PLANNING MALAYSIA:
Journal of the Malaysian Institute of Planners
VOLUME 19 ISSUE 3 (2021), Page 411 – 422

MACHINE LEARNING FOR PROPERTY PRICE PREDICTION AND PRICE VALUATION: A SYSTEMATIC LITERATURE REVIEW

Nur Shahirah Ja'afar[1], Junainah Mohamad[2], Suriatini Ismail[3]

[1] *Faculty of Architecture, Planning and Surveying*
UNIVERSITI TEKNOLOGI MARA SHAH ALAM MALAYSIA
[2] *Department of Built Environment Studies & Technology,*
Faculty of Architecture, Planning and Surveying
UNIVERSITI TEKNOLOGI MARA PERAK BRANCH, MALAYSIA
[3] *Faculty of Architecture and Ekistics*
UNIVERSITI MALAYSIA KELANTAN, MALAYSIA

Abstract

Machine learning is a branch of artificial intelligence that allows software applications to be more accurate in its data predicting, as well as to predict current performance and improve for future data. This study reviews published articles with the application of machine learning techniques for price prediction and valuation. Authors seek to explore optimal solutions in predicting the property price indices, that will be beneficial to the policymakers in assessing the overall economic situation. This study also looks into the use of machine learning in property valuation towards identifying the best model in predicting property values based on its characteristics such as location, land size, number of rooms and others. A systematic review was conducted to review previous studies that imposed machine learning as statistical tool in predicting and valuing property prices. Articles on real estate price prediction and price valuation using machine learning techniques were observed using electronics database. The finding shows the increasing use of this method in the real estate field. The most successful machine learning algorithms used is the Random Forest that has better compatibility to the data situation.

***Keyword**:* machine learning, real estate, property price prediction, valuation

[2] Lecturer at Universiti Teknologi MARA, Seri Iskandar Campus. Email: mjunainah@uitm.edu.my

INTRODUCTION

Machine learning (ML) was first found in the early years and has since been further developed and vastly applied to date. ML technologies have grown and raised its capabilities across a suited of application (Ja'afar & Mohamad, 2021). ML is a computer program and a branch of artificial intelligence which is applied to identity, acquire and improve data performance carrying out its roles as a prediction model (Kamalov & Gurrib, 2021). In other words, ML learns from previous experiences to predict current performance and improve for future data. In ML, there are several categories learning such as supervised and unsupervised. Every learning category has several algorithms that studies different patterns. How ML works is done by studying previous pattersn using selected algorithms and predicts future result upon observations (Oladunni, 2016).

The benefits of ML includes to improve the performance of iterative algorithms by caching the previous accessed datasets Park & Kwon (2015) that avoids the overfitting on datasets which contain noise or many other features, able to predict unstable and unpredictable market with reasonable accuracy Sarip (2015), manufacturing, education, financial modelling, policing Jordan (2015), medicine (Christodoulou, 2019), transport system (Maalel, 2011), healthcare Ghassemi (2018) and engineering (Begel, 2019). This proved ML has a wide application and is being used in many sectors in addressing various issues.

METHODOLOGY

Based to the Preferred Reporting Items for Systematic Review (PRISMA) and Meta-Analyses, there are four steps involved in reviewing methodology e.g. starting with identification, screening, eligibility and finally, the inclusion (Lalu, Li, & Loder, 2021). In identification step, authors conduct literature searches by using electronic databases with variety of keywords to identify related articles. The objective of PRISMA is to ease the identification of the literature review. Authors studies several previous journals in producing a comprehensive literature review. Referring to the checklist studied by Lalu et al. (2021), it comprises items to be analysed such as defining clear research questions, identifies inclusion and exclusion criteria, besides examine a few database for scientific literature. In addition, in conducting literature review, researchers used two databases of peer-reviewed publication database which is Scopus and Web Science, it is the most profound database, used by many, to be the primary competitor database for citation analysis and journal ranking statistics. This subsection is classified as phase one in identification of literature review.

