Rapid Modelling of Machine Learning in Predicting Office Rental Price

Thuraiya Mohd[1], Muhamad Harussani[2], Suraya Masrom[3]

GreensAFE (GreSFE) Research Group-Faculty of Architecture-Planning and Surveying-Department of Built Environment Studies and Technology, Universiti Teknologi MARA, Perak Branch, Seri Iskandar Campus, 32610 Perak, Malaysıa[1]

Centre of Graduate Studies, Universiti Teknologi MARA, Perak Branch, Seri Iskandar Campus, 32610 Perak, Malaysıa[2]

Malaysia Machine Learning and Interactive Visualization (MaLIV) Research Group Computing Sciences Study-College of Computing-Informatics and Media, Universiti Teknologi MARA, Perak Branch, Tapah Campus, 35400 Perak, Malaysia[3]

***Abstract*—This study demonstrates the utilization of rapid machine learning modelling in an essential case of the real estate industry. Predicting office rental price is highly crucial in the real estate industry but the study of machine learning is still in its infancy. Despite the renowned advantages of machine learning, the difficulties have restricted the inexpert machine learning researchers to embark on this prominent artificial intelligence approach. This paper presents the empirical research results based on three machine learning algorithms namely Random Forest, Decision Tree and Support Vector Machine to be compared between two training approaches; split and cross-validation. AutoModel machine learning has accelarated the modelling tasks and is useful for inexperienced machine learning researchers for any domain. Based on real cases of office rental in a big city of Kuala Lumpur, Malaysia, the evaluation results indicated that Random Forest with cross-validation was the best promising algorithm with 0.9 R squared value. This research has significance for real estate domain in near future, by applying a more in-depth analysis, particularly on the relevant variables of building pricing as well as on the machine learning algorithms.**

***Keywords*—Random forest; decision tree; support vector machine; rapid prediction modelling; office rental price**

I. Introduction

Now-a-days, real estate is becoming more digital, automated, and integrated. The fusion of industry 4.0 and digital 4.0 includes connected buildings, wearable technology, data management for buildings and infrastructure, and smart cities. The transformation of the real estate industry was improved due to the advancement in data science technologies such as analytic technologies [1]. The analytic technologies mentioned include Computational Statistics, Artificial Intelligence (AI) and Machine Learning. Machine learning is a sub-field of AI that can learn and re-learn from data exploration and inferences. Nowadays, these analytic technologies have successfully transformed the real estate industry to discover various opportunities, particularly by developing prediction applications that involve fundamental tasks that uncover hidden patterns, unknown correlations, and preferences [2]. Despite opportunities, some challenges appeared, including gaining adequate skills instantly that involves varying knowledge of AI concepts, mathematics, programming, and computer technologies. Thus, rapid software is useful to them and at the same time benefits the expert in accelerating the preliminary analytic tasks.

Considering real estate markets in general, office building markets are more synchronized in terms of exposure to macro-effects and performance of the real estate within the market. The heterogeneity of the office markets makes them more complex to analyse [3], [4]. It can be challenging to understand the market, for which the property's price might be determined on the market, but it may not always equate with the valuation of property in the market [5]. Office markets often relate to good investment opportunities since it draws much capital but with a substantial return [6]. Despite being a well-established investment industry, it has a highly complex market structure due to the lack of a central marketplace and the individuality of each building.

Numerous econometric models have been proposed to predict the office market performance, especially the rental property market. These include office market econometric models [5], and the hedonic regression model [6]. Sadly, limited success was achieved in finding a reliable and consistent model to predict rental property market movements over a five-to-ten-year time frame [7]. It was expected that lacking market data can be the main problem to fault the unreliability of prediction model. Based on the preliminary statistical analysis, the collected data of office rentals has a few problems of variance insufficient, imbalance with very skewed data distribution and most of them are having low dependencies to the target data (dependent variable) to be relied by the prediction model in generating high accurate results.

Acknowledging the advantages of artificial intelligence computing approach that able to learn and redevelop knowledge to self-improve the output target from the given data, the used of machine learning technique in solving issues of real estate industry has started to begin. Even with low-association dataset, machine learning with the intelligent and leaning ability, will use mathematical and heuristics projection to self-improve their performances continuously during the training stage. Despite the wider used of machine learning in various domains of problems, there is still limited work that can be found for the real estate industry. This research attempted to fill the gap by focusing on the flexible and rapid modelling machine learning approach for office rental prediction problem.

