American Journal of Management Science and Engineering
2023; 8(2): 50-55
http://www.sciencepublishinggroup.com/j/ajmse
doi: 10.11648/j.ajmse.20230802.12
ISSN: 2575-193X (Print); ISSN: 2575-1379 (Online)

Machine Learning-Based House Rent Prediction Using Stacking Integration Method

Kainuo Wang[1, *], Huiyi Zhao[2], Jingzhong Li[1]

[1]School of Resource and Environmental Science, Wuhan University, Wuhan, China

[2]School of Urban Design, Wuhan University, Wuhan, China

Email address:

2020302051136@whu.edu.cn (Kainuo Wang), 2021302092032@whu.edu.cn (Huiyi Zhao), 00009232@whu.edu.cn (Jingzhong Li)

[*]Corresponding author

To cite this article:

Kainuo Wang, Huiyi Zhao, Jingzhong Li. Machine Learning-Based House Rent Prediction Using Stacking Integration Method. *American Journal of Management Science and Engineering*. Vol. 8, No. 2, 2023, pp. 50-55. doi: 10.11648/j.ajmse.20230802.12

Received: March 7, 2023; **Accepted**: April 10, 2023; **Published**: April 18, 2023

Abstract: With the advancement of urbanization and the gradual increase of the rental population, the housing rental market is growing rapidly. It is important to achieve accurate housing rent prediction in order to stabilize the rental housing market. The influence of spatial and temporal factors has led to the complexity of house rent prediction, so it has always been difficult to find an appropriate method. In recent years, machine learning models have been widely studied and applied in various fields, which may provide a promising solution to it. In this paper, a stacking-based ensemble learning model is proposed to solve the problem of house rent prediction. First, the raw data are preprocessed, including decomposing hybrid features, removing rent outliers using scatterplot, removing uncorrelated features, and applying one-hot encoding to transform categorical features into numerical features. Second, the pre-processed data is normalized to unify the magnitudes. Then, the competent base predictive models are selected from all the trained base predictive models and integrated into a comprehensive ensemble model using the stacking integration method to make the final prediction. Finally, the various models are evaluated by some metrics. The experimental results show that the proposed stacking integration-based machine learning method outperforms the individual machine learning methods in solving the house rent prediction problem.

Keywords: Stacking Integration, Ensemble Model, Machine Learning, House Rent, Prediction

1. Introduction

Urbanization is a major trend in the economic development of various countries today. As urbanization progresses, the transition from rural to urban populations has led to a dramatic increase in demand and prices for city-based housing [1]. The transient or population with low purchasing power is in urgent need for housing solutions, and renting house is a suitable way to solve the problem. Today, the rental market continues to grow and receives more and more attention [7]. In addition, fluctuations in urban house rent are increasingly linked to the happiness of local residents. Therefore, it is important to accurately predict house rent with scientific statistical methods, in particular, machine learning methods.

The machine learning model is a computing system, which, after being trained, can recognize specific types of patterns, e.g., make inferences to future data based on historical data [10]. In this study, Firstly, the raw data is preprocessed, including decomposition of hybrid features, removal of rent outliers, the one-hot encoding for feature numeration, and data normalization. Then, the raw data is divided into two parts, i.e., the test set and the training set. Finally, an ensemble learning model is proposed for accurate prediction of house rent through stacking integration. The proposed model deploys an ensemble algorithm on three base predictive models, including the Gradient Boosting Regression (GBR), Light Gradient Boosting Machine (LGBM) and Neural Networks (NN), and uses linear regression as the meta-regressor to enhance model performance. The dataset of house rent in India is used to validate the model, and evaluate the model performance according to some metrics.

The remainder of this study is organized as follows. Section 2 shows related work of data mining techniques and house rent prediction. Section 3 explores data preprocessing process and the modeling method. Experimental results are analyzed in

Section 4. In Section 5, conclusion of this study and future work are presented.