The first phase is to identify the related journal article relating to ML in real estate price prediction which leads to the search for systematic literature review based on the topic from two licensed database which is the Scopus and Web of Science. Through the advanced search, using the query string, the database has disclosed over 2,254 articles available. In identifying related

literature, the keywords and queries string information strategy were used during penetrating keywords. The query string search done by authors from Scopus includes TITLE-ABS-KEY ("machine learning") AND "real estate" AND "price" AND "price prediction" AND "predict" AND "real estate" OR "price predict" OR "property" OR "house" OR " housing"), ("Machine Learning" AND "Real Estate" AND "price AND predict"). Meanwhile for the query string search for Web of Science are TITLE-ABS-KEY "machine learning" AND "real estate" AND "price" AND "price prediction" OR "property" OR "house" OR "housing"), ("Machine learning" AND "Real Estate" AND "price AND predict".

The second phase is via screening identified literature; this is to choose the suitable article that is associated to this topic of study. Based on 135 literatures found, only 47 literatures are used for this study.

The third phase is the eligibility and exclusion, there are several eligibility and exclusion criterions that were decided by authors. First and foremost, will be based on the literature type, authors selected articles from journals with empirical data and theoretical data which means, non-research articles, book chapters and book series are excluded. Secondly, authors only focused on English publication and excluded non-English publication in order to avoid difficulty and confusing in translating. Third, authors only considered publications between 1999 until 2021, but emphasised more on articles published from the year 2009 and above, this is to observe the published articles development areas. This process only focused on the application of ML on real estate prediction and articles will be selected if it was social science indexed. Lastly, authors also focused the objectives of the studies that is the ML in real estate prediction in general, so that articles studies are compatible to any region.

Table 1: Criteria Selection

Criteria	Eligibility	Exclusion
Article Type	Research article and conference proceeding	Non-research article and book sections
Language	English	Non-English
Timeline	1999 until 2021	<2009
Indexes	Social Science and Web of Science	None
Countries	Any countries	None

The fourth phase is on the data abstraction and analysis. This phase consists significant information, whereby the data from the articles were assessed and analysed, this is to concentrate on specific studies to respond to this study's questions and objective. Through the collected data, the classification types of ML on the real estate price prediction are found and proceeded to analysis.

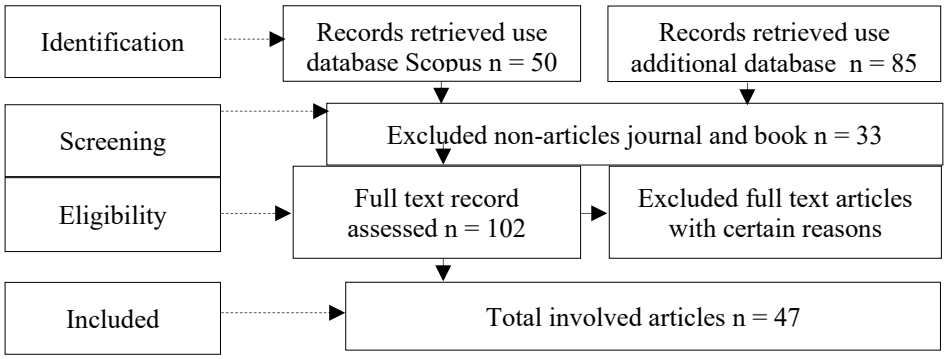

Figure 1: Articles Filtering Phase

Table 2: Previous Machine Learning on Real Estate Articles

Author	Country	Property	Involved Units	Method Supervised Machine Learning		Best Prediction
				Regression	Classification	
(Schernthanner, 2011)	Germany	Housing	74,098 Units	Random Forest	Nil	Random Forest
(Mu, Wu, & Zhang, 2014)	America		452 Units	Partial Least Square, Regression	Support Vector Machine, Least Square	Support Vector Machine
(Mccluskey & Daud, 2014)	Malaysia		313 Units	Linear Regression, Boosted Regression Tree	Nil	Boosted Regression Tree
(Oladunni & Sharma, 2015)	America		135 Units	Linear Regression, Gradient Boosting	Nil	Gradient Boosting
(Park & Kwon, 2015)	Virginia		5,359 units	Decision Trees, Ensemble	Naïve Bayesian	Ensemble
(Crosby & Davis, 2016)	UK		12,000 Units	Decision Tree, Random Forest	Nil	Decision Tree
(Oladunni, 2016)	America		2,075 Units	Principal Component Regression (PCA)	Support Vector Machine, k-Nearest Neighbors	PCA
(Valle & Crespo, 2016)	Chile		16,472 Units	Neural Network, Random Forest	Support Vector Machine	Random Forest

