SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre
p-ISSN 2964-1691 | e-ISSN 2964-8246

Optimization house price prediction model using gradient boosted regression trees (GBRT) and xgboost algorithm

Putri Susi Sundari [1], **Khafidz Putra Mahardika** [2]
[1,2] Department of Informatics Engineering, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received June 26, 2023
Revised July 22, 2023
Accepted July 28, 2023

Keywords:

GBRT
XGBoost
RMSE
Prediction algorithms
Machine learning

ABSTRACT

In this rapidly advancing technological era, the demand for the real estate industry has also increased, including in the field of house price prediction. House prices fluctuate every year due to several factors such as changes in land prices, location, year of construction, infrastructure developments, and other factors. Numerous studies have been conducted on this issue. However, the challenge lies in building a proven accurate and effective model for predicting house prices with the abundance of features present in the dataset. The objective of this research is to develop a predictive model that can accurately estimate house prices based on relevant features or variables. The researcher utilizes ensemble learning techniques, combining the Gradient Boosted Regression Trees (GBRT) and XGBoost algorithms. The dataset used in this article is titled "Ames Housing dataset" obtained from Kaggle. The predictive model is then evaluated using the Root Mean Squared Error (RMSE) method. The RMSE result from this research is 0.0047. It also means that the combination of GBRT and XGBoost algorithms successfully improves the prediction accuracy.

1. Introduction

A house or a place to live is a basic thing that becomes one of the 3 basic human needs besides clothing and food. Every year, house prices fluctuate due to several factors. Some factors that can affect changes in house prices include changes in

[1] ***Corresponding Author:***
Putri Susi Sundari,
Department of Informatics Engineering,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Kota Semarnag, Indonesia.
Email: wagiy841@students.unnes.ac.id

DOI: https://doi.org/10.52465/josre.v1i2.176

land prices and changes in infrastructure both inside and outside the region. Nowadays, predictions are indispensable and as a source to support decisions in upcoming events [1]. In predicting a single house price, a more accurate method is needed based on the type of house, location, local facilities, year of construction, and other factors that affect the demand and supply of houses [2].

The development of technology today provides various conveniences for various life problems. Humans are required to be able to use technology as well as overcome various problems, one of which is to predict house prices, where house price predictions will increase every year, therefore it is necessary to model house price predictions [3]. The real estate market has become very competitive in terms of pricing and fluctuations [4] because it has significant implications for industry and fields related to investment, construction, and public welfare [5]. Machine Learning can be utilized for purposes such as predicting house prices against several known factors. One machine learning algorithm that can be used is Gradient Boosted Regression Tree (GBRT). House price prediction is also facilitated by advances in the visual analysis of the gradient boosting regression tree (GBRT) [6].

Several previous studies have been conducted to make a prediction [7]. Some previous studies have predicted house prices by utilizing machine learning techniques [8], but they were less effective. In previous research entitled "Machine Learning Based House Price Prediction using Modified Extreme Boosting" using the Modified Extreme Gradient Boosting algorithm method. The method used by researchers can select adaptive and probabilistic models, but the method also has several weaknesses such as sensitivity to outliers in the data, fairly low model interpretability, and long training time. Other researchers have also conducted quantitative analysis activities on several variables that can affect house prices, which then produce house price predictions [9]. The selection of the XGBoost algorithm in house price prediction optimization using the GBRT method is due to its ability to cope with complex features, good performance, model interpretability, strong regulation, and proficiency in scalability and speed. Therefore, this study uses the Gradient Boosted Regression Trees (GBRT) algorithm which is a machine learning method used to predict target continuous variables, such as house prices. GBRT combines several simpler regression tree models to incrementally improve prediction accuracy. The GBRT algorithm is very popular in predictive analysis because it can overcome several problems such as overfitting and underfitting, and provides accurate and stable prediction results. Therefore, the GBRT algorithm can be used to help people predict house prices well and make it easier for users to predict an accurate time when they want to buy a house [10]. The development of this model in predicting houses is also expected to be able to help home sellers or real estate in making more informative decisions based on house price assessments [11]. To create an accurate predictive model, appropriate feature engineering is also necessary. In this study, we employed several feature engineering techniques, including data normalization, categorical data encoding, handling missing values, feature selection, and logarithmic transformation.

