachelor's degree in Estate Management from Universiti Teknologi MARA in 2001. Before joining Universiti Teknologi MARA, she served with Ismail & Co as a Valuation Executive for 2 years. Lecturing experience in UiTM, Perak Branch for 20 years in the Department of Estate Management. Lecturing experience includes teaching core courses of Valuation and Property Development to undergraduate students, Built Environment Theory to PhD students and Economics of Green Architecture to Masters students. She was awarded an Excellent University Community Transformation Centre (UCTC) Award 2015 by the Ministry of Education (MOE), MALAYSIA. Received a few grants (as leader) from the MOE and Ministry of Finance, MALAYSIA in the areas of property development, machine learning, and disaster management. Her current research interests focus on areas of green development, sustainable real estate, housing, and disaster management. She also has been involved in machine learning research that applied to real estate problems. She is currently one of the research members of the Machine Learning and Interactive Visualization Research Group at UiTM Perak Branch. She achieved Professional Qualifications as Surveyor (Sr) and is a professional member of the Board of Valuers, Appraisers, Estate Agents and Property Managers (BOVEAP). She can be contacted at email: thura231@uitm.edu.my.



Ts. Abdullah Sani Abd Rahman    obtained his first degree in Informatique majoring in Industrial Systems from the University of La Rochelle, France in 1995. He received a master's degree from Universiti Putra Malaysia in Computer Science, with specialization in Distributed Computing. Currently, he is a lecturer at the Universiti Teknologi PETRONAS, Malaysia and a member of the Institute of Autonomous System at the same university. His research interests are cybersecurity, data analytics and machine learning. He is also a registered Professional Technologist. He can be contacted at email: sani.arahman@utp.edu.my.