Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 6, No.3, July 2024, pp. 02403014-01 ~ 02403014-020
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v6i3.602

Improving the Accuracy of House Price Prediction using Catboost Regression with Random Search Hyperparameter Tuning: A Comparative Analysis

Faezal Wahyu Hartono[1], Muljono[1*], Ahmad Zainul Fanani [1]

[1] Informatics Engineering, Computer Science Faculty, Dian Nuswantoro University, Jl. Imam Bonjol No 207 Semarang 50131, Central Java, Indonesia

[*] muljono@dsn.dinus.ac.id

Abstract. Achieving a significant improvement over traditional models, this study presents a novel approach to house price prediction through the integration of Catboost Regression and Random Search Hyperparameter Tuning. By applying these advanced machine learning techniques to the King County Dataset, we conducted a thorough regression analysis and predictive modeling that resulted in a marked increase in accuracy. The baseline model, a conventional linear regression, provided a foundation for comparison, evaluating performance metrics such as R-squared and Mean Squared Error (MSE). The meticulous hyperparameter tuning of the Catboost model yielded a remarkable improvement in predictive accuracy, demonstrating the efficacy of sophisticated data science techniques in real estate and property valuation. The percentage increase in accuracy over the baseline model is explicitly stated in the abstract.

Keywords: Catboost Regression, Hyperparameter Tuning, Random Search, Regression Analysis, Predictive Modeling

(Received 2024-05-19, Accepted 2024-07-03, Available Online by 2024-07-25)

1. Introduction

House price prediction is a topic that plays an important role in investment decisions and the property market. House price prediction can provide benefits to prospective home buyers and property investors. For prospective buyers, house price prediction can provide a reference for future home prices, thus helping them to buy a home at the right time.

For investors, house price prediction can provide a range of house prices in the market, thus enabling them to determine the right price for investment [1]. In 2014-2015, data presented by the Washington Government indicated that King County had the largest population in the Washington area, with 2,052,800 residents. Due to its large population, King County has more potential home buyers and a higher demand for homes, leading to the collection of the most complete and accurate home price data for use in house price prediction.

Most home prices are determined by the shape and size of bedrooms, bathrooms, floors of the house, and the view or scenery of the house. Home prices on the King County scale were tested using regression models consisting of XGBoost Regression, CatBoost Regression, LightGBM Regression, Random Forest Regression, Polynomial Regression, Multiple Linear Regression, Lasso Regression, and Ridge Regression. To determine the most efficient regression method for prediction, specific parameters are needed to compare each regression method.

The parameters intended for comparison are algorithm values such as MSE (Mean Square Error) and RMSE (Root Mean Square Error). By comparing different machine learning algorithms in the analysis of house price prediction, the accuracy value of the algorithm on the King County Dataset can be calculated [2],[3]. Before implementing house price prediction using various machine learning algorithms applied to regression models, data analysis is first conducted, which involves creating or finding a dataset to analyze the data.

Analyzing the entire dataset using graphs or charts is an important factor in regression analysis, which includes: the number of independent variables, the shape of the regression line, and the type of dependent variables. Additionally, using a correlation matrix helps identify variables that have a strong relationship with house prices. The correlation matrix can be analyzed along with other regression methods to build a more accurate house price prediction model.

In the analysis, the correlation matrix can be used to detect redundant features. By providing two features, the correlation analyst based on the existing dataset can measure how strongly one feature indicates another [3],[4],[5],[6],[7],[8]. By applying house price estimates that take into account factors such as property size, location, and amenities. In the real estate industry, it is very important for the community to understand the historical fluctuations in the housing market, which have caused banking crises due to the adverse effects on financial stability and the real economy, resulting in cyclical rises and falls in property prices. Therefore, accurate house price predictions are made to determine house prices efficiently and to allocate house prices through an estimation model that is very important for the financial market to avoid financial economic risks in determining prices set by investors [9],[10],[11],[12],[13].

In this research, the house price prediction will perform a comparison of algorithms, namely Catboost Regression, which is one of the gradient boosting algorithms capable of applying various problems in the literature of machine learning, such as classification, regression, and prediction. Catboost is one of the implementations of gradient boosting, which uses binary decision trees as the basic predictor. Thus, this algorithm becomes an efficient machine learning method in predicting categorical features. Catboost can serve as one of the toolkit methods for the ensemble technique of decision trees from Gradient Boosting Decision Trees (GBDT), which is very efficient. Therefore, this algorithm is generally able to be implemented due to its simplicity and high performance [14],[15],[16],[17],[18].

In addition to using the Catboost Regression algorithm as a regression model, this research also applies Hyperparameter Tuning with an optimization used in this study, namely Random Search. Hyperparameter tuning refers to parameters that have been implemented before the learning process and are not parameters received through the training process. These parameters can be adjusted and can directly affect the model's performance.

Trying various hyperparameter combinations is a very suitable method used in various strategies needed for adjustment to find the most optimal hyperparameters. However, over time, many approaches have been suggested to optimize these hyperparameters, including Grid Search and Random Search. Random Search is a numerical optimization method that does not require the gradient of the function to be optimized. Random Search operates by randomly iterating through the search space.

Random Search replaces the complete selection of all combinations with the necessary Random Search. Random Search can surpass Grid Search, especially if only a few hyperparameters affect the performance of the machine learning algorithm. Random Search not only tries all available hyperparameters but can perform random searches according to the defined hyperparameter space and ends after completing the iterations required. Random Search can provide good performance to assist in deep learning. In the selection, Random Search can be used to find parameters faster with several

iterations used by Grid Search by finding pairs of each parameter to perform each evaluation [14],[19],[20],[21].

CatBoost Regression is the most suitable algorithm for house price prediction because it can efficiently process categorical data, prevent overfitting, and has the capability to integrate with hyperparameter search techniques such as grid search and random search. Therefore, CatBoost is very effective in improving the accuracy of the model to achieve predictions that are precisely and quickly adjusted compared to other boosting algorithms [11],[15],[22]. It is hoped that this research will be able to provide an improvement in the accuracy of house price predictions with the Catboost Regression algorithm model with Hyperparameter Tuning and Random Search optimization to obtain better and more accurate results.

2. Methods

This research requires detailed and complete experimental data as a method comparison between research using the Catboost Regression method in predicting house prices through the discussion provided in this research. Therefore, a methodology is needed for data collection to improve the method in order to produce an evaluation of research results using the method. Catboost Regression based on the type used.