(Nejad, Lu, & Behbood, 2017)	Australia		1,967 Units	Random Forest, Ensemble, Decision Tree	Nil	Random Forest
(Trawiński & Telec, 2017)	Poland		12,439 units	Neural Networks, Linear Regression, Decision Tree	Nil	Decision Tree
(Horino & Nonaka, 2017)	Japan		6,320,631 Posts	Nil	Support Vector Machine	Support Vector Machine
(Gu & Xu, 2017)	China		253 Units	Linear Regression, Gradient Boosting	Nil	Gradient Boosting
(Di, Satari, & Zakaria, 2017)	India		21,000 Units	Linear Regression, Multivariate Regression, Polynomial Regression	Nil	Mix all models
(Kilibarda, 2018)	Serbia		7,407 Units	Linear Regression, Random Forest, PCA	Nil	Random Forest
(Ma & Zhang, 2018)	Beijing	Warehouse	25,900 Rental listings	Linear Regression, Random Forest, Gradient Boosting	Nil	Random Forest
(Varma & Sarma, 2018)	Mumbai	Housing	Nil	Linear Regression, Neural Network, Random Forest,	Nil	Neural Network
(Pow & Janulewicz, 2018)	Montreal		25,000 Units	Linear Regression, k-Nearest Neighbors, Random Forest	Support Vector Machine	k-Nearest Neighbors
(Dellstad, 2018)	Swedish		57,974 Units	Regression, Random Forest, Neural Network	Support Vector Machine	Random Forest
(Medrano & Delgado, 2019)	China		89 Units	Linear Regression, Support Vector Regression, Neural Network	k-Nearest Neighbors	Support Vector Regression

(Chardon & Javier, 2018)	Chile		334,353 units	Artificial Neural Network, Random Forest	Support Vector Machine	Random Forest
(Niu & Feng, 2019)	China		44,113 Units	Decision Tree, Gradient Boosting, Random Forest	Nil	Random Forest
(Mohd, Masrom, & Johari, 2019)	Malaysia		19 Features	Random Forest, Decision Tree, Ridge, Lasso, Linear Regression	Nil	Random Forest
(Mohamad, Ja'afar, & Ismail, 2020)	Malaysia		19 Features	Linear Regression, Decision Tree, Random Forest, Lasso, Ridge	Nil	Random Forest
(Ja'afar & Mohamad, 2021)	Malaysia		248 Units	Ridge, Lasso, Linear Regression, Decision Tree, Random Forest	Nil	Random Forest

LITERATURE REVIEW

Method and Statistical Price Prediction

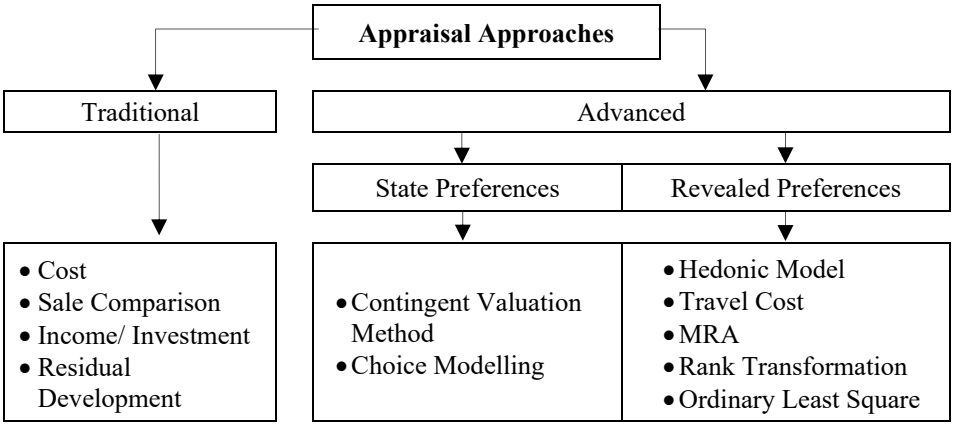

Figure 2: Diagram of Appraisal Approach

The market value of real estate is assessed through the valuation methods by following the existing procedures to reflect the nature and circumstances of the property to meet the market value definition. Every country has a different cultural and environment backgrounds, thus it has dissimilarity in determining the appropriate method for each particular property (Pagourtzi, Assimakopoulos,

