House Price Prediction using Multiple Linear Regression and KNN

Fransiskus Dwi Febriyanto
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ffebriyanto@gmail.com

Endroyono
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
endroeleven@gmail.com

Yoyon Kusnendar
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
yoyonsuprapto@ee.its.ac.id

***Abstract*—The transition of BPHTB management from central taxes to regional taxes is a continuation of the regional autonomy policy. The difference between the market value and the prevailing NJOP poses a challenge for the Sintang District Government in determining the Tax Object Acquisition Value (NPOP) as the basis for imposing BPHTB. Machine learning has been extensively explored for predictions and can be an alternative that can help predict NPOP, especially house prices. This study uses backward elimination and forward selection methods to select the features used in this study and multiple linear regression and K-Nearest Neighbor methods to make house price prediction models. The results of the model performance measurement using RMSE, Multiple Linear Regression method with feature selection using backward elimination resulted in a better model with an RMSE value of 44.02 (million rupiahs) and an R2 value of 0.707.**

Keywords— Backward Elimination, Forward Selection, House Price Prediction, KNN, Multiple Linear Regression

I. INTRODUCTION

Law (UU) Number 28 of 2009 concerning Regional Taxes and Regional Levies in lieu of Law Number 34 of 2000 increases regional authority in managing regional taxes and regional levies.[1] One of them is the transfer of authority for managing Land and Building Acquisition Fees (BPHTB) from the central tax to local taxes, which is a continuation of the regional autonomy policy.

BPHTB is a tax imposed on the acquisition of rights to land and or buildings, where the taxpayer is the party who obtains the acquisition of rights to land and or buildings. BPHTB taxpayers can be individuals or entities.

The development of a region greatly affects tax revenues in this sector, including in Sintang Regency. Based on data from the Central Statistics Agency (BPS), Sintang Regency in 2019 experienced a relatively moderate population growth of 1.31% (percent) [2]. This certainly affects the increasing demand for housing, which can potentially increase BPHTB revenues.

Based on the Sintang District Regulation Number 11 of 2011, the basis for the imposition of BPHTB is the Acquired Value of Tax Objects (NPOP). NPOP is the transaction price or market value. If the transaction price or market value is unknown or lower than the Selling Value of the Tax Object (NJOP), then the basis of imposition is NJOP [3].

The low NJOP determination as the basis for the imposition of Land and Building Tax (PBB) on market value [4] resulted in the reported acquisition value being vulnerable to manipulation [5], [6]. One of them is by using the NJOP listed on the PBB Payable Tax Return (SPPT) in reporting the acquisition value and not using the original buying and selling price, which is the result of an agreement between the seller and the buyer. This is a challenge for local governments to validate the acquisition value.

This study utilizes a machine learning approach to predict the acquisition value. Over the years, machine learning techniques have been extensively explored for prediction. Prediction of house prices has been widely used for taxation, the housing market, banking (as loan guarantees), or insurance [7].

Machine learning builds algorithms and builds models from data, then applies them to new data to make predictions. Linear Regression, KNN, Neural Networks, and Deep Learning are some of the popular machine learning algorithms.[8]–[12].

A study on forecasting home sales in Lamongan [13] used the multiple linear regression method to predict the number of buyers using 144 data consisting of one dependent variable and two independent variables resulting in a model with a mean squared error of 5.557. The next study uses Multiple Linear Regression to predict house prices in Mumbai City in 2018[14] and gives a minimum prediction error of 0.3713. The study [15] used backward elimination in multiple linear regression, which increased estimation accuracy with fewer parameters and smaller RMSE (reduced to 14.81%).

W. Zhao, C. Sun, and J. Wang apply KNN and weighted-KNN to predict used house prices [12]. And in research [16], using a combination of KNN and forward selection increases the estimation results.

Based on previous research, this study tries to provide an alternative house price prediction in the Sintang sub-district using the Multiple Linear Regression and KNN methods. The best prediction model will be determined with the smallest Root Mean Square Error (RMSE) value. This research differs from similar research because it was conducted in a different area, precisely in the Sintang District. It is hoped that this research will be helpful for the Regional Government of Sintang Regency as a material consideration in the process of evaluating houses that are used as a primary reference for tax imposition.

II. METHODS

The methods and workings of this study include 5 (five) main stages, namely data collection, preprocessing, feature selection, predictive modeling, and model performance evaluation, as shown in Fig. 1.

 How to cite: Febriyanto, F. D., Endroyono & Kusnendar, Y. (2023). House Price Prediction using Multiple Linear Regression and KNN. *JAREE (Journal on Advanced Research in Electrical Engineering), 7(1).*

A. Data Collection

The data comes from the BPHTB database and files in the housing sector at the Badan Pengelola Pendapatan Daerah of Sintang Regency from 2017 to 2019. This data consists of 20 variables and 941 observations.