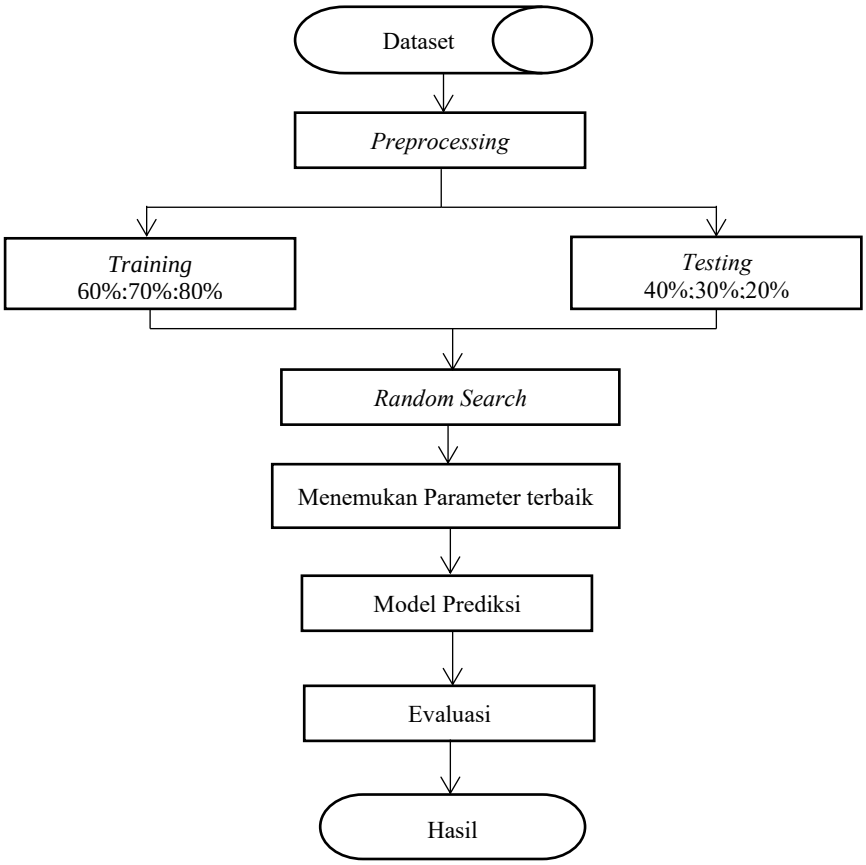

Figure 1. Proposed method experiment scheme

In the use of the CatBoost method for improving house price accuracy, there are several important hyperparameters behind the selection during tuning. These include setting the learning rate value where smaller values require more iteration determining the optimal tree depth, and selecting L2 regularization by trying different values for the regularizer to find the best fit. By using Random Search, it is possible to determine parameter distributions more quickly compared to other hyperparameter tuning methods.

2.1. Proposeed Method

From the methods used, there are stages explained in the research. The method applied in this research project is using regression methods to make house price predictions. The proposed method is expected to facilitate research problems and complete the research objectives.

In this study, the Catboost Regression method with Random Search Hyperparameter Tuning will be proposed as the method. Catboost is a machine learning algorithm that can be used for regression, classification, or ranking. Catboost has several advantages, such as being able to handle categorical data automatically, having definable important features, and having high and stable performance.

Hyperparameter tuning is the process of finding the optimal combination of parameters that can affect the model's outcome. And with random search as part of hyperparameter tuning that performs random searches within a defined parameter space. Thus, random search can save time and resources better compared to other methods. Below are the proposed method steps for implementing house price prediction using the Catboost Regression algorithm with Random Search Hyperparameter Tuning, describing Figure 1 as follows :

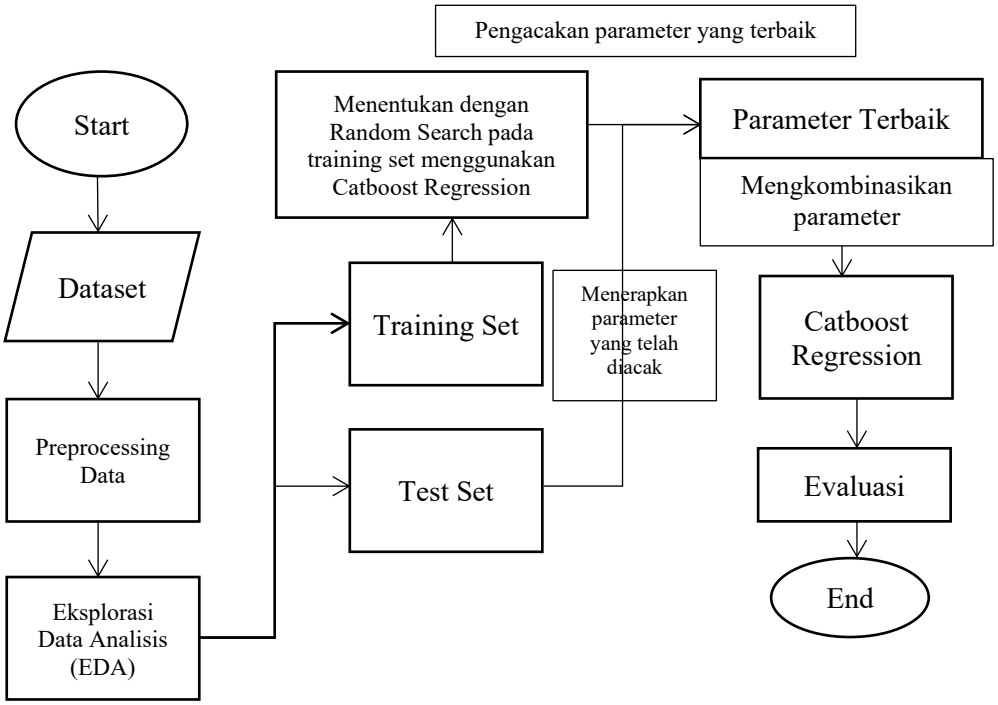

Figure 2. Proposed method experiment scheme

2.2. Dataset

The data used are primary data or more commonly known as public data, as they have been obtained through Kaggle and downloaded from the House Sales in King County database. The data received in this study are qualitative and quantitative sales data from 2014-2015. The dataset contains 21,613 entries, among which there are 20 house features with 8 of them being continuous numerical variables that visualize the broad dimensions through measurements and the geographical position of the house.

The 8 continuous numerical variables are as follows: "sqft_living", "sqft_lot", "sqft_above", "sqft_basement", "lat", "long", "sqft_living15", "sqft_lot15". Additionally, there are other attributes, namely discrete variables that provide clearer information about the components of the house. Discrete variables are predominantly used to count the number of items in the house, such as the number of bedrooms, bathrooms, waterfront, and floors.