2. Method

Researchers use the flowchart of the method contained in Figure 1, as follows:

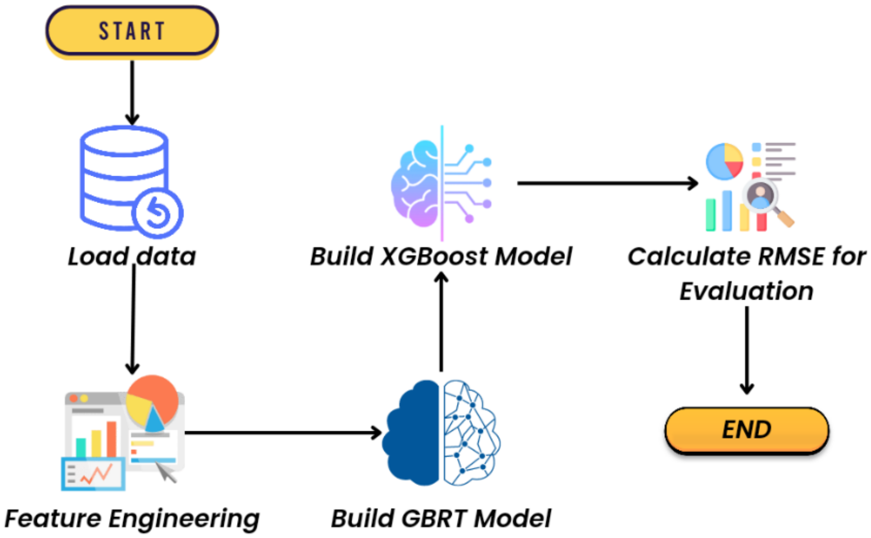

Figure 1. Flowchart of house price prediction method

2.1. Dataset

This research uses the help of Google Collaboratory software in making models [12]. The dataset used in the research to predict house prices comes from the Kaggle Ames Housing dataset which consists of 2,930 samples, where about 80% of the data is used as training data and 20% is used as testing data. The dataset contains housing located in Ames, United States, and is often used as a context in house price prediction. The dataset contains various variables, such as SalePrice (the target variable), OverallQual (the level of quality and finishing of the house, which is given a scale of 1 to 10), GarageCars (the area of the car garage), and various other variables.

2.2. Feature Engineering

Feature engineering in machine learning is very important. Feature engineering involves manipulating, transforming, and creating new features from raw data to improve the quality and representation of relevant information in the dataset. Feature engineering in predicting house prices is explained as follows.

- The first step is to visualize the SalePrice distribution on the dataset. This aims to provide an overview of the data distribution and help understand the characteristics of the data. Figure 2 shows the distribution of SalePrice variables while Figure 3 shows the distribution of variables in each neighborhood.