2. Related Work

In recent years, machine learning techniques have been widely applied in multiple fields to help people extract information from historical data and predict the future more accurately. Lim et al. compared the performance of the Autoregressive Integrated Moving Average (ARIMA) model with Artificial Neural Network (ANN) model using mean square error (MSE), and confirmed the ANN's superiority over the traditional statistical model such as ARIMA in price prediction of the Singapore condominiums, which was important for potential buyers to make an informed decision [9]. Garzon et al. proposed a machine learning model that can predict the free energy of compound binding to SARS-CoV-2 major protease, in which, the optimized multiple linear regression models were trained on 226 natural compounds and achieved the reliable prediction performance [5]. Sharma and Prabha investigated the machine learning effect in cancer prediction and prognosis, and found that machine learning methods can substantially improve the predictive accuracy of cancer susceptibility, recurrence and mortality by 15-25% [11]. Ahmad et al. proposed an ANN-based model for predicting electricity consumption in New Zealand residences, and the result demonstrated that the model can predict the energy consumption of an Oakland home 24 hours in advance, which is more accurate than the benchmark methods [2].

With the emergence of the real estate rental market, machine learning techniques have been gradually applied to the field of house rent prediction. Zhang et al. combined eXtreme Gradient Boosting (XGBoost), LGBM and CatBoost together into a joint model for predicting house rent using linear weighted least squares method [16]. Fikire discovered that homeowners and renters prefer properties with good structure, location and community features using machine learning [4]. It was also found that renters were more concerned about accessibility to employment centers and public transportation. Housing structure characteristics such as building age, land size, access to water, floor and kitchen size were factors that influence rent. Yoshida et al. used regression-based and machine learning-based methods to spatially predict rent with a large dataset, by adding new empirical evidence and considering the spatial dependence of observations [15].

Previous literature has demonstrated that machine learning methods have good applications in medicine, electricity consumption, house rent and other fields. However, in housing rent prediction, most studies tended to use a single machine learning model or just considered the simple combination of various base models, which affects the robustness and stability of models. In this study, three machine learning models are selected as competent base predictors for predicting house rent. To further improve the performance of the predictive models, the stacking ensemble method is used to integrate the three competent base predictors.

3. Methodology

3.1. Data Preprocessing

The data used in the study is obtained from the Kaggle[1] (a Melbourne-based company founded by Co-founder and CEO Anthony Goldbloom in 2010), including house rent information in India. As shown in Table 1, the dataset consists of 4747 instances with 10 main features and 1 prediction label, where the features are related to specific information about the rental houses and the prediction label is house rent. The original data contains both textual and numerical data. There are also some data outliers and features that are not relevant to the experiment. Therefore, it is important to pre-process the dataset before experiment.

Table 1. *The description of main features.*

Features	Description
Posted on	Release date
BHK	The number of bedrooms, hall or kitchen.
Bathroom	The number of bathrooms.
Size	The area of the houses/apartments/flats in square feet.
Floor	Houses/apartments/flats on which floor and total number of floors
Area type	Type of the houses/apartments/flats (super area, carpet area or floor area etc.)
City	City where the houses/apartments/flats are located.
Furnishing status	The furnishing status of the houses/ apartments/flats
Tenant preferred	Type of tenant preferred by the owner or agent.
Point of contact	Whom should you contact for more information regarding the houses/apartments/flats.
Rent	Rent of the houses/apartments/flats.

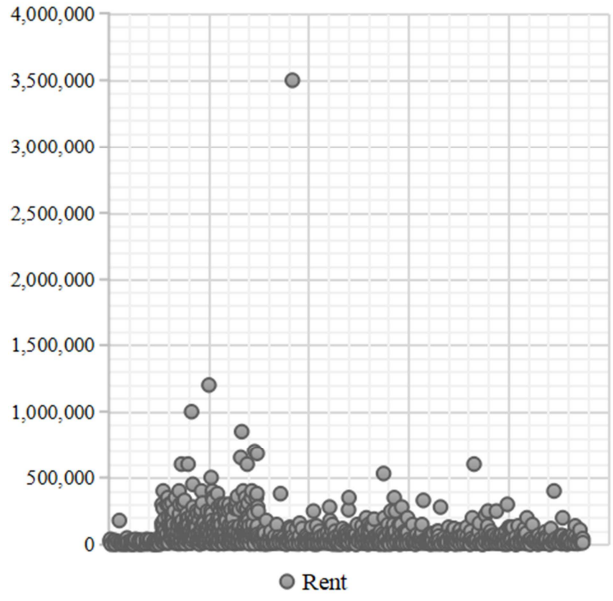

Figure 1. *Result of the scatterplot.*

The steps for data preprocessing are as follows:

1 https://www.kaggle.com/

1) Decomposition of hybrid features

In the raw data, the feature “Floor” is a hybrid feature that represents which floor the houses/apartments/flats is on and also the total number of floors (e.g., Ground out of 3, 2 out of 4, etc.). It should be decomposed into two machine-understandable simple features, i.e., “Current floor” and “Total floors”. For example, after the hybrid feature “floor” with value “2 out of 4” is decomposed, 2 is stored in the new feature “Current floor” and 4 is stored in the new feature “Total floors”.