There is also background information about the house such as the year of construction, year of renovation, and the previous year's sale price. Moreover, there are 2 evaluation values to assess the overall condition of the house based on different scales and standards. collection data taken from the following link: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction. Below is the dataset of variable data from the house price prediction in the table and figure below.

Table 1. Number of House data in Bedrooms

No	Bedroom Quantity	Number of Bedroom Quantities
1	2 Bedrooms	2760 Houses
2	3 Bedrooms	9824 Houses
3	4 Bedrooms	6882 Houses
4	5 Bedrooms	1601 Houses
5	Others	546 Houses
	Total	21613 Houses

Table 2. Number of House data in Bathrooms

No	Bathroom Quantity	Number of Bathroom Quantities
1	1 Bathrooms	3852 Houses
2	1.5 Bathrooms	1446 Houses
3	1.75 Bathrooms	3048 Houses
4	2 Bathrooms	1930 Houses
5	2.25 Bathrooms	2047 Houses
6	2.5 Bathrooms	5380 Houses
7	Others	3910 Houses
	Total	21613 Houses

Table 3. Number of House data on House Floors

No	House Floor Quantity	Number of House Floor Quantities
1	1 House Floor	10680 Houses
2	1.5 House Floors	1910 Houses
3	2 House Floors	8241 Houses
4	2.5 House Floors	161 Houses
5	3 House Floors	613 Houses
6	3.5 House Floors	8 Houses
	Total	21613 Houses

Table 4. Number of House data on Waterfronts

No	Yes/No Waterfront	Number of Waterfronts on the house
1	No Waterfront	21450 Houses

2	There is a waterfront	163 Houses
	Total	21613 Houses

Table 5. Number of House data in View

No	House View Quantity	House View Quantity Amount
1	No House View	19489 Houses
2	1 House View	332 Houses
3	2 House Views	963 Houses
4	3 House Views	510 Houses
5	4 House Views	319 Houses
	Total	21613 Houses

Table 6. Number of House data on House Conditions

No	House Quantity index of house condition	Total Quantity index of house condition
1	house condition index, namely 1	30 Houses
2	house condition index, namely 2	172 Houses
3	house condition index, namely 3	14031 Houses
4	house condition index, namely 4	5679 Houses
5	house condition index, namely 5	1701 Houses
	Total	21613 Houses

Table 7. Number of House data on House Grade

No	House Quantity house grade index	Total Quantity index of house grade
1	house grade index, namely 6	2038 Houses
2	house grade index, namely 7	8981 Houses
3	house grade index, namely 8	6068 Houses
4	house grade index, namely 9	2615 Houses
5	house grade index, namely 10	1134 Houses
6	Others	777 Houses
	Total	21613 Houses

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	3	7	1180	0
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	3	7	2170	400
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	3	6	770	0
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	5	7	1050	910
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	3	8	1680	0

Figure 3. Dataset of house price predictions

2.3. Data Preprocessing

In the preprocessing stage, data is processed early and well. The preprocessing stage also has various steps, including removing missing or irrelevant data. It involves data analysis exploration, data visualization, and the deletion of unused variable columns. In the initial preprocessing stage after the

dataset collection has been shown, a dataframe object in Python is called, which displays information about an object, a data structure containing a table or data matrix. In Figure 3 above, the output from this dataframe object has 21,363 rows and 21 columns.

```
Out[4]: id               0
        date             0
        price            0
        bedrooms         0
        bathrooms        0
        sqft_living      0
        sqft_lot         0
        floors           0
        waterfront       0
        view             0
        condition        0
        grade            0
        sqft_above       0
        sqft_basement    0
        yr_built         0
        yr_renovated     0
        zipcode          0
        lat              0
        long             0
        sqft_living15    0
        sqft_lot15       0
        dtype: int64
```

Figure 4. Data Cleaning

Next, the dataframe object shows descriptive statistics for each numeric column, such as the average value (mean), standard deviation (std), quartiles (including 25%, 50%, 75%), minimum value (min), and maximum value (max). Figure 4 below shows the output of descriptive statistics for preprocessing data.

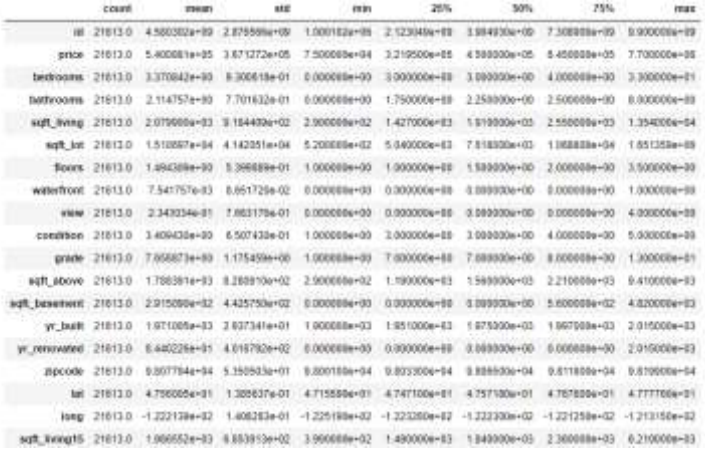

Figure 5. Descriptive Statistics of house prices.

Next, Exploratory Data Analysis (EDA), which is a process for recognizing and analyzing data descriptively before presenting statistical models or more complex algorithms. For this research, the author carried out variable statistics by implementing categorical and numerical variables such as frequency tables, bar charts, and histograms. By also implementing outliers with boxplots and scatterplots. As well as, patterns and relationships on various specific charts.

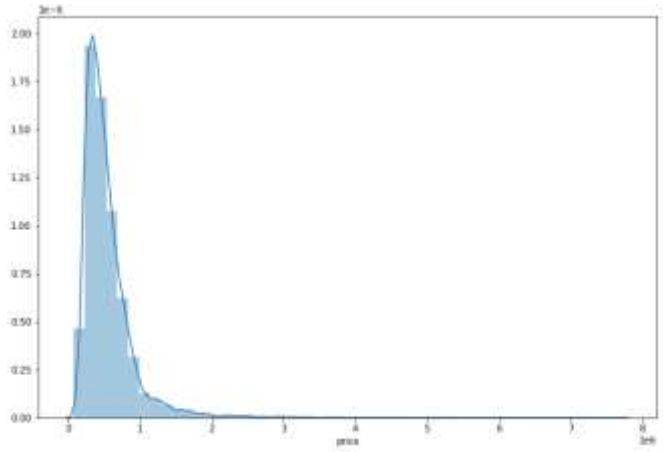

Figure 6. Data visualization on Histogram.