& Thomas, 2003). In the valuation process, there are several methods used such as traditional and advanced method (Mohamad & Ismail, 2019). Traditional valuation method has been practiced in Malaysia and it has several methods as stated in the diagram. Due to the nature of the method which has several limitations and restrictions to produce accurate value, the advance method has been adopted in carrying out valuation prediction (Olanrewaju & Lim, 2018). Advanced method is better than traditional method in terms of the availability and amount of data to run which is less time consuming. In addition, in determining the price of real estates, valuation also have a different purpose in valuing market transaction to compulsory purchase, the purposes of this valuation are for sale report, accounting purpose, loan security, auction, insurance, taxation and investment (Pagourtzi et al., 2003).

In reference to the study by Nejad et al., (2017), authors indicated the accurate sale price of property transaction process is significant in price prediction. It is important to achieve fair value during transactions of valuable property for home sellers and buyers, similar to the price assessment which is also important for investors to learn better decision making in obtaining investment opportunities and as well as to avoid risk. However, there are also difficulties in producing accurate value because the prices of residential property are affected by various factors including the property location, neighbourhood, market condition and many other internal and external factors that should be considered in value real estate property.

The current ML is found to be a practical application in various fields (Jordan, 2015). As stated by Park & Kwon (2015), ML is a statistical learning of predictive analysis which is a subset of artificial intelligence used in various task such as modelling, designing, programming and recognition. Besides that, ML is about extracting knowledge from data input to deliver output. A study from Kilibarda (2018) stated that researchers have concluded in this twentieth century, ML is known as an alternative in model predicting.

Based on Figure 3 of ML, it shows ML has two types of learning algorithms. These two learnings process has different algorithms due to dissimilar method learn datasets condition. From the previous studies, the most common learning being used is supervised machine learning, this learning dataset is structured by human then this learning will learn and train the dataset automatically in producing result based from the previous data experiences. This learning consists of two types of problem such as regression and classification. In short, regression was used to identify or learn the value output of “distance” and “kilogram”, while classification was used for output e.g., “colour”, “talking” and “thinking”.

The application of ML must provide complete dataset for further prediction, ML will learn the given dataset in the entire system from start until the produced result. Previous study has applied ML techniques to predict various

study to observe the possibility to get an accurate result (Varma & Sarma, 2018). The hedonic price regression is mainly been used for inferences. In contrast, ML has a great potential in predictions.