2) Removing the rent outliers

Outliers that do not fit the overall data trend usually exist within the dataset for various reasons. In this study, the outliers of the rent were found and removed through the scatterplot. The result of the scatterplot is shown in Figure 1, which can be used to remove the maximum value from the graph [8, 13].

3) Transforming the data

Considering that the dataset is based on rental information within four months and that housing rents do not fluctuate significantly in the short term, the feature "posed on" is removed. Then, the categorical features such as “Area type”, "City", "Furnishing status", "Tenant preferred" and "Point of contact" are mapped into Euclidean space and transformed into numeric features using one-hot encoding, which facilitates the calculation of distance or similarity between features in regression-based machine learning algorithms.

4) Normalizing the data

Since the magnitudes of different features are different, the data normalization approach is adopted, so as to reduce the impact of different order of magnitudes and improve the performance of the model.

3.2. The Proposed Ensemble Model

The flowchart of the proposed ensemble model is shown in Figure 2, and the details are as follows.

1) Dividing the dataset

To evaluate the accuracy of the model, the dataset is divided into two parts, i.e., 80% as the training set and 20% as the test set. The training set is used to train and optimize the predictive model, while the test set is used to test the accuracy, robustness and efficiency of the model [12].

2) Training the base predictive models

Some promising base predictive models are trained and optimized with training set, including GBR, Support Vector Regression (SVR), NN, K-Nearest Neighbor (KNN), and LGBM.

3) Stacking integration

The Stacking integration method first obtains some promising base predictive models from different algorithms through training [14], and then integrates the output results of competent base predictive models by training a meta-model through ensemble learning. In the experiment, three base predictive models with better performance, i.e., GBR, LGBM and NN, are selected as competent base regressors in the ensemble model. Linear Regression (LR) is used as a meta-regressor to combine the results of the competent base regressors, and its output is obtained as the final prediction result of the ensemble model.