From Figure 5 above, visualize the data in a histogram regarding the house price data column (price) which shows the frequency or number of houses sold within a certain price range.

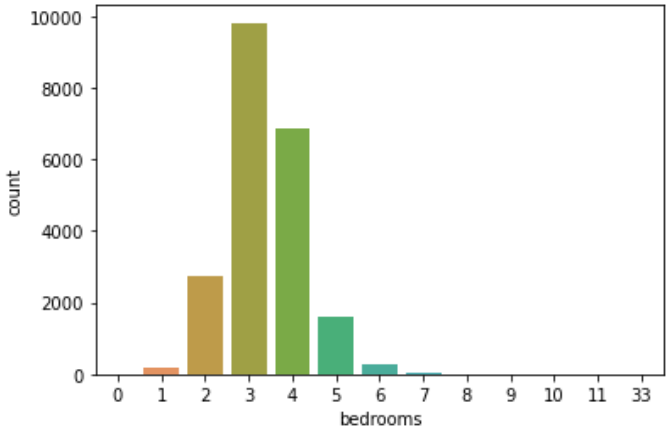

Figure 7. Data visualization on bar charts.

From Figure 6 above, the data in the bar chart relates to the data column for the number of bedrooms which shows the frequency or number of houses sold with a certain number of bedrooms.

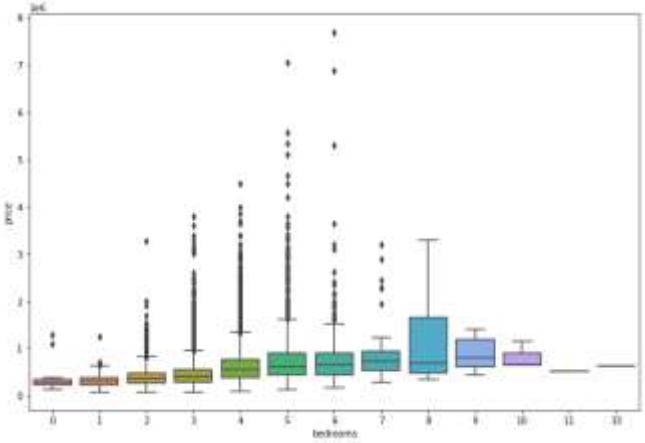

Figure 8. Data visualization in box plots.

From Figure 7 above, the data in the data box plot regarding the data column between house price (price) and number of bedrooms (bedrooms) shows the distribution of data based on five statistical values, namely minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

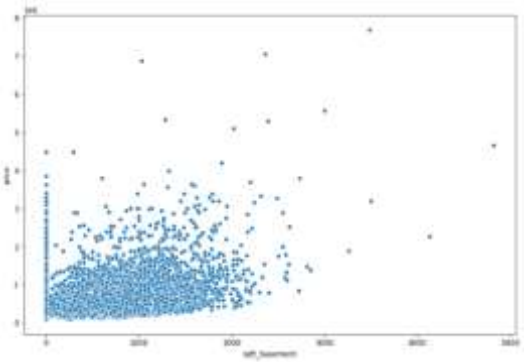

Figure 9. Data visualization on a scatter plot.

From Figure 8 above, the data in the scatter plot data regarding the data column between house price (price) and basement area (sqft_basement) which shows the relationship between two numerical variables which can help see patterns, correlations, outliers from this research data.

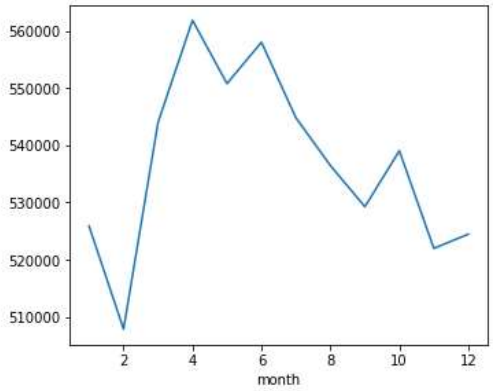

Figure 10. Data visualization on line graphs.

From Figure 9 above, the data in the line graph relates to the house price data column (price) with the month time variable (month) which shows the relationship between two variables, usually the time variable and other variables by displaying a sub plot to create a line graph of the relationship between prices house (price) and sales month (month) in the house price prediction.

2.4. Train Test Split

To begin the experiment in this study, a train test split is performed, which is a data division method consisting of two parts: training data and testing data. The training data can be used as a predictive model, while the testing data can be used to evaluate the performance of the predictive model. In this study, house price prediction will use Catboost Regression by determining the ratio of training data to testing data according to needs, such as 80:20, 70:30, and 40:60.

2.5. Hyperparameter Tuning

After performing the data division (train test split), the next step will be Hyperparameter tuning to find the best parameters using the Random Search method. Hyperparameters play a very important role in adjusting the model to be predicted. In applying parameters, they must be determined before the training process, such as learning rate, depth, iterations, and L2 regularization. Hyperparameters can also be combined with values from parameters to improve accuracy and quality, as well as the performance of the model that provides the best parameter values. The parameter values are chosen from the initial and final limits of the hyperparameters shown in the table below:

Table 8. The limit value of the intended parameter

Parameter Name	Intended Value
learning rate	[0.1, 0.05, 0.01]
iterations	[1000, 2000, 3000]
depth	[4, 5, 6]
l2_leaf_reg	[3, 5 ,7]

2.6. Random Search

After determining the parameters to be used for hyperparameter tuning, the process continues with a random search to process the search for combinations of hyperparameter values randomly to achieve the best parameter values. In the optimization process, random search uses the Catboost Regression prediction model in house price prediction, applying hyperparameters that will be optimized to determine the possible value range for each parameter. To test this optimization, data is first divided to determine the best hyperparameters by comparing catboost regression with data division ratios in testing, including 80:20 (0.2), 70:30 (0.3), and 60:40 (0.4). Below is a table that is the result of a random search to find the best parameters that have been shown:

Table 9. Test data division = 0.2 (ratio 80:20) of the best parameters

Data Split Ratio (Testing Data)	Parameter Name	Value
80:20 (0,2)	learning rate	0.05
	depth	5
	iterations	3000
	l2_leaf_reg	4

Table 10. Test data division = 0.3 (ratio 70:30) of the best parameters

Data Split Ratio (Testing Data)	Parameter Name	Value
70:30 (0,3)	learning rate	0.1
	depth	5
	iterations	1000
	l2_leaf_reg	5

Table 11. Test data division = 0.4 (ratio 60:40) of the best parameters

Data Split Ratio (Testing Data)	Parameter Name	Value
60:40 (0,4)	learning rate	0.05
	depth	5
	iterations	3000
	l2_leaf_reg	4

2.7. Catboost Regression

For this research, the most important thing to do in predicting house prices is to use the Catboost Regression algorithm. Catboost Regression is the most optimal boosting algorithm used, especially for house price predictions. Catboost also provides advantages, such as providing various cost functions and evaluation metrics namely MAE, MSE, MAPE, RMSE, R2, Adj R2, etc. By using the Catboost

model, it is able to predict the price of the house being tested using optimal hyperparameter tuning with random search for randomization in training and testing data.