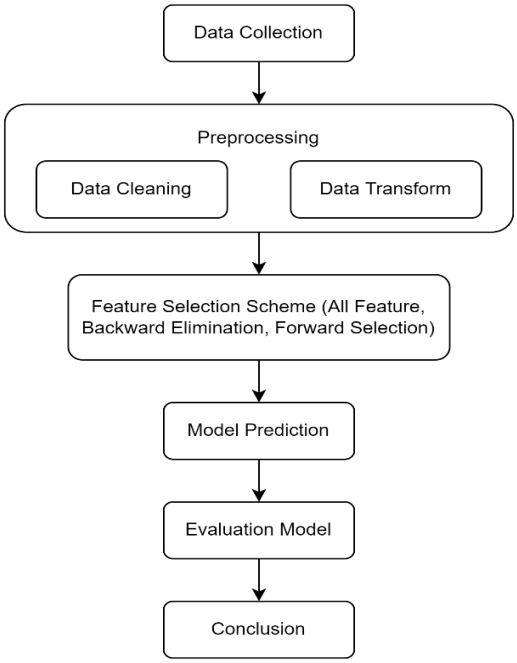

Fig. 1. Research Methodology

. The data variables are based on several factors influencing house prices, including physical properties, legality, location, and time. The variable "price" is the numerical dependent variable, which is the value of the house-buying and selling transaction (Table I). While other variables will be used as independent variables/predictors.

TABLE I. VARIABLES USED

Category	Variable (Initial Code)	Description
Dependent Variable	Price (y)	the house price (million rupiahs)
Time	Year (x_1)	year of acquisition of land and buildings
Location	Road width (x_2)	the width of roads to access the object of land and buildings (meters)
	road function (x_3)	the function of roads to access the object of land and buildings (environment:1, local:2, collector:3 dan artery:4)
	School distance (x_4)	the distance of the object of land and buildings to educational facilities (meters)
	Hospital distance (x_5)	the distance of the object of land and buildings to health facilities (meters)
	shop distance (x_6)	the distance of the object of land and buildings to the market/shops (meters)
Physical properties	Land area (x_7)	The land area of the object (square meter)
	Building area (x_8)	the total floor area of the building, including terraces, balconies, and other outbuildings (square meters)
	Number of floors (x_9)	The number of building floors
	Building condition (x_{10})	the general condition of the building (Poor:1, Moderate:2, Good:3, and Very Good:4)
	Construction (x_{11})	the main building construction materials (Wood:1, Brick:2, Concrete:3, or Steel:4)
	Roof (x_{12})	the main building roof materials (Zinc:1, Asbestos:2, Ordinary/Shingle Tiles:3, Concrete/Aluminum Tiles:4, or Decrabon/Concrete/Glazed Tiles:5)
	Wall (x_{13})	the main materials used on the walls (Zinc:1, Wood:2, Bricks/Conblocks:3, Concrete:4, Glass/Aluminum:5)
	Floor (x_{14})	the main materials used on the floor (Cement:1, PC/Board Tiles:2, Terrazzo:3, Ceramic:4 or Marble:5)
	Ceiling (x_{15})	materials used on the ceiling of the building (none:1, plywood/asbestos:2, acoustic/teak:3)
	Electrical power (x_{16})	installed electrical power (WATT)
	Year of construction (x_{17})	Year of construction
	Year of renovation (x_{18})	Last year of building construction
Legality	Ownership (x_{19})	the type of house ownership (land certificate:1, building use rights:2 or property rights:3)

TABLE II. HOUSE DATA

Year	road width	road function	school_ distance	hospital_ distance	shop_ distance	land_ area	building_ area	...
2017	4	Local	320	788	610	418	54	...
2017	7	Local	40	470	435	3629	112	...
2017	4	Local	100	225	130	120	36	...
2017	2	Environment	40	725	230	170	55	...
2017	4	Local	320	775	130	896	96	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
2019	3	Environment	170	875	130	198	36	...

ownership	number_ floor	building_ condition	construction	roof	wall	...
Right of ownership	1	Moderate	Brick	Ordinary Tiles/Shingle	Wood	...
Right of ownership	1	Moderate	Brick	Zinc	Zinc	...
Right of ownership	1	Moderate	Brick	Zinc	Brick/ Conblocks	...
Right of ownership	1	Very good	Concrete	Zinc	Brick/ Conblocks	
Land certificate	1	Good	Brick	Zinc	Brick/ Conblocks	...
⋮	⋮	⋮	⋮	⋮	⋮	...
Right of ownership	1	Good	Brick	Zinc	Brick/ Conblocks	...