Machine Learning Algorithms

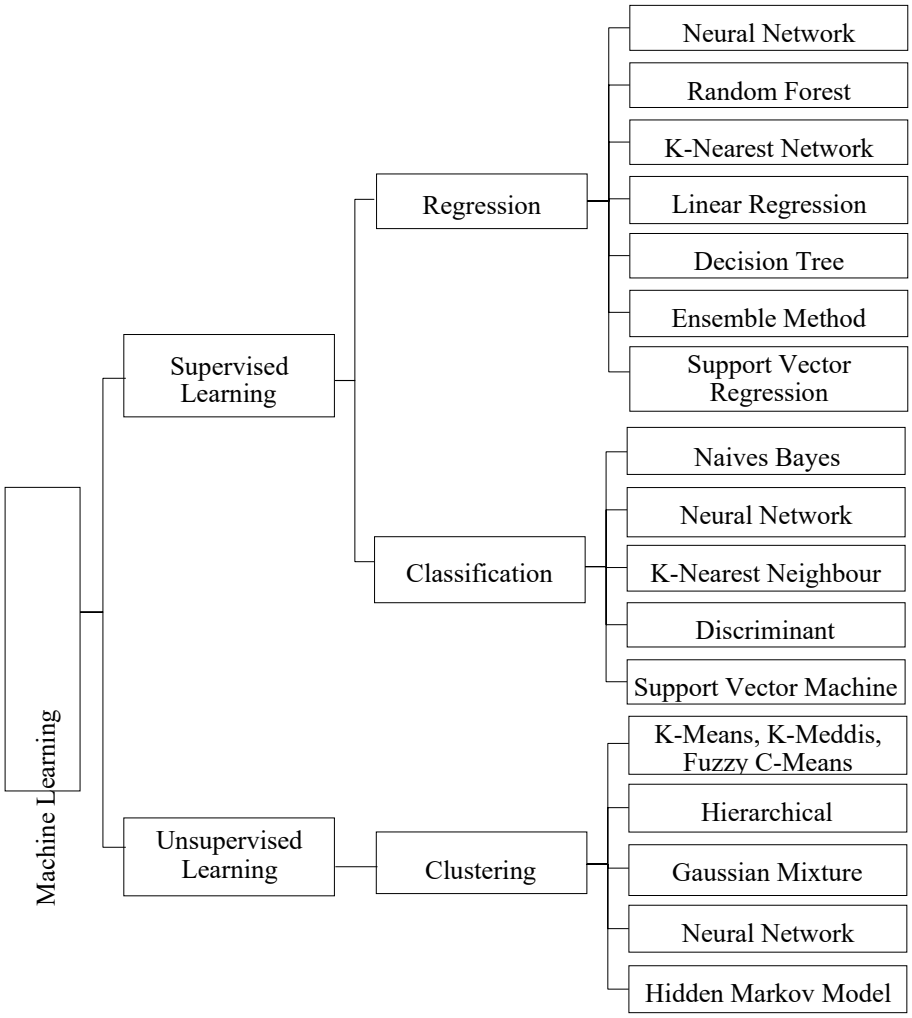

Figure 3: Types of Machine Learning Algorithms

RESULT AND DISCUSSIONS

Table 2 has presented 24 articles that have been selected in the ML price prediction studies. From previous studies of real estate price prediction using ML, most researchers applied supervised ML techniques. From previous literature

studies, authors have concluded that the ML techniques have been discovered in the real estate field in the past 5 years as property price prediction. The most popular learning used are supervised learning which is refer to regression and classification. Furthermore, from all the studies, there are few algorithms that has always been selected as the most popular model in prediction which is the algorithm of Random Forest, Decision Tree, Gradient Boosting, Neural Network and Linear Regression. The best model in prediction is the Random Forest, this algorithm can adapt well with dataset situation and produce accurate and effective results, it is also recommended by previous researchers to consider in study prediction.

The decision made by ML algorithm is based on its previous data experiences. The application of ML has been applied in many fields as predicting analysis. However, the use of ML in real estate industry in Malaysia is not often been discovered. Thus, authors have considered to use ML in real estate as a tool in predicting price of property to observe the significant improvement that can be achieved in price prediction. However, there may be some restrictions when exploring the ML algorithms in achieving accuracy.

There are several suggestions when it comes to ML algorithms for real estate price predictions. The first would be the most selected algorithm which is Random Forest (RF). This algorithm is known as one of ensemble method used in supervised, it is usually used as prediction in analysing the price of property and decision making in real estate (Shinde & Gawande, 2018). Random Forest has the same concept like Decision Tree by reproducing a large number of trees in forest then this algorithm will restart training sample and randomly choose features and observe to build decision tree and choose in improving the accuracy (Ja'afar & Mohamad, 2021). Besides that, Random Forest has less error than the other algorithms when it comes as property price prediction, this algorithm are among popular ML algorithm (Shinde & Gawande, 2018).

The second most used algorithm is the gradient boosting (GB) algorithm, it also known as another ensemble method created in ML. GB algorithm were built to construct new base learners to be optimal and relevant in learning the data. This algorithm used as prediction by convert weak learners to strong learner, usually in decision tree. From previous studies, GB generate accurate result as prediction of house price and it also shows good performance than LR algorithm, but RF is the best in estimating rental (Borde & Rane, 2017).

The third algorithm is the Support Vector Machine (SVM). This algorithm been applied in regression and classification learning, however it mainly used in classification. SVM were imposed to plot each data item or variable as a point in two dimensional spaces with the value of each feature of particular coordinate, besides that it also can reduce the bias problem during forecast and deliver a better simplification after consider the item of DV and IV

(Phan, 2019; Sarip, 2015). For the summary, SVM is among the best algorithm in supervised ML and mostly been used as recognition in classification learning.