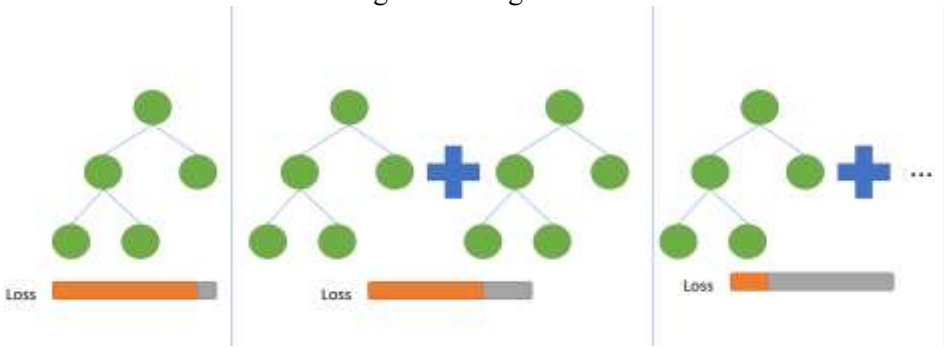

Figure 11. Catboost Regression Structure.

2.8. Analyze the Evaluation of Results

After making predictions with the model, the prediction results are analyzed by calculating metric values from the tested data. The performance evaluations used to measure are MAE, MSE, MAPE, RMSE, R2, and Adj R2. This aims to identify predicted house prices in order to produce optimal accuracy. The following is the formula used to calculate these various evaluation metrics:

$$\text{MAE} = \frac{|\text{p}_1 - \text{b}_1| + \cdots + |\text{p}_\text{n} - \text{b}_\text{n}|}{\text{n}} \ldots (1)$$

$$\text{MSE} = \sum \frac{(\text{y}_\text{i}' - \text{y}_\text{i})^2}{\text{n}} \ldots (2)$$

$$MAPE = \frac{\sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right|}{n} X\ 100\% \ldots (3)$$

$$\text{RMSE} = \sqrt{\sum \frac{(\text{y}_\text{i}' - \text{y}_\text{i})^2}{\text{n}}} \ldots (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i')^2} \ldots (5)$$

$$R_{adj}^2 = 1 - \left(\frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i'\right)^2} \right) \left(\frac{n-1}{n-k-1} \right) \ldots (6)$$

3. Results and Discussion

In these results, it shows a comparison of the evaluation metric outcomes using random search hyperparameter tuning with 3 other algorithms, namely Catboost Regression, Decision Tree, and XGBoost Regression through training data and testing data. The evaluation metrics that will be used to accurately predict house prices are MAE, MSE, MAPE, RMSE, R^2, and Adj R^2 for predicting house prices with high accuracy.

3.1. Catboost Regression

The following are the results of evaluating house price predictions using Catboost Regression which was processed on training data as follows:

Table 12. Evaluation results from Catboost Regression house price predictions on training data

Data Split Ratio (Testing Data)	Evaluation Metrics	Results
80:20 (0,2)	MAE	48795.91
	MSE	5091220542.88
	MAPE	10.39
	RMSE	71352.79
	R2	0.96
	Adj R2	0.96
70:30 (0,3)	MAE	47323.64
	MSE	4667469149.95
	MAPE	10.18
	RMSE	68318.88
	R2	0.96
	Adj R2	0.96
60:40 (0,4)	MAE	46701.95
	MSE	4502132230.77
	MAPE	10.06
	RMSE	67097.93
	R2	0.96
	Adj R2	0.96

Table 1 shows the evaluation results of the catboost regression model using hyperparameter tuning which selected the best parameters using random search in predicting house prices. For data sharing, the best results from training data on all evaluation metrics, namely a ratio of 60:40 (0.4). After carrying out comparisons, the ratio of the training data will be continued with the testing data. The following are the results of evaluating house price predictions using Catboost Regression which was processed on test data as follows:

Table 13. Evaluation results from Catboost Regression house price predictions on testing data

Data Split Ratio (Testing Data)	Evaluation Metrics	Results
80:20 (0,2)	MAE	66156.36
	MSE	16087591527.14
	MAPE	12.19
	RMSE	126836.87
	R2	0.894
	Adj R2	0.893

70:30 (0,3)	MAE	65102.15
	MSE	13577608352.87
	MAPE	12.25
	RMSE	116522.99
	R2	0.906
	Adj R2	0.906
60:40 (0,4)	MAE	65332.27
	MSE	14754893571.89
	MAPE	12.30
	RMSE	121409.72
	R2	0.901
	Adj R2	0.901

Table 2 shows the evaluation results of the Catboost Regression model using hyperparameter tuning which selected the best parameters using random search in predicting house prices. For data sharing, the best results from training data on the overall evaluation metric, namely 70:30 (0.3).

3.2. Decision Tree

The following are the results of evaluating house price predictions using Decision Tree which was processed on training data as follows:

Table 14. Evaluation results from Decision Tree house price predictions on training data

Data Split Ratio (Testing Data)	Evaluation Metrics	Results
80:20 (0,2)	MAE	67556.71
	MSE	15222113498.35
	MAPE	12.44
	RMSE	123377.93
	R2	0.88
	Adj R2	0.88
70:30 (0,3)	MAE	67987.01
	MSE	15232739285.19
	MAPE	12.52
	RMSE	123420.98
	R2	0.88
	Adj R2	0.88
60:40 (0,4)	MAE	68715.59
	MSE	15461859736.28
	MAPE	12.73
	RMSE	124345.73
	R2	0.88
	Adj R2	0.88

Table 3 shows the results of evaluating the decision tree model using hyperparameter tuning which selected the best parameters using random search in predicting house prices. For data sharing, the best results from training data on all evaluation metrics, namely a ratio of 80:20 (0.2). After carrying out comparisons, the ratio of the training data will be continued with the testing data. The following are the results of evaluating house price predictions using Decision Tree which was processed on test data as follows:

Table 15. Evaluation results from Decision Tree house price predictions on testing data

Data Split Ratio (Testing Data)	Evaluation Metrics	Results
80:20 (0,2)	MAE	90010.28
	MSE	31250447165.96
	MAPE	16.17
	RMSE	176777.96
	R2	0.793
	Adj R2	0.792
70:30 (0,3)	MAE	89855.20
	MSE	30303615786.07
	MAPE	16.32
	RMSE	174079.34
	R2	0.79
	Adj R2	0.79
60:40 (0,4)	MAE	91338.26
	MSE	30253824760.14
	MAPE	16.57
	RMSE	173936.27
	R2	0.798
	Adj R2	0.797

Table 4 shows the results of evaluating the decision tree model using hyperparameter tuning which selected the best parameters using random search in predicting house prices. For data sharing, the best results from training data on all evaluation metrics, namely a ratio of 60:40 (0.4).

3.3. XGBoost Regression

The following are the results of evaluating house price predictions using XGBoost Regression which was processed on training data as follows:

Table 16. Evaluation results from XGBoost Regresssion house price predictions on training data

Data Split Ratio (Testing Data)	Evaluation Metrics	Results
80:20 (0,2)	MAE	36366.76
	MSE	2644034032.61
	MAPE	8.12
	RMSE	51420.17
	R2	0.98
	Adj R2	0.98

70:30 (0,3)	MAE	35675.77
	MSE	2542945598.95
	MAPE	7.99
	RMSE	50427.63
	R2	0.98
	Adj R2	0.98
60:40 (0,4)	MAE	34197.59
	MSE	2305117363.13
	MAPE	7.70
	RMSE	48011.64
	R2	0.98
	Adj R2	0.98

Table 5 shows the results of evaluating the XGBoost Regression model using hyperparameter tuning which selected the best parameters using random search in predicting house prices. For data sharing, the best results from test data on all evaluation metrics, namely a ratio of 60:40 (0.4). After carrying out comparisons, the ratio of the training data will be continued with the testing data. The following are the results of evaluating house price predictions using XGBoost regression which was processed on test data as follows:

Table 17. Evaluation results from XGBoost Regresssion house price predictions on testing data

Data Split Ratio (Testing Data)	Evaluation Metrics	Results
80:20 (0,2)	MAE	66983.77
	MSE	18725667841.13
	MAPE	12.26
	RMSE	136841.76
	R2	0.876
	Adj R2	0.876
70:30 (0,3)	MAE	67079.56
	MSE	17990072069.77
	MAPE	12.45
	RMSE	134127.07
	R2	0.875
	Adj R2	0.875
60:40 (0,4)	MAE	67018.61
	MSE	17640995113.08
	MAPE	12.45
	RMSE	132819.41
	R2	0.882
	Adj R2	0.882

Table 6 shows the evaluation results of the XGBoost Regression model using hyperparameter tuning which selected the best parameters using random search in predicting house prices. For data sharing, the best results from test data on all evaluation metrics, namely a ratio of 60:40 (0.4).

3.4. Results Comparison of evaluation metrics

Based on the results that have been tested by comparing several other prediction models, using public datasets via Kaggle on the website, namely "House sales in King County" by applying 3 machine learning algorithm models such as Catboost, Decision Tree, and XGBoost Regression. As well as using Hyperparameter Tuning according to each model algorithm. To determine the prediction results obtained by carrying out evaluation metrics as the results that will be determined by this prediction model, namely MAE, MSE, MAPE and RMSE.

From these results, a table and bar graph will be shown which displays the metric evaluation of the model prediction results through data distribution. Parameter values have been determined using random search by looking for the best parameters as follows:

Table 18. Comparison of house price prediction models of the best parameters based on the overall evaluation metrics results

	MAE	MSE	MAPE	RMSE
Catboost	65102.15	13577608352.87	12.19	116522.99
Decision Tree	89855.2	30253824760.14	16.17	173936.27
XGBoost	67018.61	17640995113.08	12.26	132819.41

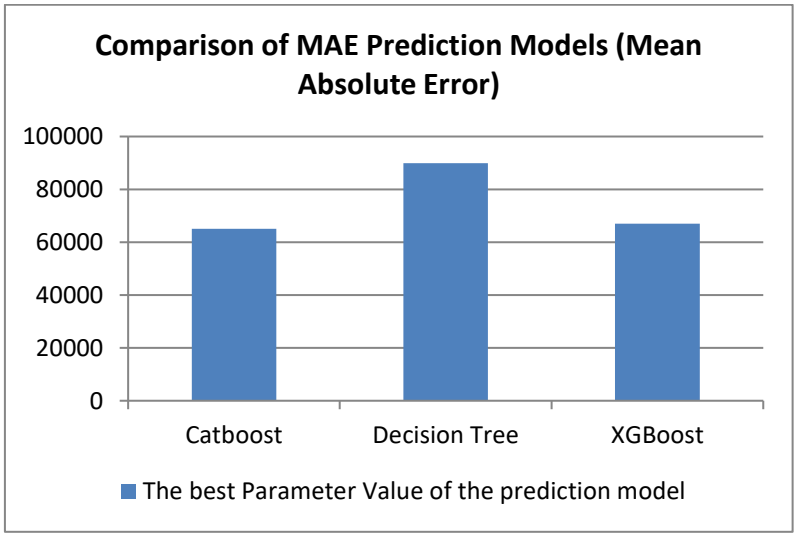

Figure 12. Graph based on comparison of prediction models of the best parameters for MAE evaluation metric results