floor	ceiling	electrical power	year construction	year renovation	Price (million rupiah)
PC/Board Tiles	Plywood/ Asbestos/Bamboo	450	1994	0	250.00
PC/Board Tiles	Plywood/ Asbestos/Bamboo	1300	2003	0	900.00
Terrazzo	Plywood/ Asbestos/Bamboo	900	2000	0	275.00
Ceramic	Plywood/ Asbestos/Bamboo	1300	2017	0	230.00
Ceramic	None	1300	2015	0	151.20
⋮	⋮	⋮	⋮	⋮	⋮
Cement	Plywood/ Asbestos/Bamboo	450	1998	0	100.00

B. Preprocessing

Preprocessing is the stage of preparing data to be ready for analysis. The steps taken include data cleaning, data transformation, and outlier detection. This stage is processed using the help of the R-Studio software version 2021.09.2

1) Data Cleaning

Data cleaning in this study is to overcome missing or invalid data. It begins with searching for the data by displaying a summary, as shown in fig. 2.

```
> summary(house)
      Year        road_width    road_function      school_distance
 Min.   :2017   Min.   :1.000   Length:941         Min.   :  40.0
 1st Qu.:2018   1st Qu.:3.000   Class :character   1st Qu.: 270.0
 Median :2019   Median :3.000   Mode  :character   Median : 420.0
 Mean   :2018   Mean   :2.998                      Mean   : 522.3
 3rd Qu.:2019   3rd Qu.:3.000                      3rd Qu.: 620.0
 Max.   :2019   Max.   :7.000                      Max.   :1820.0
                                                   NA's   :40
 hospital_distance shop_distance     land_area      building_area
 Min.   :  45      Min.   :  30.0   Min.   :  51.0   Min.   : 15.00
 1st Qu.:1025      1st Qu.: 180.0   1st Qu.: 150.0   1st Qu.: 36.00
 Median :2025      Median : 230.0   Median : 177.0   Median : 36.00
 Mean   :2058      Mean   : 420.5   Mean   : 283.2   Mean   : 60.11
 3rd Qu.:2525      3rd Qu.: 530.0   3rd Qu.: 258.0   3rd Qu.: 64.00
 Max.   :5025      Max.   :2030.0   Max.   :6740.0   Max.   :610.00
 NA's   :40        NA's   :40
  ownership         number_floor    building_condition construction
 Length:941         Min.   :1.00    Length:941         Length:941
 Class :character   1st Qu.:1.00    Class :character   Class :character
 Mode  :character   Median :1.00    Mode  :character   Mode  :character
                    Mean   :1.07
                    3rd Qu.:1.00
                    Max.   :3.00
    roof              wall              floor             ceiling
 Length:941         Length:941         Length:941         Length:941
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character

 electrical_power year_construction year_renovation      price
 Min.   : 450     Min.   :1964      Min.   :   0.0   Min.   :  48.5
 1st Qu.: 900     1st Qu.:2007      1st Qu.:   0.0   1st Qu.: 110.0
 Median :1300     Median :2016      Median :   0.0   Median : 135.0
 Mean   :1149     Mean   :2011      Mean   : 119.8   Mean   : 164.2
 3rd Qu.:1300     3rd Qu.:2018      3rd Qu.:   0.0   3rd Qu.: 160.0
 Max.   :6600     Max.   :2019      Max.   :2019.0   Max.   :2500.0
```

Fig. 2. Summary data

From the data summary, three data variables are empty (null) in the numeric data. It will be filled with the missing data variables average value (mean). Equation (1) is used to find the average value.

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad (1)$$

where

$\bar{x}$ = mean

x_i = i[th] sample value

m = number of samples

2) Data Transform

At this stage, the transformation process will be carried out for categorical data using the label encoder technique, namely the variables road_function, ownership, building_condition, construction, roof, wall, floor, and ceiling. As in the road_function variable, if it is in the form of an "environment" road, it will be transformed into 1, "local" into 2, "collector" into 3, and "artery" into 4. Overall, this process is shown in Table III

TABLE III. CATEGORIAL VARIABLE ENCODING

Variable	Description	Code
Road function	Environment	1
	Local	2
	Collector	3
	Artery	4
Ownership	Land Certificate	1
	Building Use Rights	2
	Right of Ownership	3
Building condition	Poor	1
	Moderate	2
	Good	3
	Very good	4
Construction	Wood	1
	Brick	2
	Concrete	3
	Steel	4
Roof	Zinc	1
	Asbestos	2
	Ordinary Tiles/ Shingle	3
	Concrete tiles/Aluminum	4
	Decrabon/Concrete/Glaze Tiles	5
Wall	Zinc	1
	Wood	2
	Brick/Conblocks	3
	Concrete	4
	Glass/Aluminum	5
Floor	Cement	1
	PC/Board Tiles	2
	Terrazzo	3
	Ceramic	4
	Marble	5
Ceiling	None	1
	Plywood/Asbestos/Bamboo	2
	Acoustic/Teak	3