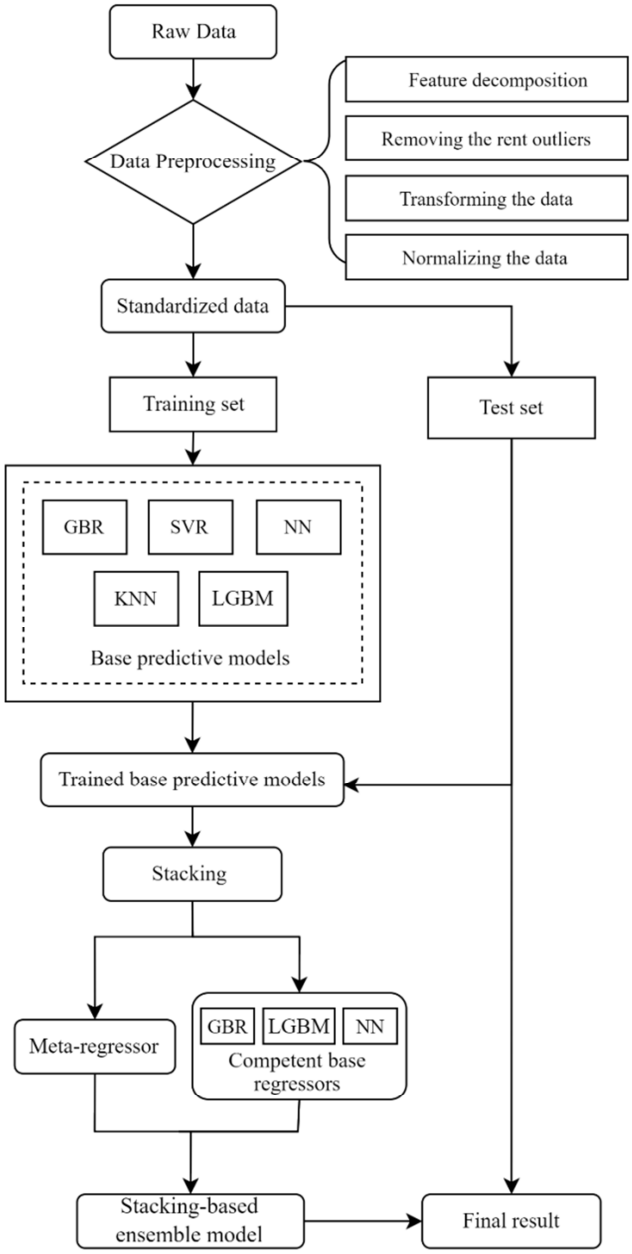

Figure 2. *Flow chart of the proposed ensemble model.*

4. Experimental Results

This section evaluates the effectiveness of the proposed stacking integration-based machine learning method and its advantage over individual machine learning methods. Firstly, the evaluation metrics involved in the experiment are presented. Then the experimental results of all predictive models in terms of house rent prediction are compared and analyzed. The Python programming language is used to implement all predictive models and methods.

4.1. Evaluation Metrics

In order to evaluate the predictive performance of the

predictive models, four widely used credible statistical metrics are adopted, consisting of root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and coefficient of determination (R^2). MAE and RMSE are used to measure the absolute magnitude of the deviation of the true values from the predicted values, MAPE measures the relative magnitude of the deviation (i.e., percentage), and R^2 measures the fitness of the linear regression [3]. The definition and calculation process of these four indicators are as shown in Eqs. (1) - (4).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2} \quad (1)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{x}_i - x_i| \quad (2)$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{x}_i - x_i}{x_i}\right| \quad (3)$$

$$R^2 = \left[1 - \frac{\sum_{i=1}^{N}(\hat{x}_i - x_i)^2}{\sum_{i=1}^{N}(\bar{x}_i - x_i)^2}\right] \quad (4)$$

where $x_i(1 \le i \le N)$ and $\hat{x}_i(1 \le i \le N)$ denote the practical and predicted values of the house rent respectively, and N denotes the size of the test samples. $\bar{x}_i$ denotes the average house rent of *N* test samples. It is worth noting that for RMSE, MAE and MAPE, the lower value indicates the better predictive performance [6]. For R^2, the closer the number is to 1, the better the predictive performance.

4.2. Experimental Results Analysis

Table 2 shows the evaluation results obtained by the five promising base predictive models. It is found that GBR, LGBM and NN are the three well-performing models, which achieve the highest R^2 at 0.8608, 0.8486 and 0.8558 respectively, and they achieve the lowest RSME at 0.3859, 0.4024 and 0.3927 respectively. In general, GBR achieves the best overall performance. On the contrary, the SVR and KNN have the relatively fair performance, i.e., achieve 0.7047 and 0.7423 in R^2 respectively, 0.1664 and 0.1808 in MAE respectively, and 0.5620 and 0.5251 in RSME respectively. Therefore, GBR, LGBM and NN are more suitable for house rent prediction and selected as competent base regressors to form the ensemble model through stacking integration. Moreover, to present the evaluation results more intuitively, the predicting results of all models are depicted in column charts, as shown in Figure 3.