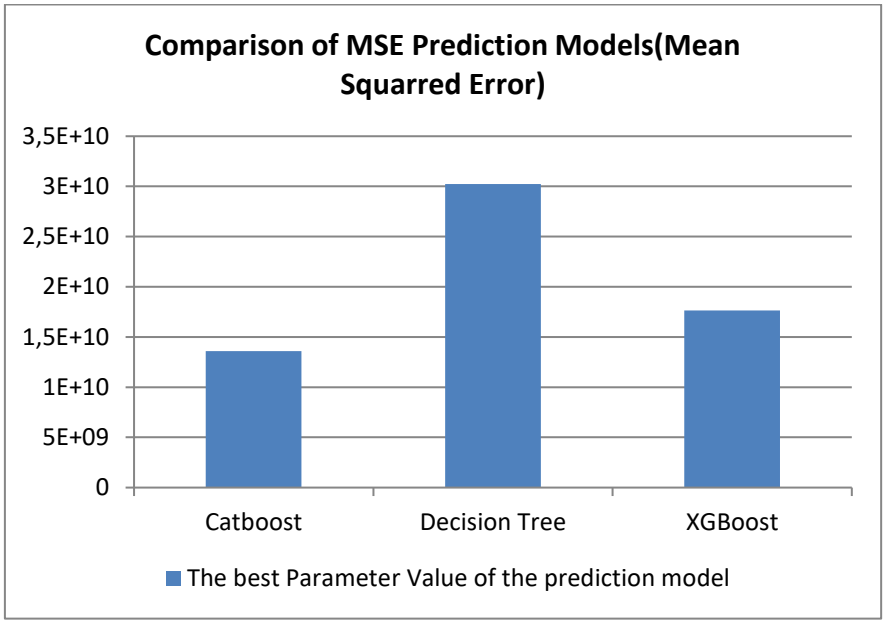

Figure 13. Graph based on comparison of prediction models of the best parameters for MSE evaluation metric results

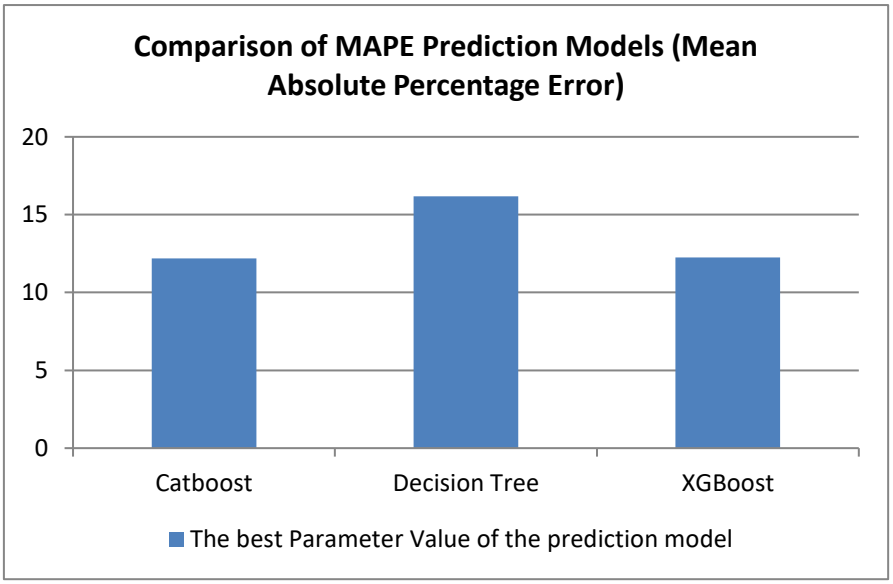

Figure 14. Graph based on comparison of prediction models of the best parameters for MAPE evaluation metric results

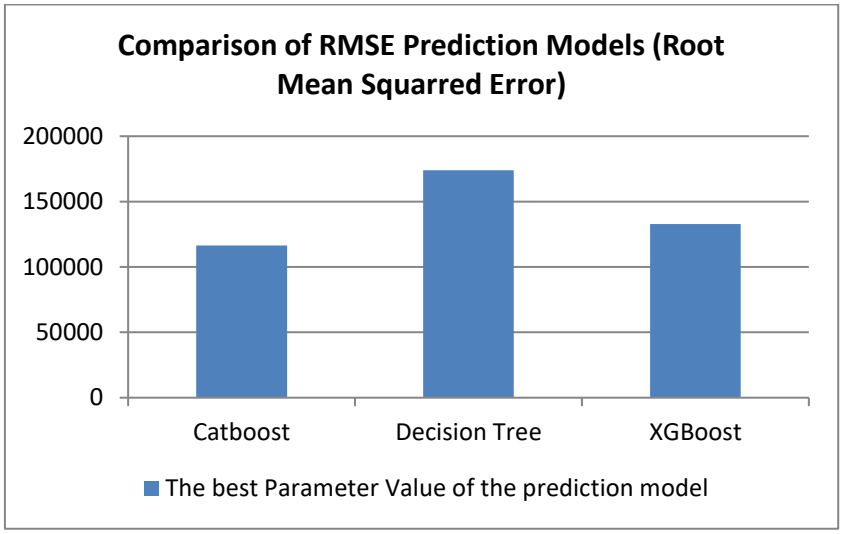

Figure 15. Graph based on comparison of prediction models of the best parameters for RMSE evaluation metric results

From the overall results, the best evaluation metric is determined based on the lowest value. In Figure 1, it shows that Catboost Regression is the MAE value from the results of the best evaluation metric in the prediction model from the best parameters, namely 65102.15. In Figure 2, it shows that Catboost Regression is the MSE value from the results of the best evaluation metric on the prediction model from the best parameters, namely 13577608352.87.

In Figure 3 shows that Catboost Regression is the MAPE value from the results of the best evaluation metric in the prediction model from the best parameters, namely 12.19. In Figure 4, it shows that Catboost Regression is the RMSE value from the results of the best evaluation metric on the prediction model from the best parameters, namely 116522.99. In this discussion, it can be concluded that the results of the model comparison experiments in this test prove that Catboost Regression is the most optimal prediction model and is able to provide increased accuracy in predicting house prices compared to other prediction models such as Decision Tree and XGBoost.

In the overall comparison results of evaluating house price prediction metrics using CatBoost regression and comparing it with other models, it has been found beneficial for various real-world applications. For instance, in the context of predicting house prices, a real estate agent aims to predict house prices based on features such as land area, number of bedrooms, location, and amenities. By optimizing hyperparameters in the CatBoost model, the real estate agent can improve the accuracy of house price predictions, assisting both buyers and sellers in decision-making.

4. Conclusion

For this research, predict house prices using a dataset via Kaggle, namely "House Sales In King County" and for the prediction model that will be used, namely Catboost Regression with Random Search optimization. These steps are aimed at comparing the Catboost Regression algorithm prediction model with other prediction models, such as Decision Tree and XGBoost Regression. This is done in order to determine which algorithm is best at predicting house prices. And with parameter values that have been combined optimally using random search to randomize parameter values to determine the best parameters. Therefore, the following conclusions have been drawn:

1. Catboost Regression using Hyperparameter Tuning Random Search is able to provide good model performance and increased accuracy that is more accurate compared to other prediction models.