Here, the fourth algorithm is Decision Tree (DT), this algorithm is a supervised learning that covered in both regression and classification. DT algorithm are used in visualise the decision, decision tree was drawn from root, branches and bottom. Commonly, DT used for variable selection, assess the significant connection of variables, monitor the missing value, prediction and data management (Song & Lu, 2015).

Lastly but not least, the fifth algorithm is Linear Regression (LR), this algorithm is one of the most frequent algorithms been used in predicting. This algorithm also known as Ordinary Least Square (OLS), Simple Linear Regression (SLR) and Multiple Linear Regression (MLR). The one independent in predicting will be SLR, while more than two independents will use MLR in predicting the linear between Y and X. From previous studies, LR is popular in house price prediction (Mayer, Bourassa, & Hoesli, 2019). Basically LR is used for prediction, forecasting and seeking to study plus to analyses the straight line relationship between variables (Borde & Rane, 2017).

CONCLUSION

This study has reviewed the application of machine learning techniques in real estate prediction analysis studied by previous researchers, from the result studied there are several existing machine learning algorithm techniques that have been applied. The types of real estate forecasting in appraisal approach and statistical learning been discussed in above sections. Besides that, the overview types of machine learning algorithms, number of transactions involved and result of several previous studies are also reported in this study. This study summaries few suggestions of potential machine learning algorithm for future studies. Authors strongly believe that this study is desirable, because of the outlined basic information of previous studies especially in discovering methods price prediction using machine learning techniques.

ACKNOWLEDGEMENTS

This study was supported by the grant of "FRGS" project: Automated Machine Learning Pipelines in Predicting the Price of Heritage Property. Project number: FRGS/1/2018/WAB03/UITM/03/1.

REFERENCES

Begel, A. (2019). Software Engineering for Machine Learning : A Case Study. *Microsoft Research*, 1–10.

Borde, S., & Rane, A. (2017). Real Estate Investment Advising Using Machine Learning. *Journal of Engineering and Technology (IRJET)*, *04*(03), 1821–1825.

Chardon, I., & Javier, F. (2018). *Housing Prices: Testing Machines Learning Methods*.

2–15.
Christodoulou, E. (2019). A Systematic Review shows no Performance Benefit of Machine Learning over Logistics Regression for Clinical Prediction Models. *Journal of Clinical Epidemiology*, *110*, 12–22.
Crosby, H., & Davis, P. (2016). *A Spatio-Temporal , Gaussian Process Regression , Real-Estate Price Predictor*. 3–6.
Dellstad, M. (2018). *Comparing Three Machine Learning Algorithms in the task of Appraising Commercial Real Estate*. KTH Royal Institute of Technology.
Di, N. F. M., Satari, S. Z., & Zakaria, R. (2017). *Real estate value prediction using multivariate regression models*
Ghassemi, M. (2018). *A Review of Challenges and Opportunities in Machine Learning for Health*.
Gu, G., & Xu, B. (2017). Housing Market Hedonic Price Study Based on Boosting Regression Tree. *Advanced Computational Intelligence and Informatics*, *21*(6).
Horino, H., & Nonaka, H. (2017). Development of an Entropy-Based Feature Selection Method on Real Estate. *Industrial & Engineering Management*, 2351–2355.
Ja'afar, N. S., & Mohamad, J. (2021). Application of Machine Learning in Analysing Historical and Non-Historical Characteristics of Heritage Pre-War Shophouses. *Journal of the Malaysian Institute of Planners*, *19*(2), 72–84.
Jordan, M. I. (2015). Machine learning: Trends, Perspectives and Prospects. *Journal of Electrical Engineering and Computer*, *349*(6245), 255–260.
Kamalov, F., & Gurrib, I. (2021). Financial Forecasting with Machine Learning: Price Vs Return. *Journal of Computer Science*, *17*(3), 251–264.
Kilibarda, M. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *Geo-Information*, 1–16.
Lalu, M. M., Li, T., & Loder, E. W. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Surgery*, *88*, 1–11.
Ma, Y., & Zhang, Z. (2018). Estimating Warehouse Rental Price using Machine Learning Techniques. *Journal of Computers Communications and Control*, *13*(April), 235–250.
Maalel, A. (2011). *The Contribution of Machine Learning to Analyze and Evaluate the Safety of Automated Transport Systems*. (May 2016), 1–7.
Mayer, M., Bourassa, S. C., & Hoesli, M. (2019). Estimation and Updating Methods for Hedonic Valuation. *Journal of European Real Estate Research*.
Mccluskey, W. J., & Daud, D. (2014). Boosted Regression Trees an Application for the Mass Appraisal of Residential Property in Malaysia. *Journal of Financial Management of Property and Construction*, *19*(2), 152–167.
Medrano, C., & Delgado, J. (2019). *Estimation of missing prices in real-estate market agent-based simulations with machine learning and dimensionality reduction*. *7*.
Mohamad, J., & Ismail, S. (2019). Capabilities of Revealed Preference Method for Heritage Property Valuation. *Journal of the Malaysia Institute of Planners*, *17*(1), 377–379.
Mohamad, J., Ja'afar, S., & Ismail, S. (2020). Heritage Property Valuation using Machine Learning Algorithms. *Annual Pacific Rim Real Estate Society*, 1–12.
Mohd, T., Masrom, S., & Johari, N. (2019). Machine Learning Housing Price Prediction in Selangor, Malaysia. *Recent Technology and Engineering (RTE)*, *8*(2), 542–546.