This research was funded by NAPREC under grant number 100-TNCPI/GOV 16/6/2 (027/2021).

The contributions of this paper are two-fold. First, it demonstrates the use of AutoModel RapidMiner for the office rental price prediction model as a rapid modelling in the preliminary research activities. Second, it presents a comparison of results between two machine learning training approaches (split, cross-validation) that developed with manual model in RapidMiner and provides discussion on the significant findings in the context of office rental domain.

This paper is organized as follow. The next section II presents the literature review of rapid modelling machine learning and the state-of-the-arts of research for office rentals prediction. The methodology of research is given in section II followed with the results and discussion in section IV. The last section V provides the conclusion remarks.

II. Literature Review

A. Rapid Modelling of Machine Learning

Rapid modelling has been a long-time concern by researchers to solve real-world problems, mainly when complex computing techniques are required. In machine learning, to support rapid modelling, several tools have been introduced, and the most popular is script programming with Python or R programming languages [7]. Although this two-scripting language is considerably easy to utilize with their built-in programming libraries, they still need some help for a non-computing expert who never learns to program.

Graphical User Interface (GUI) based rapid modelling is easier than programming to implement machine learning. To date, some of the popular rapid tools are Weka [8] and RapidMiner [9]. Automated machine learning (AML) [10], [11] is a promising module in Weka and RapidMiner. Auto-WEKA is the AML in Weka, while AutoModel is the one provided in RapidMiner. AML helps to accelerate the modelling tasks by optimizing the machine learning pipelines from the input variables (features selections) step, algorithm suitability selection and hyper-parameters optimal setting. Literature on research that used Auto-WEKA is more available than AutoModel RapidMiner. Thus, focusing on Auto-Model RapidMiner for this study can provide additional benefits to the AML scholarly, mainly to inexpert data scientists in various domains.

B. Office Rent Predictions with Machine Learning

A number of researchers have used machine learning in the real estate or property industry. Researchers in [12] presents the real estate opportunities with machine learning technique modelling. It was reported that one of the advantages of machine learning is for assisting the stakeholders in making important decisions related to commercial or office building. A significant of finding has been presented by researchers in [13] that used machine learning technique to estimate the warehouse rental price. More interesting, by utilising social media web scraping technique, the collected data were analysed with machine learning prediction algorithms and hedonic modelling to monitor the building rental prices in Shenzhen, China [14]. To ensure the development of holistic smart building control effectively, researchers in [15] have utilised deep reinforcement learning, a recent advance method of machine learning. Internet of Things (IoT) is the main elements of smart building hence researchers in [16] introduced an algorithm named as Random Neural Network (RNN) to make used the IoT data to predict the consumption of energy of the smart building. Random Forest machine learning for forecasting shop rents in Guangzhou, China. To conduct mass appraisal in an urban residential area where commercial properties are available, researchers in [17] used multiple regression and random forest as the proposed methods. Similarly, the performances of random forest and multiple regression has been reported from the research findings in [18]. How the neighbourhood environment can influence peer-to-peer accommodation when using random forest is the finding reported in [19]. It seems that random forest is very promising, and it is also one of the suggested algorithms from AutoModel used in this research together with Decision Tree [20] and Support Vector Machine [21]. To the best of our knowledge, rapid modelling on office rental prediction has not been reported yet in the current literature. This research filled the gap by presenting the precise steps and the comparison of results.

III. Methodology

A. The Dataset

The dataset used in this study is a collection of office rentals from the year 2015-2021 in Kuala Lumpur, Malaysia. Table I shows the set of features of developing the machine learning prediction model. This study uses 21 attributes or features as independent variables for office rental prediction.

TABLE I. Features of the Office Rental Prediction Model

Office Rent Determinants	Description
Building Appearance and Design	Physical building appearances and Design
Building Age	Age of Building
Amenities and In-house Services	Amenities and Services Provided by the building
Occupancy	Occupancy rate
Distance to the city centre	Distance to the city centre
Building Frontage	Building allocation towards the main road
Neighbourhood Characteristics	Surrounding areas
Traffic Condition	Traffic Condition (Congested, Free)
Nearest Public Transport	Availability of Public Transport
Transaction Date	Date of Transactions
Floor Level	Rented Floor Level
Rentable Area	Area of premises being rented
Tenancy Duration	Lease duration by the tenant
Service Charge	Service Charge towards the building's occupants
Employment Rate	Employment rate by year
Inflation	Inflation by year
Gross Domestic Product (GDP)	Gross Domestic Product by Year
Finance, Insurance and Real Estate (FIRE)	Business purposes of building
Green Certificate	Green Certification of the building
MSc Status	MSc Certification of the building
Building Grade	Office Building Grading