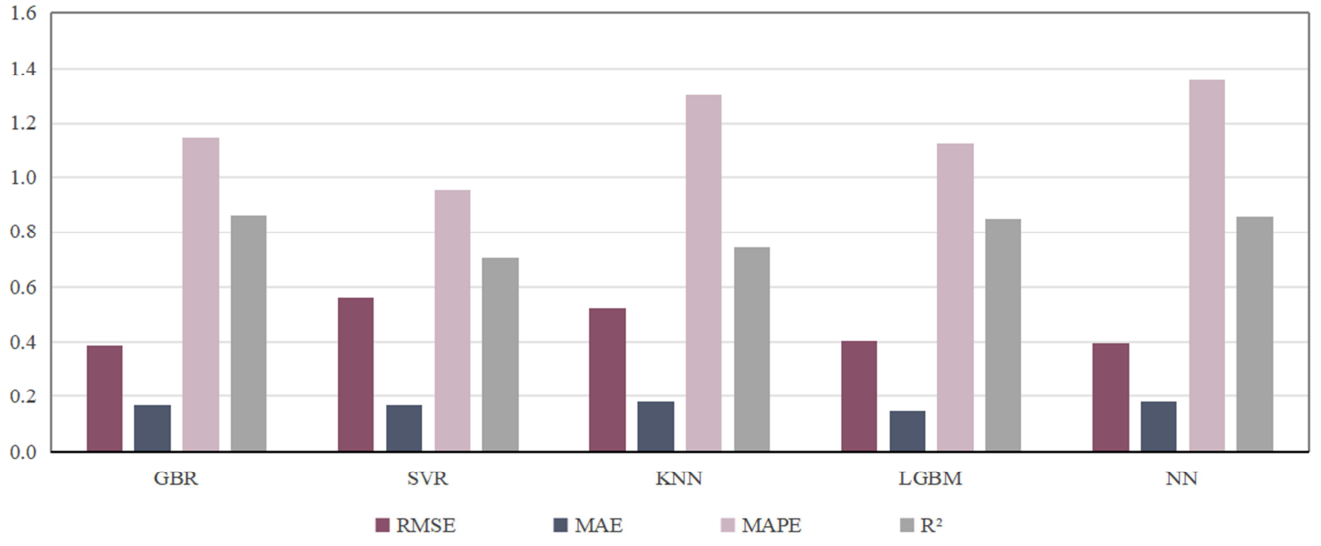

Figure 3. *Evaluation results of five base predictive models.*

Table 2. *Evaluation results of five base predictive models.*

Model	RMSE	MAE	MAPE	R²
GBR	0.3859	0.1659	1.1470	0.8608
SVR	0.5620	0.1664	0.9527	0.7047
KNN	0.5251	0.1808	1.3020	0.7423
LGBM	0.4024	0.1458	1.1261	0.8486
NN	0.3927	0.1783	1.3575	0.8558

In this study, the best-performing GBR among the five base predictive models was selected for comparison with the proposed stacking-based ensemble model to verify the predictive accuracy of the proposed model.

Table 3 shows the evaluation results obtained by the proposed ensemble model (Stacking) and GBR respectively. It shows that the proposed ensemble model outperforms GBR in terms of all metrics. This means that the proposed stacking-based ensemble model has a better fitness for the house rent prediction. Moreover, to present the evaluation results more intuitively, the predicting results of GBR and Stacking model are depicted in column charts, as shown in Figure 4.

Table 3. *Evaluation results of GBR and Stacking model.*

Model	RMSE	MAE	MAPE	R²
GBR	0.3859	0.1659	1.1470	0.8608
Stacking	0.3712	0.1485	1.0952	0.8712

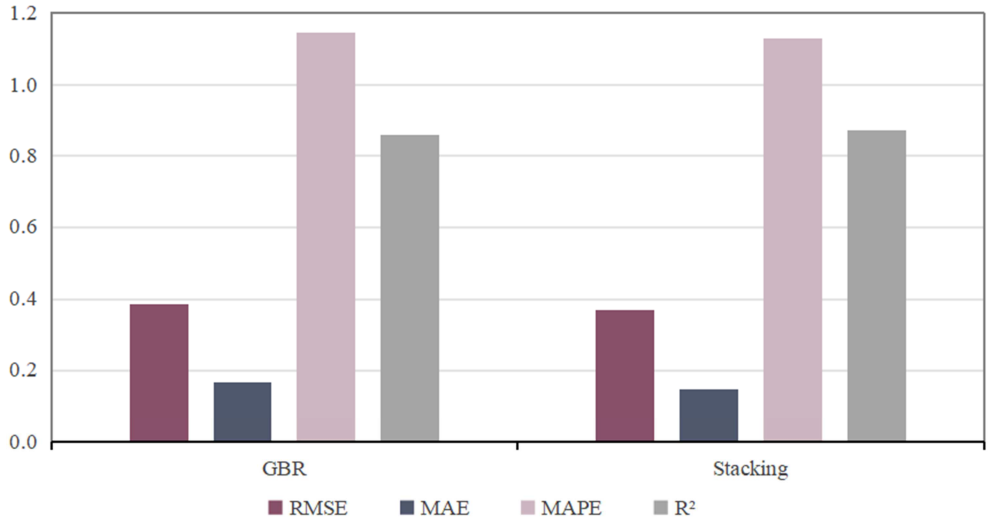

Figure 4. *Evaluation results of GBR and Stacking model.*

To sum up, the proposed stacking-based ensemble model achieves the better predicting results than five base predictive models in most metrics, thus it can be concluded that the stacking integration-based machine learning method is more accurate than individual machine learning methods in predicting house rent.