3) Outlier

Outliers are data that do not follow the overall data pattern. One method to find outliers is the standardized residual. The outlier detection method uses this Standardized Residual by checking the residuals. The i[th] residual can be seen in (2).

$$e_i = y_i - \hat{y}_i \qquad (2)$$

Where e_i (i[th] residual) is the result of subtracting from y_i (i[th] actual data) to $\hat{y}_i$ (i[th] prediction). According to the i[th] residual above, the i[th] standardized residual can be defined in (3).

$$e_{is} = \frac{e_i}{\sqrt{\frac{\sum_{i=1}^{m}(e_i)^2}{m}}} \quad (3)$$

where :
e_{is} = i[th] *standardized residual*
e_i = i[th] *residual*
m = number of data

According to Singgih Santoso, the data are considered an outlier if the standardized residual value is more than 2.5 or less than -2.5.[17]

```
> standard_res <- rstandard(lm_house_02)
> standard_res[which(abs(standard_res)>2.5)]
       1        25        82       145       237       290
3.310314 16.025549 15.846582  3.056517  6.299022  6.156525
     352       432       502       552       680       688
-2.691840 -2.524019 -2.502360 -2.762601 -2.812176 -2.791015
     738       794       842       852       886
-3.043940 -3.093132 -3.215069 -3.280489 -3.979454
```

Fig. 3. Outlier Detection Using Standardized Residual

Based on the calculation of standardized residuals on the house data with R shown in Fig. 3. The green rectangle displays the results of calculating the standardized residual where the value varies from the smallest -3.97 to the most significant 16.02. the red rectangle indicates the location. These outlier data are located at data 1, 25 82, 145, 237, 290, 352, 432, 502, 552, 680, 688, 738, 794, 842, 852, and 886. The total number of data considered outliers is 17. These data are then deleted so that the previous number of data was 941 lines after the detection process was carried out. Outlier, the amount of data used for the next stage is 924.

4) Split Data

The amount of training data used in predictive analysis affects the level of accuracy. The greater the training data can increase the accuracy[18], [19]. In this study, the training data and test data will be divided using the 90:10 scheme. Where 90% is training data and 10% is test data. The distribution of training and test data is done randomly.

C. Feature Selection Scheme

The feature selection process is used in determining the independent variables that are relevant to the dependent variable, which will be a reference in the formation of the model [20]. Feature selection in this study uses three feature schemes. The first scheme uses all features, the second scheme uses backward elimination, and the third scheme uses forward selection.

1) Backward Elimination

Backward Elimination is a method to optimize a model by eliminating insignificant variables. The process starts by using all the variables then the variables that have a low correlation are eliminated/excluded from the model. This process continues until a model is found that only contains variables that significantly affect the dependent variable.[21], [22]

2) Forward Selection

Forward Selection is the opposite of the Backward Elimination method, where features or independent variables are selected in stages. The independent variables are entered into the model in stages, starting with the variable with the highest correlation to the positive and negative dependent variables. This process continues until the new independent variable cannot increase the effect on the dependent variable.

That's why the forward selection procedure is one of the best model selection procedures in regression by eliminating the independent variables that build the model in stages [23].

D. Prediction Model

1) Multiple Linear Regression

Multiple Linear Regression aims to model the relationship between several independent variables (predictors/ x_1, x_2, …, x_n) and the dependent variable (response/y).

The form of the multiple linear regression equation can be seen in (4)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon. \quad (4)$$

Notes:
Y = dependent variable
$X_1, X_2, \ldots, X_n$ = independent variable
$\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ = regression coefficient
ε = error in the model

TABLE IV. MULTIPLE LINEAR REGRESSION DATA TABLE

i	y	x_1	x_2	…	x_n
1	y_1	x_{11}	x_{12}	…	x_{1n}
2	y_2	x_{21}	x_{22}	…	x_{2n}
3	y_3	x_{31}	x_{32}	…	x_{3n}
m	y_m	x_{m1}	x_{m2}	…	x_{mn}

Equation (4) can be rewritten as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_0 x_{i2} + \cdots + \beta_n x_{in} \quad (5)$$

If the number of regression equation y is m as in (6)

$$\begin{bmatrix} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_n x_{1n} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_n x_{2n} + \varepsilon_2 \\ \vdots \\ y_m = \beta_0 + \beta_1 x_{m1} + \beta_2 x_{m2} + \ldots + \beta_n x_{mn} + \varepsilon_m \end{bmatrix} \quad (6)$$

In the form of a matrix in (7).