B. Machine Learning Rapid Modelling Framework

Fig. 1 presents the rapid modeling framework used in this research. On the pre-processing data that was ready to be read in RapidMiner, AutoModel was firstly executed. The purpose of executing AutoModel is to get suitable machine learning algorithms for the dataset and the optimal hyper-parameters. The suggested machine learning algorithms are Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees and Support Vector Machine. The six sets of experiments can be described as listed in Table II. Table III is the optimal hyper-parameters setting suggested by the AutoModel.

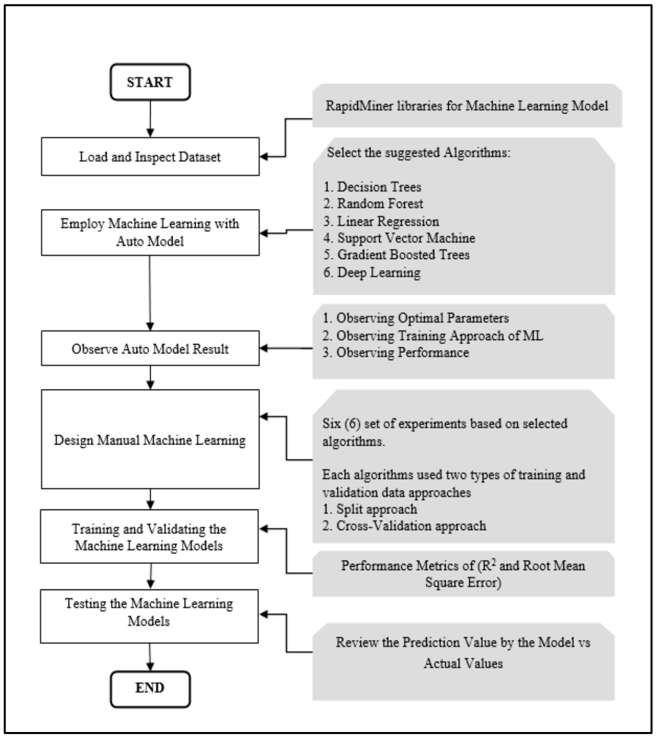

Fig. 1. Machine learning rapid modeling framework.

TABLE II. DESCRIPTION OF EXPERIMENT SET

Experiment set	Algorithm	Training
1	Decision Tree	Split
2	Random Forest	Split
3	Support Vector Machine	Split
4	Decision Tree	CV
5	Random Forest	CV
8	Support Vector Machine	CV

TABLE III. ALGORITHMS SELECTION AND OPTIMAL HYPER-PARAMETERS

List of Algorithms	Maximal Depth	Number of Trees	RBF	C	Error Rate
Decision Tree	15	NA	NA	NA	10.7%
Random Forest	20	7	NA	NA	26%
Support Vector Machine	NA	NA	0.050	1000	43.5%

As listed in Table III, the lowest error rate for the Decision Tree can be achieved for the office rental dataset if the maximal depth of the tree is 15. maximal depth defines the maximum level of the tree minus 1. AutoModel has identified that the worst error rate was achieved if the maximal depth was set to 2. Random Forest has additional hyper-parameters as it is an advancement of Decision Tree. Besides maximal depth, Random Forest has number of trees. Imagine a forest that consists of more than one tree. AutoModel suggested that the optimal values for the number of trees and maximal depth were 20 and 7, respectively. The worst error rate achieved was 42.9% at 140 trees and two maximal depths. Support Vector Machine has Radial Basis Function (RFB) Kernal and C value. The C parameter instructs the SVM optimization to avoid misclassifying each training example. The kernel function is used to transform the data, increasing its dimensional. This enhancement causes the data to be split with a hyperplane with a significantly greater probability and establishes a minimal prediction probability error measure [22]. The configuration of optimal parameters suggested by Auto Model for the Support Vector Machine algorithm is at RBF = 0.050, C = 100, and 43.5% error rate.

C. Training Approaches

AutoModel used a split training approach that allowed the researcher to look at another training approach, namely the cross-validation approach. Therefore, based on the AutoModel findings, manual machine learning modelling in RapidMiner was developed to compare the three machine learning algorithms with the optimal setting but with different training approaches; split and cross-validation (CV). Fig. 2 presents the RapidMiner process for split training, while Fig. 3 presents the CV approach.