2. From the distribution of data that has been tested by dividing the data into two, 3 experiments were carried out, namely 80:20 (0.2), 70:30 (0.3), 60:40 (0.4) using Catboost Regression by analyzing the overall results Evaluation metrics show that the best ratio is 70:30 (0.3).
3. Parameters that are suitable for finding the best results through comparing the data sharing ratios from this research are 'iterations=1000', 'learning rate=0.1', 'depth=5', and 'l2_leaf_reg=5'.
4. To predict house prices, a model that is suitable for use in the future is Catboost Regression using Hyperparameter Tuning Random Search.

In this study, it was able to achieve the best accuracy in predicting house prices. Therefore, to provide the best model performance for this case study, the following recommendations are given:

1. By using Catboost Regression via Hyperparameter Tuning Random Search, it is able to predict several different data to analyze the data, resulting in fast and accurate accuracy values.
2. Can carry out further analysis to find out what factors influence house prices, so that you can determine relevant features and eliminate irrelevant features.
3. It is recommended to explore optimizations in Hyperparameter Tuning in addition to Random Search to determine optimal parameter values efficiently and accurately.

References

[1] R. E. Febrita, A. N. Alfiyatin, H. Taufiq, and W. F. Mahmudy, "Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Bali: IEEE, Oct. 2017, pp. 351–358. doi: 10.1109/ICACSIS.2017.8355058.

[2] Y. Wang and Q. Zhao, "House Price Prediction Based on Machine Learning: A Case of King County," *Business and Management Research*, vol. 211.

[3] Ch. 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)*, Chennai, India: IEEE, Mar. 2019, pp. 1–5. doi: 10.1109/ICSSS.2019.8882834.

[4] V. S. Rana, J. Mondal, A. Sharma, and I. Kashyap, "House Price Prediction Using Optimal Regression Techniques," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India: IEEE, Dec. 2020, pp. 203–208. doi: 10.1109/ICACCCN51052.2020.9362864.

[5] M. Chakraborty, A. Mukhopadhyay, and U. Maulik, "A Comparative Analysis of Different Regression Models on Predicting the Spread of Covid-19 in India," in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India: IEEE, Oct. 2020, pp. 519–524. doi: 10.1109/ICCCA49541.2020.9250748.

[6] Kavitha S, Varuna S, and Ramya R, "A comparative analysis on linear regression and support vector regression," in *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, Coimbatore, India: IEEE, Nov. 2016, pp. 1–5. doi: 10.1109/GET.2016.7916627.

[7] V. Chouvatut and S. Wattanapairotrat, "Feature Reduction from Correlation Matrix for Classification of Two Basil Species in Common Genus," in *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Chonburi, Thailand: IEEE, Jul. 2019, pp. 375–380. doi: 10.1109/JCSSE.2019.8864221.

[8] P. Kapoor, P. K. Singh, and A. K. Cherukuri, "IT Act Crime Pattern Analysis using Regression and Correlation Matrix," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India: IEEE, Jun. 2020, pp. 1102–1106. doi: 10.1109/ICRITO48877.2020.9197835.

[9] N. Peng, K. Li, and Y. Qin, "Leveraging Multi-Modality Data to Airbnb Price Prediction," in *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*, Chongqing, China: IEEE, Nov. 2020, pp. 1066–1071. doi: 10.1109/ICEMME51517.2020.00215.

[10] C. Wang and Q. Gao, "High and Low Prices Prediction of Soybean Futures with LSTM Neural Network," in *2018 IEEE 9th International Conference on Software Engineering and Service*

Science (ICSESS), Beijing, China: IEEE, Nov. 2018, pp. 140–143. doi: 10.1109/ICSESS.2018.8663896.
[11] J. Ding, Z. Chen, L. Xiaolong, and B. Lai, "Sales Forecasting Based on CatBoost," in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, Guangzhou, China: IEEE, Dec. 2020, pp. 636–639. doi: 10.1109/ITCA52113.2020.00138.
[12] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," in *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, Sydney, Australia: IEEE, Dec. 2018, pp. 35–42. doi: 10.1109/iCMLDE.2018.00017.
[13] A. Varma, A. Sarma, S. Doshi, and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore: IEEE, Apr. 2018, pp. 1936–1939. doi: 10.1109/ICICCT.2018.8473231.
[14] M. R. Mubarok and R. Herteno, "HYPER-PARAMETER TUNING PADA XGBOOST UNTUK PREDIKSI KEBERLANGSUNGAN HIDUP PASIEN GAGAL JANTUNG," vol. 09, 2022.
[15] A. A. Ibrahim, R. L., M. M., R. O., and G. A., "Comparison of the CatBoost Classifier with other Machine Learning Methods," *IJACSA*, vol. 11, no. 11, 2020, doi: 10.14569/IJACSA.2020.0111190.
[16] M. J. M. M., U. S., M. B. P., and S. G. Sandhya, "Detection of ransomware in static analysis by using Gradient Tree Boosting Algorithm," in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India: IEEE, Jul. 2020, pp. 1–5. doi: 10.1109/ICSCAN49426.2020.9262315.
[17] S. V. Boyapati, M. S. Karthik, K. Subrahmanyam, and B. R. Reddy, "An Analysis of House Price Prediction Using Ensemble Learning Algorithms," *Research Reports on Computer Science*, pp. 87–96, May 2023, doi: 10.37256/rrcs.2320232639.
[18] M. Massaoudi, S. S. Refaat, H. Abu-Rub, I. Chihi, and F. S. Wesleti, "A Hybrid Bayesian Ridge Regression-CWT-Catboost Model For PV Power Forecasting," in *2020 IEEE Kansas Power and Energy Conference (KPEC)*, Manhattan, KS, USA: IEEE, Jul. 2020, pp. 1–5. doi: 10.1109/KPEC47870.2020.9167596.
[19] Fatihah Rahmadayana and Yuliant Sibaroni, "Sentiment Analysis of Work from Home Activity using SVM with Randomized Search Optimization," *RESTI*, vol. 5, no. 5, pp. 936–942, Oct. 2021, doi: 10.29207/resti.v5i5.3457.
[20] A. Nugroho and H. Suhartanto, "Hyper-Parameter Tuning based on Random Search for DenseNet Optimization," in *2020 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, Indonesia: IEEE, Sep. 2020, pp. 96–99. doi: 10.1109/ICITACEE50144.2020.9239164.
[21] P. Liashchynskyi and P. Liashchynskyi, "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS." arXiv, Dec. 12, 2019. Accessed: Jan. 23, 2024. [Online]. Available: http://arxiv.org/abs/1912.06059
[22] R. Shiller, "Understanding Recent Trends in House Prices and Home Ownership," National Bureau of Economic Research, Cambridge, MA, w13553, Oct. 2007. doi: 10.3386/w13553.