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix},$$
$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix} \quad (7)$$

In matrix notation, it can be written as $Y = X\beta + \varepsilon$

Where obtain the value of the regression coefficient (β) by minimizing the number of error squared, it is known as OLS (*Ordinary Least Square*), shown by the (5)

$$\beta = (X^T X)^{-1} X^T Y \quad (8)$$

Notes:

β = the estimated parameter vector in the form of a matrix of (n+1) x 1 size

X = independent variable matrix of m x (n+1) size
Y = the observation vector of the dependent variable of m x 1 size

2) K-Nearest Neighbor

K-Nearest Neighbor (KNN) has the ability to perform classification and regression. The principle works by using the shortest distance between the sample and test data. This is similar to real estate agents estimating the value of a house by comparing previous home buying and selling transactions with the same characteristics. Where the KNN algorithm calculates the shortest distance between the target and the dataset case. The measurement of the distance between the test data and the training data uses the Euclidean distance [24], shown by (8)

$$D(\bar{x},\bar{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (8)$$

where:

$D(\bar{x},\bar{y})$ = distance between x data and y data
x_i = training data
y_i = testing data
n = number of features or variables

3) Evaluation Model

At this stage, performance measurement will be performed on each combination of the feature selection scheme and prediction model. This study uses Root Mean Square Error (RMSE)[25], Coefficient of Determination (R2), and Mean Absolute Percentage Error (MAPE) to measure model performance. The smallest RMSE value will indicate the best model and is defined in (9). R2 is used to assess the effect of the independent variables used in the best model on the "price" variable and is mathematically shown in (10). MAPE is the average value of the absolute percentage error of the actual value against the predicted value to assess the eligibility criteria of the best prediction model shown in (11)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m}} \quad (9)$$

$$R^2 = \left[1 - \left(\frac{\sum_i^m e_i^2}{\sum_{i=1}^m y_i^2 - \frac{(\sum_{i=1}^m y_i)^2}{m}}\right)\right] \quad (10)$$

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \quad (11)$$

III. Results and Discussion

A. Feature Selection Result

The feature selection process uses three schemes, where the first scheme uses all independent variables (19 variables). The second scheme using the backward elimination method produces 16 independent variables and eliminates 3 other variables (roof, floor, and electrical_power). The third scheme using the forward selection method produces 14 independent variables and does not include 5 independent variables (electrical_power, roof, floor, hospital_distance, and shop_distance) in the model formation. As shown in Table V.

Table V. Feature Selection Scheme

Feature Selection Method	Number of features	Description
All Feature	19	year, road_width, road_function, school_distance, hospital_distance, shop_distance, land_area, building_area, ownership, number_floor, building_condition, construction, roof, wall, floor, ceiling, electrical_power, year_construction, year_renovation
Backward Elimination	16	Year, road_width, road_function, school_distance, hospital_distance, shop_distance, land_area, building_area, ownership, number_floor, building_condition, construction, wall, ceiling, year_construction, year_renovation
Forward selection	14	Building_area, road_function, number_floor, land_area, road_width, year_renovation, wall, year_construction, ownership, building_condition, year, construction, school_distance, ceiling

B. Prediction Model Result

Based on the feature selection process in Table V, a prediction model will be formed using the Multiple Linear Regression and KNN methods. Formation of this model using training data. The model produced by Multiple Linear Regression using the all feature scheme can be seen in (12), backward elimination (13), and forward selection (14)

$$\begin{aligned}\hat{y}_{All} = {} & -11293{,}62 + 7{,}01x_1 + 34{,}69x_2 + 34{,}69x_3 \\ & - 0{,}015x_4 - 0{,}0032x_5 + 0{,}0087x_6 + 0{,}056x_7 \\ & + 0{,}79x_8 + 11{,}33x_9 + 84{,}59x_{10} + 10{,}52x_{11} \\ & + 12{,}42x_{12} + 1{,}55x_{13} + 11{,}28x_{14} - 0{,}703x_{15} \\ & + 10{,}893x_{16} + 0{,}008x_{17} - 1{,}57x_{18} - 0{,}029x_{19}\end{aligned} \quad (12)$$

$$\begin{aligned}\hat{y}_{BE} = {} & -11308{,}75 + 6.980511x_1 + 34{,}66323x_2 \\ & + 69{,}85511x_3 - 0{,}01603805x_4 - 0{,}003056098x_5 \\ & + 0{,}008703363x_6 + 0{,}05559184x_7 + 0{,}8111657x_8 \\ & + 11{,}26369x_9 + 85{,}85475x_{10} + 10{,}51984x_{11} \\ & + 13{,}09488x_{12} + 11{,}88260x_{14} + 10{,}74023x_{16} \\ & - 1.538618x_{18} - 0{,}02932039x_{19}\end{aligned} \quad (13)$$