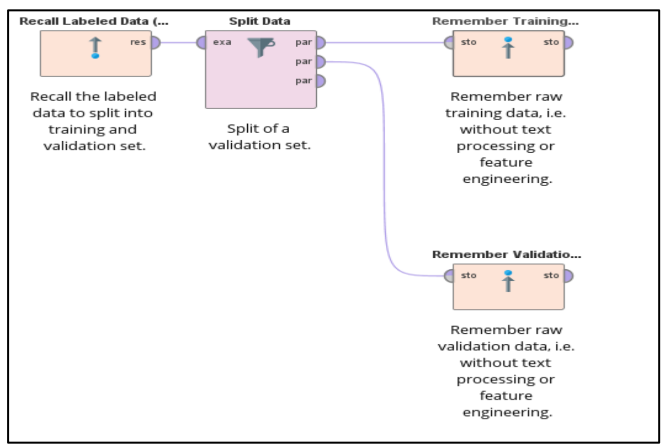

Fig. 2. Split approach.

The "Split Data" operator is a custom operator for splitting a dataset into training and testing datasets [23]. To configure the parameter in the split approach, the researcher will specify the ratio of all partitions. The sum of all partitions' ratios should be equal to 1. As for this study, the 0.8:02 ratio was used for setting the partitions. Therefore, the training and testing datasets constructed from the original dataset were 80% and 20% of the data, respectively. The next step is to include the algorithms in the office rental price prediction model by connecting all nodes from the chosen parameters.

Fig. 4, Fig. 5 and Fig. 6 present the manual modelling process for the tree algorithms with split training. The Split Data operator can be changed to Cross Validation operator for implementing CV on each model.

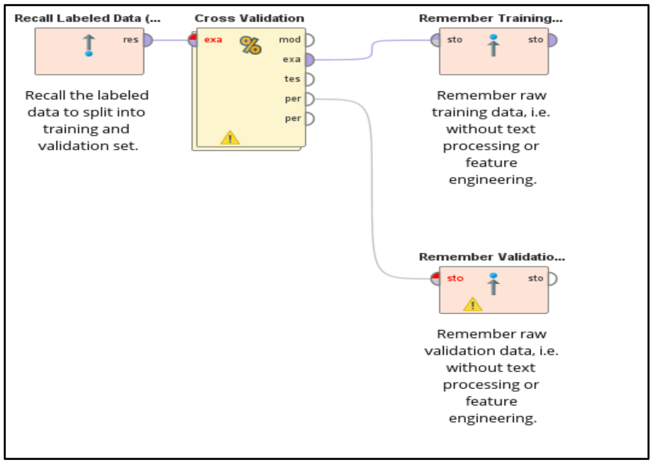

Fig. 3. Cross-validation approach.

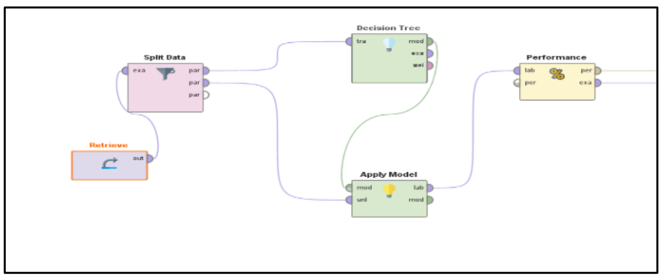

Fig. 4. Decision tree with split training.

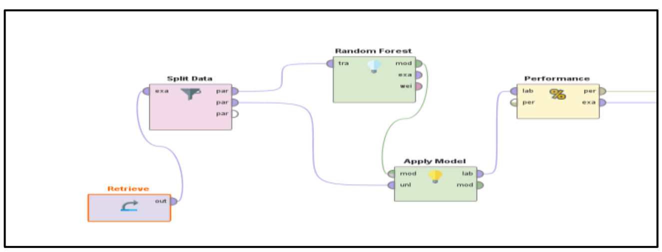

Fig. 5. Random forest with split training.

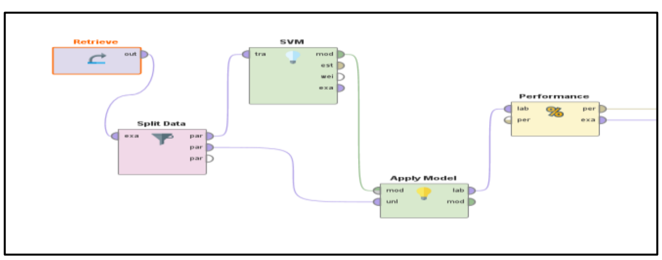

Fig. 6. Support vector machine with split training.