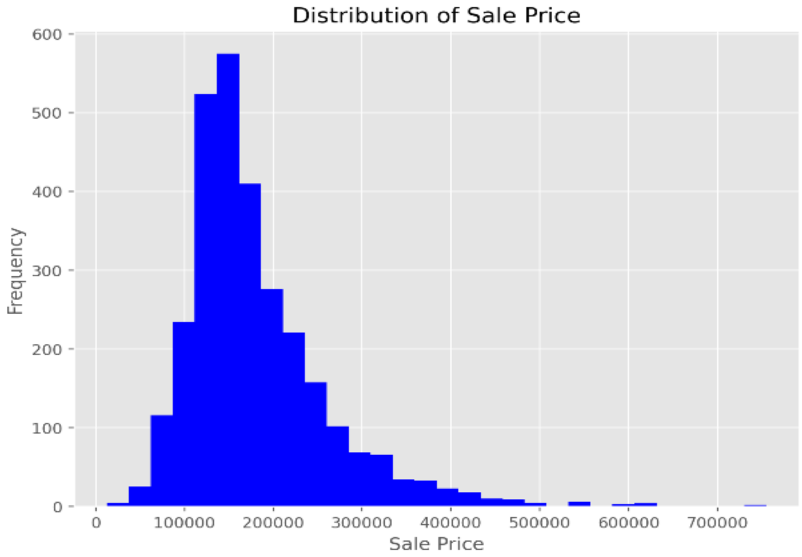

Figure 2. Distribution of the sale price variable

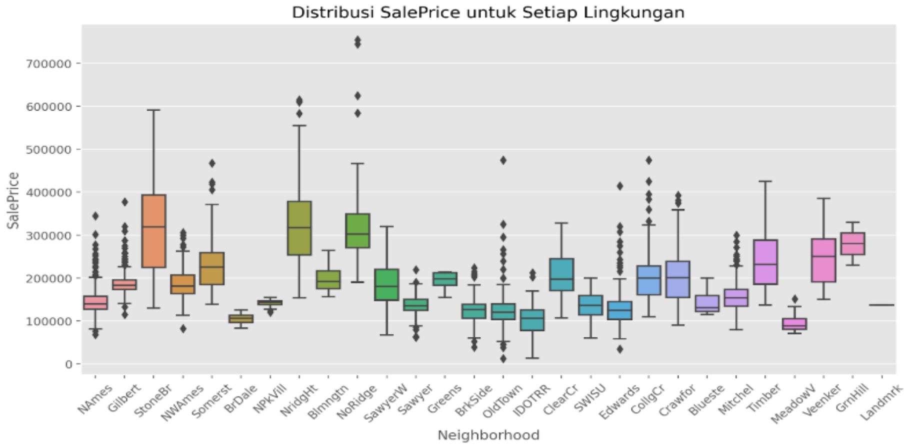

Figure 3. Distribution of sale price for each neighborhood

- The normalisation of 'Overall Qual', 'Gr Liv Area', and 'Garage Cars' features in the dataset.
- Encoding categorical variables which in this case are 'Neighbourhood', 'Exterior 1st', and 'Sale Type' using 'LabelEncoder()'.
- Handling missing values. For numerical features, it is filled with the mean or average. While for categorical features, it is filled with the mode value.
- Feature selection using correlation where we look for features that have a high level of correlation with the 'SalePrice' feature.
- Transform the selected features with the logarithm method.

2.3. Gradient Boosting Regression Tree (GBRT)

The method in this study uses an empirical formula, namely the machine learning gradient boosted regression trees (GBRT) algorithm [13]. GBRT is a learning algorithm based on gradient enhancement that has been proposed by Friedman [14]. The gradient-boosted regression trees (GBRT) algorithm is also a machine learning method that works by combining iterative trees that are hunted over several regression trees [15] in an orderly manner that is used to improve prediction results. Machine learning can facilitate analyzing, identifying patterns, and making predictions that can help users when making decisions [16]. The machine learning process will involve data which will then train the computer using a machine learning model with the help of the GBRT algorithm [17]. The gradient-boosted regression trees (GBRT) algorithm concentrates on reducing various kinds of prediction errors in the previous model by taking a new regression tree. The gradient-boosted trees will also incorporate some of the decision trees but in contrast to random forests. The various regression trees will then be bootstrapped sequentially, where each tree will attempt to correct the errors in the prediction results that have not been resolved by the previous trees. The GBRT algorithm also has the advantage of dealing with numerical, categorical features, and can produce fairly accurate predictions in handling regression problems.

The following are several stages of the method used in predicting house prices using the gradient-boosted regression trees (GBRT) algorithm; The first thing the researcher does is collect data, where the collection of this dataset includes several variables, such as building area, number of bathrooms, number of bedrooms, land area, house price, location, etc. Then at the next stage, the researcher conducted a data pre-processing stage, where this was done to clean the data to overcome invalid values and transform the data if needed, such as standardization or normalization. Furthermore, data sharing is done by dividing the dataset into two subsets, consisting of training data, and testing data. After the data division is complete, training on the GBRT model will be carried out, where the GBRT model will be built by measuring parameters, namely tree depth, learning rate, and number of trees. Training on the GBRT model is done using training data, this is done so that each tree can correct prediction errors in the previous tree. Furthermore, GBRT model training will maximize the model with a technique called cross-validation or grid search to find better accuracy. The parameter used in this study is the learning rate, where the learning rate parameter can control the contribution contained in each ensemble tree, so that a lower value will lead to a conservative model with a much smaller tree contribution value, otherwise if the learning rate is higher, it will produce a much more complex model with a much larger contribution value. The benchmark used in the learning rate parameter is to get a value that can produce a prediction level with a more minimal error or a much more optimal metric result. Evaluation and optimization of the learning rate are done by trying different values and choosing the one that gives optimal results in terms of prediction accuracy.