$$\begin{aligned}\hat{y}_{FS} = {} & -12476{,}89 + 0{,}8055994x_8 + 71{,}36800x_3 \\ & + 85{,}42456x_{10} + 0{,}05529192x_7 + 33{,}55137x_2 \\ & - 0{,}02935082x_{19} + 12{,}07665x_{14} - 1.571539x_{18} \\ & + 10{,}71641x_9 + 11{,}48274x_{11} + 7.591164x_1 \\ & + 12{,}27210x_{12} - 0{,}01389931\,x_4 + 11{,}00202x_{16}\end{aligned} \quad (14)$$

Notes:

$\hat{y}_{All}$ = The Multiple Linear Regression prediction model uses all features.
$\hat{y}_{BE}$ = The Multiple Linear Regression prediction model uses backward elimination
$\hat{y}_{FS}$ = The Multiple Linear Regression prediction model uses forward selection.

The resulting model is then applied to make predictions using test data. Furthermore, performance measurement is carried out on the model.

TABLE VI. MULTIPLE LINEAR REGRESSION MODEL PERFORMANCE MEASUREMENT TABLE

Method	Feature Selection	RMSE (million rupiahs)	R^2
Multiple Linear Regression	All Features	44.87	0.696
	Backward elimination	44.02	**0.707**
	Forward selection	44.47	0.701
Average		44.53	0.701

Table VI shows that the prediction model using Multiple Linear Regression produces the smallest RMSE value in the backward elimination feature selection scheme of 44.02 (million rupiahs) compared to the forward selection scheme and all features, each with a value of 44.47 (million rupiahs). and 44.53 (million rupiah). This also applies to the R2 value, backward elimination produces the largest R2 value of 0.707, followed by forward selection and all-feature schemes of 0.701 and 0.696.

TABLE VII. K VALUE RESULT

K	RMSE (million rupiahs)		
	All feature	**Backward Elimination**	**Forward Selection**
1	69.24474	74.72764	72.51015
2	61.41302	63.41222	62.26008
3	59.16537	60.73112	60.26455
4	58.57380	57.60658	57.83208
5	57.38664	56.31964	56.29308
6	57.41488	56.02238	56.28193
7	57.50303	55.87130	55.53427
8	57.30330	55.88698	55.67774
9	57.51970	56.22013	56.32706
10	58.20345	56.65249	56.38003

Table VII displays the KNN model for various feature schemes, using k based on the smallest RMSE value. Conduct training by entering k values from 1 to 10 in each feature scheme. In the scheme of all features, the best model has an RMSE value of 57.303, and the number of nearest neighbors (k) is 8. The best backward elimination model feature scheme is based on the smallest RMSE value of 55.871 with the nearest neighbors (k) value of 7, and in the forward selection feature scheme, the number nearest neighbor (k) is 7. This value of k will be used as the basis for determining the predicted value of the test data based on the average value of the k closest training data samples.

Table VIII displays the results of the prediction model performance measurement using the KNN algorithm. The smallest RMSE value is generated by a combination with forward selection, which is 47.20 (million rupiahs), followed by backward elimination of 48.84 (million rupiahs) and in the all-features scheme of 50.67 (million rupiahs). The most significant R2 value is generated by forward selection of 0.664, followed by backward elimination of 0.640, and the smallest is the all-feature scheme of 0.612.

TABLE VIII. KNN MODEL PERFORMANCE MEASUREMENT TABLE

Method	Feature Selection	RMSE (million rupiahs)	R^2
KNN	All Feature	50.67	0.612
	Backward elimination	48.84	0.640
	Forward selection	47.20	0.664
Average		48.90	0.638

Besides reducing the computational dimensions by filtering out variables that have less effect on house prices (Table IV), the feature selection process using backward elimination and forward selection can also improve the performance of the resulting model. Tables V and VII show that the RMSE and R2 values generated using backward elimination and forward selection are better than using all features.

A prediction model has a remarkably accurate performance if the MAPE value is below 10%. The interpretation of MAPE values are shown in Table IX MAPE value interpretation.[26]

TABLE IX. MAPE VALUE INTERPRETATION

MAPE (%)	Interpretation
<10	Very accurate prediction
10-20	Good prediction
20-50	Decent prediction
>50	Inaccurate prediction

Prediction results using linear regression with the feature selection of backward elimination are included in the decent prediction interpretation. It is shown in Table X that the MAPE value of the model is 20.46%. The resulting model is feasible to use as a basis for predicting house/property prices.