D. Performances Metric

In developing the office rent prediction model, this study deployed two (2) performance measurements, namely the Coefficient of Determination "R^2" and "Root Mean Square Error". R-Squared is a statistical metric that indicates the proportion of the variation explained by the independent variables for the dependent variable. The greater the R-squared, the better the model matches the dataset under consideration. R-Squared may be calculated mathematically as Equation (1).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - y)^2} \tag{1}$$

The metric of root means square error (RMSE) is a common way to calculate a model's error in quantitative data prediction. *E* is no absolute good or bad error level but can be identified based on the dependent variable. Generally, the range from 0 to 1000 is classified as small, but if the range is from 0 to 1, it is classified as no longer being small. In evaluating the model, the smaller the value of the root mean square error the better the model has produced [24]. Equation (2) is to calculate root means square error.

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

IV. RESULTS AND DISCUSSION

A. Machine Learning Prediction Results based on Split Approach

Table IV compares the prediction results generated by the three different algorithms using the split approach. The "observed value" indicates the actual value/the raw rental data acquired. The prediction value indicates the value generated after considering the optimal parameters, processes, and factors involved.

TABLE IV. SPLIT-DATA PREDICTION RESULTS

Observed Value	**Prediction**		
	Decision Tree	**Random Forest**	**Support Vector Machine**
RM24.53	RM29.20	RM42.67	RM37.50
RM27.55	RM31.89	RM34.60	RM37.57
RM24.58	RM35.47	RM27.37	RM41.53
RM19.85	RM22.42	RM43.09	RM43.80
RM21.37	RM35.47	RM29.06	RM41.58
RM21.98	RM32.45	RM26.52	RM39.48
RM76.84	RM77.47	RM73.72	RM63.35
RM28.00	RM28.72	RM37.16	RM34.77
RM29.85	RM23.97	RM35.37	RM41.99
RM21.37	RM36.28	RM28.44	RM38.78

An extensive result of the prediction with details is provided in the prediction chart, as shown in Fig. 4, 5 and 6. The prediction chart depicted in Fig. 4 to 6 was analysed from the triangle patterns and the diagonal line. The diagonal line indicates the real data/actual office rent values while the triangle patterns indicates the prediction generated by the algorithms used in this study. The accumulated pattern at the diagonal line shows a good prediction. In contrast, the deviated pattern from the line showed less accuracy in predicting the rent and was considered a prediction outlier. The illustrated chart demonstrated that Random Forest (Refer Fig. 8) provides the best predictions can be seen through the high accumulation of the triangle patterns to the diagonal line. In comparison, Decision Tree algorithms also provide a good prediction, as illustrated in Fig. 7. The prediction chart for Support Vector Machine, however, depicts scattered triangle patterns which indicate high frequencies of outliers when predicting office rents, as seen in Fig. 9. To justify Random Forest algorithms as the best predictor, this study employed another approach in Machine Learning called Cross-Validation.

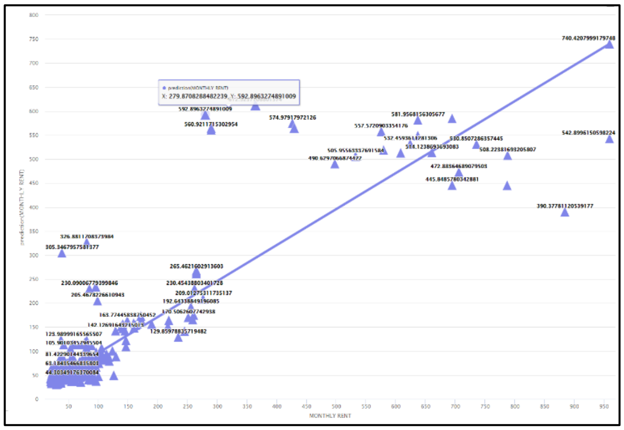

Fig. 7. Prediction chart of decision tree (split).