5. Conclusion

Every year, house rents are in flux, so it is necessary to establish a mechanism to predict the rent in the future. An accurate and reliable predictive model will be of great convenience for renters to solve this problem. Homeowners, renters and policy makers can apply the predictive model to calculate the acceptable rent and facilitate the fair transaction. In this study, a stacking-based ensemble learning model for predicting the rent based on house features is proposed. Specifically, a pre-processing work is performed on the dataset, including decomposition of hybrid features, removal of rent outliers, deletion of uncorrelated features, transformation of data, and standardization. Secondly, the stacking integration method is used to integrate the competent base predictive models to achieve the better performance. Finally, the proposed ensemble model is compared with base predictive models using evaluate metrics such as RMSE, MAE, MAPE and R^2 to verify the effectiveness of the stacking integration method.

In terms of the experimental results, the proposed model in the study achieves the better predictive performance, but there are still disadvantages and improvement spaces. First, the ensemble model can involve more machine learning algorithms such as Bayesian optimization and genetic algorithm for optimizing the hyper-parameters of the model. Secondly, other stacking integration methods (such as bagging method and propulsion method) can be considered for model construction. Thirdly, more evaluation metrics can be used to evaluate the model performance, which contributes to a sufficiently accurate comparison and analysis among different models. Finally, generalization ability of models need be enhanced to facilitate its deployment in the other fields.

References

[1] Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique. *Procedia Computer Science*, 199, 806-813.

[2] Ahmad, A., Anderson, T. N., & Rehman, S. U. (2018). Prediction of electricity consumption for residential houses in New Zealand. In *Proceedings of 3rd International Conference of Smart Grid and Innovative Frontiers in Telecommunications (SmartGIFT)*, Auckland, New Zealand, April 23-24, 2018, pp. 165-172.

[3] Cameron, A. C., & Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, 77 (2), 329-342.

[4] Fikire, A. H. (2021). Determinants of residential house rental price in Debre Berhan Town, North Shewa Zone, Amhara Region, Ethiopia. *Cogent Economics & Finance*, 9 (1), 1904650.

[5] Garzon, M. B., Blazek, R., Neteler, M., de Dios, R. S., Ollero, H. S., & Furlanello, C. (2006). Predicting habitat suitability with machine learning models: the potential area of Pinus sylvestris L. in the Iberian Peninsula. *Ecological modelling*, 197 (3-4), 383-393.

[6] Huang, S., Chang, J., Huang, Q., & Chen, Y. (2014). Monthly streamflow prediction using modified EMD-based support vector machine. *Journal of Hydrology*, 511, 764-775.

[7] Jordan, E. J., & Moore, J. (2018). An in-depth exploration of residents' perceived impacts of transient vacation rentals. *Journal of Travel & Tourism Marketing,* 35 (1), 90-101.

[8] Kumatani, S., Itoh, T., Motohashi, Y., Umezu, K., & Takatsuka, M. (2016, July). Time-varying data visualization using clustered heatmap and dual scatterplots. In *Proceedings of 2016 20th International Conference of Information Visualisation (IV)*, (pp. 63-68), Lisbon, Portugal.

[9] Lim, W. T., Wang, L., Wang, Y., & Chang, Q. (2016). Housing price prediction using neural networks. In *Proceedings of the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Changsha, China, August 13-15, 2016, pp. 518-522.

[10] Mitchell, T. M. (2006). *The Discipline of Machine Learning* (Vol. 9). Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department.

[11] Sharma, G., & Prabha, C. (2021). Applications of machine learning in cancer prediction and prognosis. In *Cancer Prediction for Industrial IoT 4.0: A Machine Learning Perspective* (Eds: Gupta, M., Jain, R., Solanki, A. & AI-Turjman, F.), Chapman and Hall/CRC, pp. 119-135.

[12] Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert systems with applications*, *28* (1), 127-135.

[13] Verma, S. P. (1997). Sixteen statistical tests for outlier detection and rejection in evaluation of international geochemical reference materials: Example of microgabbro PM-S. *Geostandards Newsletter,* 21 (1), 59-75.

[14] Ves, A. V., Ghitescu, N., Pop, C., Antal, M., Cioara, T., Anghel, I., & Salomie, I. (2019, September). A stacking multi-learning ensemble model for predicting near real time energy consumption demand of residential buildings. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 183-189). IEEE.

[15] Yoshida, T., Murakami, D., & Seya, H. (2022). Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset. *The Journal of Real Estate Finance and Economics*, DOI: https://doi.org/10.1007/s11146-022-09929-6.

[16] Zhang, K., Shen, L., & Liu, N. (2019). House rent prediction based on joint model. In *Proceedings of the 8th International Conference on Computing and Pattern Recognition*, Beijing, China, October 23-25, 2019, pp. 507-511.