Mu, J., Wu, F., & Zhang, A. (2014). Housing Value Forecasting Based on Machine Learning Methods. *Abstract and Applied Analysis*, *2014*, 1–7.

Nejad, M. Z., Lu, J., & Behbood, V. (2017). *Applying Dynamic Bayesian Tree in Property Sales Price Estimation*.

Niu, W., & Feng, Z. (2019). Neural Network, Extreme Learning Machine and Support Vector Machine in Deriving Operation Rule of Hydropower Reservoir. *Journal of Water*, *11*(88). https://doi.org/10.3390/w11010088

Oladunni, T. (2016). Hedonic Housing Theory A Machine Learning Investigation. *Journal of Machine Learning and Application*, (December).

Oladunni, T., & Sharma, S. (2015). *Predictive Real Estate Multiple Listing System Using MVC Architecture and Linear Regression 1 Introduction*.

Olanrewaju, A. L., & Lim, X. Y. (2018). Factors Affecting Housing Prices in Malaysia: Analysis of the Supply Side. *Journal of the Malaysian Institute of Planners*, *16*(2), 225–235.

Pagourtzi, E., Assimakopoulos, V., & Thomas, H. (2003). Real Estate Appraisal : A review of Valuation Methods. *Property Investment & Finance*, *21*(4), 383–401.

Park, B., & Kwon, J. (2015). Using Machine Learning Algorithms for Housing Price Prediction Fairfax. *Journal of Expert System with Applications*, *42*(6), 2928–2934.

Phan, T. D. (2019). Housing Price Prediction using Machine Learning Algorithms : The Case of Melbourne, Australia. *Machine Learning and Data Engineering*, 35–42.

Pow, N., & Janulewicz, E. (2018). *Applied Machine Learning Project 4 Prediction of Real Estate Property Prices in Montreal*.

Sarip, A. G. (2015). Fuzzy Logic Application for House Price Prediction. *Journal of Property Sciences*, *5*(1), 24–30.

Schernthanner, H. (2011). *Spatial modeling and geovisualization of rental prices for real estate portals*.

Shinde, N., & Gawande, K. (2018). Valuation of House Prices using Predictive Techniques. *Journal of Advances in Electronics Computer Science*, *5*(6), 34–40.

Song, Y., & Lu, Y. (2015). Decision Tree Methods : Applications for Classification and Prediction. *Journal of Biostatistics in Psychiatry*, *27*(2), 130–136.

Trawiński, B., & Telec, Z. (2017). Comparison of Expert Algorithms with Machine Learning Models for Real Estate Appraisal. *Journal of Science and Technology*.

Valle, M. A., & Crespo, R. (2016). Property Valuation using Machine Learning Algorithms. *Journal of Modelling and Simulation*, 97–105.

Varma, A., & Sarma, A. (2018). House Price Prediction Using Machine Learning And Neural Networks. *Journal of Inventive Communication and Computational Technologies (ICICCT)*, 1936–1939.

Received: 12th July 2021. Accepted: 7th Sept 2021