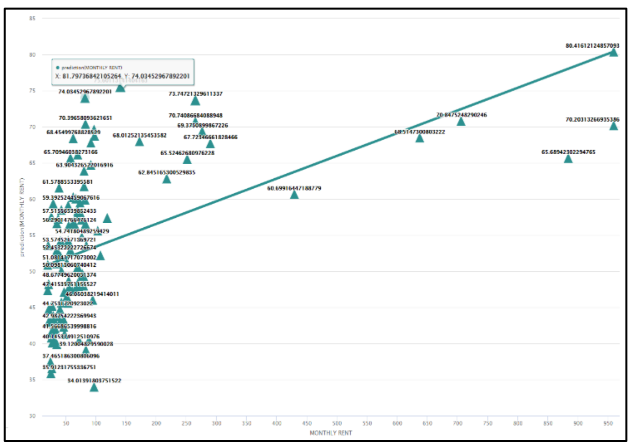

Fig. 8. Prediction chart of random forest (split).

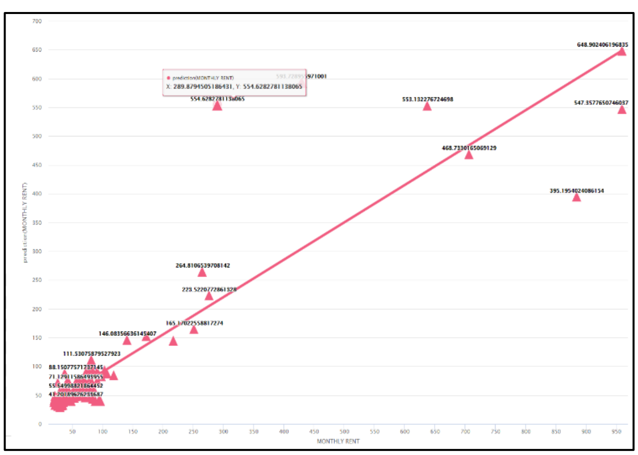

Fig. 9. Prediction chart of support vector machine (split).

B. *Machine Learning Prediction Results based on CV Approach*

Table V lists some of the prediction results generated by each algorithm using the CV approach. An extensive result of the prediction with details is provided in the prediction charts, as shown in Fig 10 to 12.

TABLE V. CROSS-VALIDATION PREDICTION RESULTS

Observed Value	**Prediction**		
	Decision Tree	**Random Forest**	**Support Vector Machine**
RM24.53	RM29.20	RM46.34	RM22.42
RM27.55	RM31.89	RM34.89	RM39.57
RM24.58	RM35.47	RM31.37	RM41.53
RM19.85	RM22.42	RM43.09	RM35.47
RM21.37	RM35.47	RM29.06	RM41.58
RM21.98	RM32.45	RM26.52	RM39.48
RM76.84	RM77.47	RM73.72	RM63.35
RM28.00	RM28.72	RM37.16	RM34.77
RM29.85	RM23.97	RM35.37	RM28.72
RM21.37	RM36.28	RM23.97	RM38.78

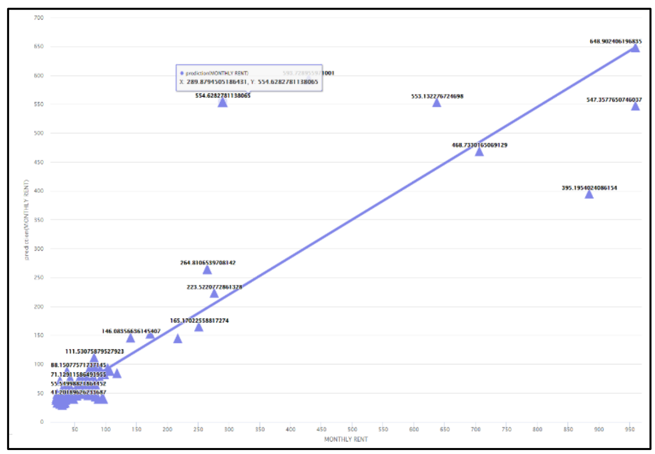

Fig. 10. Prediction chart of decision tree (CV).

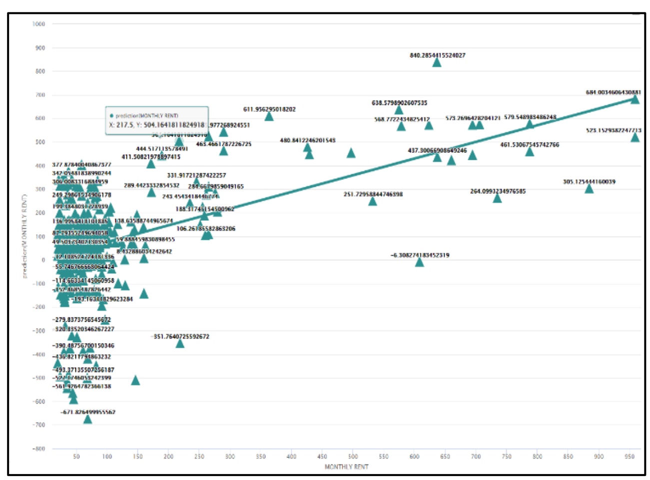

Fig. 11. Prediction chart of random forest (CV).