2.4. XGBoost

XGBoost is an algorithm that combines several weak models into a stronger prediction model. In the context of house price prediction, XGBoost will add prediction capability by incrementally building a new tree. Before that, GBRT will perform the calculation, where the difference between the predicted value and the actual predicted value will occur be the target for building the next model.

2.5. Evaluation Matrix

The evaluation matrix used is Root Mean Squared Error (RMSE). The RMSE metric gives an idea of the extent to which the actual value differs from the predicted value. This metric measures the error value between the predicted value and the actual value on the same scale as the target variable. That means, the smaller the value, the more accurate the prediction model will be. The formula of the RMSE metric is as Equation 1.

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

Where $\hat{y}_i$ is the i-th, y_i predicted value i-th, and n is the number of data.

3. Results and Discussion

3.1 Feature Engineering

Feature Engineering starts by normalizing the 'Overall Qual', 'Gr Liv Area', and 'Garage Cars' features in the dataset. The three features are categorical features that have values with non-comparable scales. With normalization, the categorical features can be well integrated to understand the relationship between the features and the target variable. The histogram of 'Neighbourhood', 'Exterior 1st', and 'Sale Type' features can be seen in Figure 4.

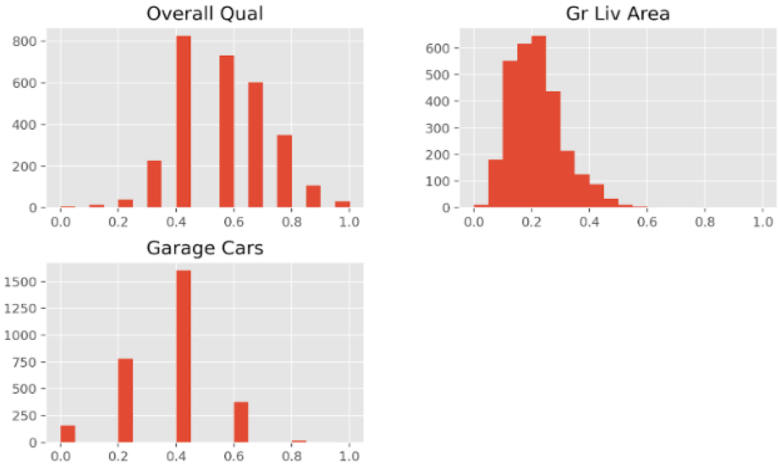

Figure 4. Histogram of normalized features

After normalization is complete, the next step is to encode the categorical variables. This is intended for the model to better understand the dataset. Encoding converts categorical features into numerical representations. Features such as 'Neighbourhood', 'Exterior 1st', and 'Sale Type' are encoded using the 'LabelEncoder()' method from 'sklearn'.

The next process in feature engineering is handling missing values. This is an important part of feature engineering. This is because the presence of missing values in the dataset can affect the quality and accuracy of the analyses performed. When there are missing values, it can interfere with statistical calculations, modeling, and data visualization. In addition, if missing values are not addressed, it can cause bias in the analysis and interpretation of the results. In the dataset used, there are still many missing values that need to be handled. For numerical features, it is filled with the mean or average. While for categorical features, it is filled with the mode value.

The next step is feature selection. Feature selection is used according to the type of data object that is owned [18]. Feature selection is based on the calculation of the correlation matrix.The correlation coefficient in a correlation matrix has a range of values between -1 and 1. In the case of house prediction, we use "absolute value" because we want to know the strength of the linear relationship between the variables regardless of the positive or negative direction. The selected features are those that have an absolute value of the correlation matrix above 0.5.This value was chosen because 0.5 is the middle value of 0-1. Therefore, we assume that a coefficient value greater than 0.5 can be categorized as a strong value.