TABLE X. ERROR PREDICTION CALCULATION WITH MAPE

Method	**Feature Selection**	**RMSE (million rupiahs)**	**MAPE**
Multiple Linear Regression	*Backward Elimination*	44,02	20,46%

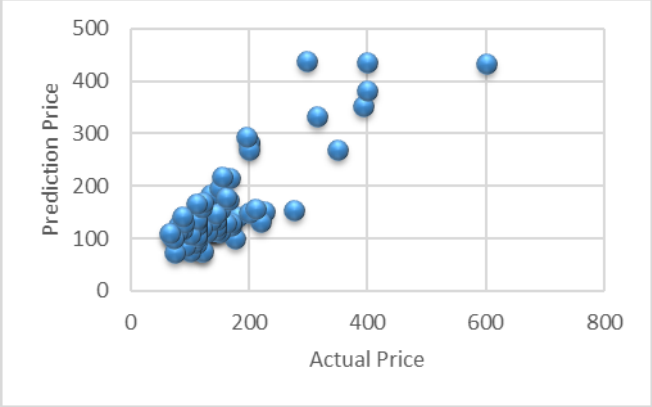

Fig. 4. House price prediction and the actual price

IV. Conclusion

In general, the Multiple Linear Regression algorithm on various feature schemes produces a better model performance than KNN in implementing house price predictions in the Sintang district. A backward elimination combination in Multiple Linear Regression produces the best model with the smallest Root Mean Square Error (RMSE) value of 44.02 (million rupiahs). The R2 value of 0.707 in the model shows that 70.7% of house prices in the Sintang sub-district are influenced by the variables year, road_width, road_function, school_distance, hospital_distance, whop_distance, land_area, building_area, ownership, number_floor, building_condition, construction, wall, ceiling and year_renovation, while 29.3% is influenced by other variables outside this study. Based on the error rate using MAPE of 20.46%, it is interpreted that the model produces feasible or even close to good predictions.

The use of backward elimination and forward selection as feature selection in this data, in addition to reducing the number of variables used, can also improve the prediction model's performance.

Further research can use other predictive methods and pay attention to other variables affecting house prices.

Acknowledgment

The author would like to thank the Human Resources Research and Development Agency of the Ministry of Communication and Information of the Republic of Indonesia, which has provided funds for the Domestic Master's Scholarship Program, and thank the Sintang District Government and my supervisor.

References

[1] UU RI, "Undang-Undang Nomor 28 Tahun 2009 Tentang Pajak Daerah dan Retribusi Daerah." Indonesia, 2009.

[2] Badan Pusat Statistik, "Kabupaten Sintang Dalam Angka 2020," 2020. doi: 10.22146/mgi.34838.

[3] Pemerintah Kabupaten Sintang, *Peraturan Daerah No 1 Tahun 2011 Tentang Bea Perolehan Hak Atas Tanah dan Bangunan*. Sintang, 2011.

[4] T. Afriyandi, "Kewenangan Pemerintah Daerah Dalam Menentukan Harga Jual Dalam Transaksi Jual Beli Tanah Dan Atau Bangunan," *J. Huk. Volkgeist*, vol. 3, no. 1, pp. 29–43, Dec. 2018, doi: 10.35326/volkgeist.v3i1.112.

[5] D. G. Raja Guk-Guk, I. Isnaini, and M. C. Ramadhan, "Efektifitas Validasi Bea Perolehan Hak atas Tanah dan Bangunan terhadap Ketidaksesuaian Nilai Objek Pajak dalam Akta Jual Beli dengan Harga Sebenarnya," *J. Educ. Hum. Soc. Sci.*, vol. 4, no. 2, pp. 875–885, Oct. 2021, doi: 10.34007/jehss.v4i2.763.

[6] D. Erlinda, F. Wisnaeni, and N. Maharani Sukma, "Pelaksanaan Verifikasi Peralihan Hak Atas Tanah Dan Bangunan (Bphtb) Di Kabupaten Bogor," *Notarius*, vol. 13, no. 2, pp. 946–960, 2020, doi: 10.14710/nts.v13i2.31305.

[7] M. D. Mankad, "Comparing OLS based hedonic model and ANN in house price estimation using relative location," *Spat. Inf. Res.*, vol. 30, no. 1, pp. 107–116, Feb. 2022, doi: 10.1007/s41324-021-00416-3.

[8] R. B. Abidoye and A. P. C. Chan, "Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network," *Pacific Rim Prop. Res. J.*, vol. 24, no. 1, 2018, doi: 10.1080/14445921.2018.1436306.