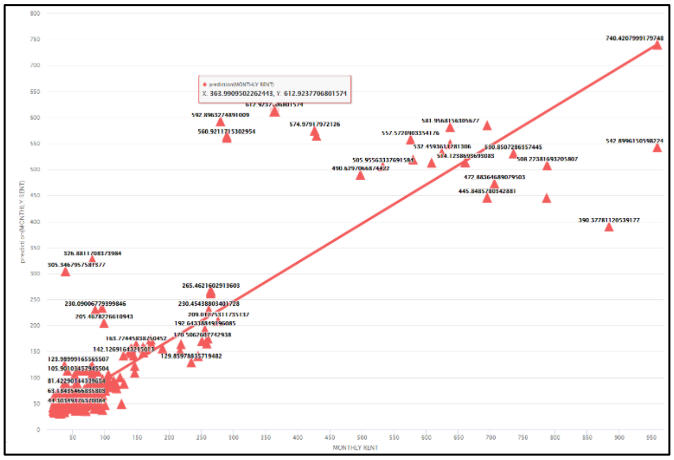

Fig. 12. Prediction chart of support vector machine (CV).

The prediction charts of each algorithm with the CV approach demonstrate that Random Forest provides the best predictions can be seen through the high accumulation of the triangle patterns to the diagonal line. In comparison, Decision Tree algorithms also provide a good prediction, as illustrated in Fig. 4. However, the prediction chart for the Support Vector Machine depicts scattered triangle patterns, which indicate high frequencies of outliers when predicting office rents.

C. Performances Comparisons

Table VI compares the performances of each algorithm concerning the training approaches. Overall, the result shows that the CV approach provides better results with a higher R^2 correlation value and lower RMSE value than the split data approach. The Random Forest algorithm outperformed other algorithms regarding office rental price prediction value. It has generated the most accurate prediction compared to the decision trees and support vector machine. The accuracy of prediction generated by random forest was supported through the observation of the RMSE value. The lowest RMSE values generated by random forest justified a lower error rate when predicting a model evaluation; the smaller the value of root means square error, the better the model produced.

TABLE VI. MACHINE LEARNING ALGORITHMS COMPARISONS

Algorithms	**Split**		**CV**	
	R^2	**RMSE**	**R^2**	**RMSE**
Decision Tree	0.852	80.663	0.876	61.294
Random Forest	0.883	75.695	0.906	51.859
Support Vector Machine	0.479	155.465	0.725	120.738

V. CONCLUSIONS

This paper presents the review and findings of using machine learning algorithms for real data of office building rent in Bandar Kuala Lumpur, Malaysia. The rigorous steps for rapid machine learning models initiated with AutoModel preliminary findings were given in this paper. The suggested setting from AutoModel has been used to improve the machine learning by using the suggested split training with cross-validation technique. The performances of all machine learning algorithms can be improved with the cross-validation technique. However, the findings from this study are limited to the tested datasets and therefore require further investigation for different types of problems. Notably, this study provides new knowledge on the application of machine learning in analyzing real estate data, particularly on rental values of the office building. By exploring the machine learning methods, this study will greatly assist future research in solving problems involving prediction and forecasting real estate data.

REFERENCES

[1] J. Barnett, W. Serrano, P. Treleaven, and A. Knight, “Real Estate Data Marketplace,” SSRN Electronic Journal, no. January, 2021, doi: 10.2139/ssrn.3745816.

[2] F. Lorenz, J. Willwersch, M. Cajias, and F. Fuerst, “Interpretable machine learning for real estate market analysis,” Real Estate Economics, 2022, doi: 10.1111/1540-6229.12397.

[3] D. Jaffee, R. Stanton, and N. Wallace, “Energy Factors, Leasing Structure and the Market Price of Office Buildings in the U.S.,” Journal of Real Estate Finance and Economics, vol. 59, no. 3, pp. 329–371, Oct. 2019, doi: 10.1007/s11146-018-9676-x.