After that, the selected features are inserted into a new data frame named 'selected_feature'. The selected features include 'Overall Qual', 'Year Built', 'Year Remodelled/Add', 'Mas Vnr Area', 'Total Bsmt SF', '1st Flr SF', 'Gr Liv Area', 'Full Bath', 'Garage Yr Blt', 'Garage Cars', and 'Garage Area'. The Correlation value of selected features with SalePrice can be seen in Table 1.

Table 1. Correlation value of selected features with SalePrice

Features	abs(Correlation)
Overall Qual	0.799262
Gr Liv Area	0.706780
Garage Cars	0.647861
Garage Area	0.640385
Total Bmst SF	0.632105
1st Flr SF	0.621676
Year Built	0.558426

Full Bath	0.545604
Year Remod/Add	0.532974
Garage Yr Blt	0.510684

Next, the data were transformed using the logarithm method. Data transformation can normalize unsymmetrical data distribution. Data transformation is applied to the selected features. This helps to fulfil the assumptions of statistical models that expect normally distributed residuals and improves the visual interpretation and performance of the model in making house price predictions.

3.2 Prediction Model with GBRT and XGBoost

The data used in the model is data that has undergone the previous feature engineering process. The target variable of the prediction model this time is 'SalePrice'. Furthermore, the dataset is divided into training data and test data with a composition of 80:20. This training data is used to train the model, while the test data is used to evaluate the model.

The Gradient Boosting Regressor and XGBoost models are initialized in the next step. The models were then trained using the training data that had previously been created. The prediction results of the models are stored in their respective prediction variables. Then the results of the two predictions are combined by taking the average value of the two variables and evaluating bith model.

3.3 Evaluation Metrics

From the process that has been carried out, the RMSE value of the GBRT model only is 0.0048 and XGBoost only is 0.0070. However, when the two models are combined, the RMSE value becomes 0.0047. This value is smaller than when we use only one model. This means that when we combine the GBRT and XGBoost models, the prediction model becomes more accurate.This value is also smaller than previous studies[1]. The research managed to get the best RMSE value of 0.1126, namely by combining the Lasso and XGBoost methods. The comparison table can be seen in Table 2.

Table 2. Comparison table

Model	RMSE
Lasso+XGBoost [2]	0.1126
XGBoost+Lasso+Ridge [19]	0.1201
GBRT	0.0048
XGBoost	0.0070
bGBRT+XGBoost	0.0047

4. Conclusion

In conclusion, this study aimed to develop a predictive model for accurately estimating house prices based on relevant features or variables in the rapidly advancing real estate industry. By utilizing machine learning techniques, specifically the Gradient Boosted Regression Trees (GBRT) and XGBoost algorithms, a predictive model was built and evaluated using the Root Mean Squared Error (RMSE) method. The obtained RMSE result of 0.0047 shows the successful improvement of prediction accuracy from previous studies. This research contributes to addressing the challenge of predicting house prices with a large number of features on Kaggle's "Ames Housing dataset", and provides insight into the application of ensemble learning algorithms in the field of house price prediction.

REFERENCES

[1] D. R. Damayanti, S. Wicaksono, M. F. Al Hakim, J. Jumanto, S. Subhan, and Y. N. Ifriza, "Rainfall Prediction in Blora Regency Using Mamdani′s Fuzzy Inference System," *J. Soft Comput. Explor.*, vol. 3, no. 1, pp. 62–69, Mar. 2022, doi: 10.52465/joscex.v3i1.69.

[2] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," in *2017 IEEE international conference on industrial engineering and engineering management (IEEM)*, IEEE, 2017, pp. 319–323, doi: https://doi.org/10.1109/IEEM.2017.8289904

[3] M. Thamarai and S. P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning.," *Int. J. Inf. Eng. Electron. Bus.*, vol. 12, no. 2, 2020, doi: http://dx.doi.org/10.5815/ijieeb.2020.02.03

[4] A. P. Singh, K. Rastogi, and S. Rajpoot, "House Price Prediction Using Machine Learning" in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, IEEE, 2021, pp. 203–206. Available online: https://www.irjet.net/archives/V10/i4/IRJET-V10I4194.pdf

[5] Q. Zhang, "Housing price prediction based on multiple linear regression," *Sci. Program.*, vol. 2021, pp. 1–9, 2021. DOI: https://doi.org/10.1155/2021/7678931

[6] Y. Huang, Y. Liu, C. Li, and C. Wang, "GBRTVis: online analysis of gradient boosting regression tree," *J. Vis.*, vol. 22, pp. 125–140, 2019, doi: https://doi.org/10.1007/s12650-018-0514-2

[7] A. F. Mulyana, W. Puspita, and J. Jumanto, "Increased accuracy in predicting student academic performance using random forest classifier," *J. Student Res. Explor.*, vol. 1, no. 2, pp. 94–103, Jul. 2023, doi: 10.52465/josre.v1i2.169.