[9] G. Hayrullahoglu, Y. Aliefendioglu, H. Tanrivermis, and A. C. Hayrullahoglu, "Estimation of the Hedonic Valuation Model in Housing Markets: The Case of Cukurambar Region in Cankaya District of Ankara Province," *Ecoforum*, vol. 7, no. 1, 2018.

[10] M. D. Mankad, "Comparing OLS based hedonic model and ANN in house price estimation using relative location," *Spat. Inf. Res.*, 2021, doi: 10.1007/s41324-021-00416-3.

[11] M. Doszyń, "Might expert knowledge improve econometric real estate mass appraisal?," *J. Real Estate Financ. Econ.*, 2022, doi: 10.1007/s11146-022-09891-3.

[12] W. Zhao, C. Sun, and J. Wang, "The research on price prediction of second-hand houses based on KNN and stimulated annealing algorithm," *Int. J. Smart Home*, vol. 8, no. 2, 2014, doi: 10.14257/ijsh.2014.8.2.19.

[13] I. Wahyuni, N. Nafi'iyah, and Masruroh, "Sistem Peramalan Penjualan Perumahan di Kabupaten Lamongan dengan Menggunakan Metode Regresi Linier Berganda," *Semin. Nas. Sist. Inf. 2019*, no. September, pp. 1969–1973, 2019.

[14] N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–5, 2018, doi: 10.1109/ICCUBEA.2018.8697639.

[15] H. Feng *et al.*, "Estimation of forest aboveground biomass by using mixed-effects model," *Int. J. Remote Sens.*, vol. 42, no. 22, pp. 8675–8690, 2021, doi: 10.1080/01431161.2021.1984611.

[16] I. Kurniawan and A. F. Abror, "Komparasi Metode Kombinasi Seleksi Fitur dan Machine Learning K-Nearest Neighbor pada Dataset Label Hours Software Effort Estimation," *Explor. J. Sist. Inf. dan Telemat.*, vol. 10, no. 2, Oct. 2019, doi: 10.36448/jsit.v10i2.1314.

[17] S. Santoso, *Mastering SPSS 18*. Jakarta: PT Elex Media Komputindo, 2010.

[18] D. Werdiastu, D. E. Ratnawati, and B. Rahayudi, "Estimasi Hasil Produksi Benih Berdasarkan Karakteristik Tanaman Kenaf Menggunakan Metode Backpropagation (Studi Kasus: Balai Tanaman Pemanis dan Serat Kota Malang)," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[19] B. Gunawan, H. Sasty, P. #2, E. Esyudha, and P. #3, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," vol. 4, no. 2, pp. 17–29, 2018.

[20] A. Shoddiq Bayu Asmoro, W. Sakti Gunawan Irianto, U. Pujianto, J. Semarang No, and J. Timur, "Perbandingan Kinerja Hasil Seleksi Fitur pada Prediksi Kinerja Akademik Siswa Berbasis Pohon Keputusan," *JEPIN (Jurnal Edukasi dan Penelit. Inform.* , vol. 4, no. 2, pp. 84–89, Dec. 2018, [Online]. Available: www.kaggle.com/aljarah/xAPI-Edu-Data

[21] A. Bode, "Perbandingan Metode Prediksi Support Vector Machine dan Linear Regression Menggunakan Backward elimination pada Produksi Minyak Kelapa," *Simtek J. Sist. Inf. dan Tek. Komput.*, vol. 4, no. 2, pp. 104–107, Oct. 2019, doi: 10.51876/simtek.v4i2.57.

[22] S. Haloui and A. E. El Moudden, "An optimal prediction model's credit risk: The implementation of the backward elimination and forward regression method," *Int. J. Adv. Comput. Sci. Appl.*, no. 2, pp. 457–467, 2020, doi: 10.14569/ijacsa.2020.0110259.

[23] R. Sanjaya and F. Fitriyani, "Prediksi Bedah Toraks Menggunakan Seleksi Fitur Forward Selection dan K-Nearest Neighbor," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, 2019, doi: 10.26418/jp.v5i3.35324.

[24] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012. doi: 10.1016/C2009-0-61819-5.

[25] C. Mishra, L. Mohanty, S. Rath, R. Patnaik, and R. Pradhan, "Application of backward elimination in multiple linear regression model for prediction of stock index," in *Smart Innovation, Systems and Technologies*, 2021, vol. 153. doi: 10.1007/978-981-15-6202-0_56.

[26] N. Koesoemaningroem, Endroyono, and S. Mardi Susiki Nugroho, "Peramalan Pencematan Udara di Kota Surabaya Menggunakan Metode DsArima dengan Pendekatan Percentile Error Bootstrap (PEB)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 5, pp. 987–994, 2021, doi: 10.25126/jtiik.202185216.