[4] H. H. Rong, J. Yang, M. Kang, and A. Chegut, “The value of design in real estate asset pricing,” Buildings, vol. 10, no. 10, pp. 1–26, Oct. 2020, doi: 10.3390/buildings10100178.

[5] M. Ezra, H. Mohd Ali, and Tuti Haryati Jasimin, “Valuers Behavioural Uncertainties in Property Valuation Decision Making,” 2018.

[6] D. Trojanowski, Recent trends and its analysis, no. January. 2019.

[7] Ramachandran Trichur Narayanan, “Novice Programmer to New-Age Application Developer: What Makes Python their First Choice?” 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) , 2019.

[8] Jason Brownlee, Machine learning mastery with Weka, 1st ed., vol. 1(4). 2019.

[9] J. Arunadevi, S. Ramya, and M. R. Raja, “A study of classification algorithms using Rapidminer,” 2018. [Online]. Available: https://www.researchgate.net/publication/325718529.

[10] F. Hutter, L. Kotthoff, and J. Vanschoren, “The Springer Series on Challenges in Machine Learning Automated Machine Learning Methods, Systems, Challenges,” 2019. [Online]. Available: http://www.springer.com/series/15602.

[11] L. Vaccaro, G. Sansonetti, and A. Micarelli, “An Empirical Review of Automated Machine Learning,” Computers, vol. 10, p. 11, 2021, doi: 10.3390/computers.

[12] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, and C. Afonso, “Identifying real estate opportunities using machine learning,” Applied Sciences (Switzerland), vol. 8, no. 11, Nov. 2018, doi: 10.3390/app8112321.

[13] Y. Ma et al., “Estimating Warehouse Rental Price using Machine Learning Techniques,” International Journal of Computers Communications & Control, vol. 13, no. 2, pp. 235–250, 2018.

[14] L. Hu, S. He, Z. Han, S. Su, M. Weng, and Z. Cai, “Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies,” 2018.

[15] X. Ding, W. Du, and A. Cerpa, “OCTOPUS: Deep reinforcement learning for holistic smart building control,” in BuildSys 2019 - Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, Nov. 2019, pp. 326–335. doi: 10.1145/3360322.3360857.

[16] A. Javed, H. Larijani, and A. Wixted, “Improving Energy Consumption of a Commercial Building with IoT and Machine Learning,” IT Prof, vol. 20, no. 5, pp. 30–38, Sep. 2018, doi: 10.1109/MITP.2018.053891335.

[17] S. Yilmazer and S. Kocaman, “A mass appraisal assessment study using machine learning based on multiple regression and random forest,” Land use policy, vol. 99, Dec. 2020, doi: 10.1016/j.landusepol.2020.104889.

[18] M. Čeh, M. Kilibarda, A. Lisec, and B. Bajat, "Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments," ISPRS Int J Geoinf, vol. 7, no. 5, May 2018, doi: 10.3390/ijgi7050168.

[19] H. Jiang, L. Mei, Y. Wei, R. Zheng, and Y. Guo, "The influence of the neighbourhood environment on peer-to-peer accommodations: A random forest regression analysis," Journal of Hospitality and Tourism Management, vol. 51, pp. 105–118, Jun. 2022, doi: 10.1016/j.jhtm.2022.02.028.

[20] J. Kelleher, B. mac Namee, and A. D. ' Arcy, "Fundamentals of Machine Learning for Predictive Data Analytics."

[21] D. A. Pisner and D. M. Schnyer, "Support vector machine," Machine Learning: Methods and Applications to Brain Disorders, pp. 101–121, Jan. 2020, doi: 10.1016/B978-0-12-815739-8.00006-7.

[22] M. Graczyk et al., "Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA," LNAI, vol. 5796, pp. 800–812, 2009, doi: 10.1007/978-3-642-04441-0_70.

[23] A. Massaro, V. Maritati, and A. Galiano, "Data Mining Model Performance of Sales Predictive Algorithms Based on Rapidminer Workflows," International Journal of Computer Science and Information Technology (IJCSIT) International Journal of Computer Science & Information Technology (IJCSIT) International Journal of Computer Science & Information Technology (IJCSIT), vol. 10, no. 3, 2018, doi: 10.5121/ijcsit.2018.10303.

[24] D. Ho, G. Newell, and A. Walker, "The importance of property-specific attributes in assessing CBD office building quality," Journal of Property Investment and Finance, vol. 23, no. 5, pp. 424–444, 2005, doi: 10.1108/14635780510616025.