[8] N. Ragapriya, T. A. Kumar, R. Parthiban, P. Divya, S. Jayalakshmi, and D. R. Raman, "Machine Learning Based House Price Prediction Using Modified Extreme Boosting," *Asian J. Appl. Sci. Technol.*, vol. 7, no. 1, pp. 41–54, 2023. Available online: https://ajast.net/data/uploads/83465.pdf

[9] N. Chen, "House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis" *Wirel. Commun. Mob. Comput.*, vol. 2022, 2022. ,doi: https://doi.org/10.1155/2022/9590704

[10] C. H. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House price prediction using regression techniques: A comparative study" in *2019 International conference on smart structures and systems (ICSSS)*, IEEE, 2019, pp. 1–5, doi: https://doi.org/10.1109/ICSSS.2019.8882834

[11] B. Afonso, L. Melo, W. Oliveira, S. Sousa, and L. Berton, "Housing prices prediction with a deep learning and random forest ensemble" in *Anais do XVI encontro nacional de inteligência artificial e computacional*, SBC, 2019, pp. 389–400, doi:

https://doi.org/10.5753/eniac.2019.9300

[12] A. Amalia, M. Radhi, S. H. Sinurat, D. R. H. Sitompul, and E. Indra, "PREDIKSI HARGA MOBIL MENGGUNAKAN ALGORITMA REGRESSI DENGAN HYPER-PARAMETER TUNING" *J. Sist. Inf. dan Ilmu Komput. Prima (JUSIKOM PRIMA)*, vol. 4, no. 2, pp. 28–32, 2021, doi: https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2479

[13] T. Chen, H. Shang, and Q. Bi, "A prediction method of five-axis machine tool energy consumption with GBRT algorithm" in *2019 IEEE 5th International Conference on Mechatronics System and Robots (ICMSR)*, IEEE, 2019, pp. 34–39, doi: http://dx.doi.org/10.1109/ICMSR.2019.8835459

[14] Y. Wang and Y. Tang, "A recommendation algorithm based on item genres preference and GBRT" in *Journal of Physics: Conference Series*, IOP Publishing, 2019, p. 12053, doi: https://iopscience.iop.org/article/10.1088/1742-6596/1229/1/012053

[15] P. Nie, M. Roccotelli, M. P. Fanti, Z. Ming, and Z. Li, "Prediction of home energy consumption based on gradient boosting regression tree" *Energy Reports*, vol. 7, pp. 1246–1255, 2021, doi: https://doi.org/10.1016/j.egyr.2021.02.006

[16] R.-T. Mora-Garcia, M.-F. Cespedes-Lopez, and V. R. Perez-Sanchez, "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times" *Land*, vol. 11, no. 11, p. 2100, 2022, doi: https://doi.org/10.3390/land11112100

[17] M. Jain, H. Rajput, N. Garg, and P. Chawla, "Prediction of house pricing using machine learning with Python" in *2020 International conference on electronics and sustainable communication systems (ICESC)*, IEEE, 2020, pp. 570–574, doi: http://dx.doi.org/10.1109/ICESC48915.2020.9155839

[18] W. F. Abror, A. Alamsyah, and M. Aziz, "Bankruptcy Prediction Using Genetic Algorithm-Support Vector Machine (GA-SVM) Feature Selection and Stacking" *J. Inf. Syst. Explor. Res.*, vol. 1, no. 2, Jul. 2023, doi: 10.52465/joiser.v1i2.180.

[19] C. Fan, Z. Cui, and X. Zhong, "House prices prediction with machine learning algorithms" in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, 2018, pp. 6–10, doi: http://dx.doi.org/10.1145/3195106.3195133