LAND VALUE PREDICTION MODEL IN THE URBAN FRINGE REGION USING MULTIPLE REGRESSION ALGORITHM

Ita Dwi Kisworini

[1]National Land Agency of East Java Province, Indonesia
E-mail: ita.dkisworini@gmail.com

Received: July 19, 2022 | **Accepted:** July 23, 2022 | **Published:** July 25, 2022
DOI: 10.12962/j27745449.v2i3.442 | **Issue:** Volume 2 Number 3 2021 | **E-ISSN:** 2774-5449

ABSTRACT

The existence of land has become a dream. The value continues to soar over the years, making the land become something that various parties highly desire. This phenomenon causes the demand for land also to continue to increase. This study aims to predict land values based on actual land prices in the field in the form of transaction results and bid prices with a statistical approach for mass valuation in the urban fringe area, where this area is considered an alternative area to overcome the need for housing which is getting higher and higher. Land value is estimated using a data comparison approach market and added to the cost of building construction. Multiple regression is one of the methods to get the best estimation model from big data. The magnitude of the contribution of the influence of the variables making up the Land Value can be known through the coefficient of determination (R2), which is 82.5%. Many factors affect the Land Value, such as Bid Price, Physical Land Factors, Environmental Factors, and Legal Aspects.

Keyword: Land value, building value, regression, urban fringe, prediction

Introduction

Land is a fundamental element that is needed by various living things. In addition to meeting the needs of life, the land also functions as a place to live and a place to earn a living. The existence of land is something that is very much dreamed of, whose value continues to soar from year to year, which makes it the prima donna that various parties covet. This fact causes the demand for land to continue to increase [1]. The statistical approach to mass valuation allows the valuation of prices using sampling data and standard methods to predict property prices without sampling additional variables [9]. Computer-assisted mass assessment has become popular as an automated scoring model [2]. Each predictive modeling technique has advantages and disadvantages, with prediction accuracy considered a fundamental component. Therefore, each technique should be measured using standard predictive performance measures and compared with overall prediction accuracy [2].

This paper presents predictions of land values in the Urban Fringe area, also often known as suburban areas. This area usually requires special attention from the local government because of the area's importance to the living conditions of the community, both rural and urban residents in the future. [3][8]. The rural- urban periphery is a dynamic and rapidly changing environment [4][8]. Many people consider the city's outskirts an alternative area to overcome the need for housing whose prices are increasing [8].

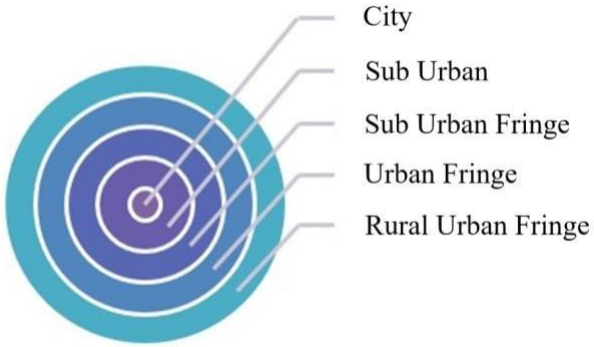

Figure 1. Urban area scheme

The purpose of this paper is to predict land values based on data that has been owned, especially for the Urban Fringe Area [5] as a prima donna when property prices in the city center are getting out of control. The data used in this study is based on actual land prices in the field in the form of transaction results (market value) in 2021 and bid prices. In this paper, we

investigate the land value prediction model, one of which is by using multiple regression equations.

Regression analysis is carried out when the relationship between two variables is a causal relationship (cause- effect) or functional. This model requires an instrument in the form of several variables that will be used to predict land values for an area. In this case, the market/transaction value is played as the dependent variable (Y), and the determining factor of land value functions as the independent variable (X) [9]. The factors used in this assessment includes: (a) soil physical variables (soil shape, front width, land area, elevation of the road and ground), (b) environmental variables (road class, accessibility, drainage, utilities, facilities, area), and (c) land ownership status [10].

Methodology

The rapid development in Gununganyar District is due to the high demand for housing [6]. The development and conversion of land in Gununganyar at this time, which was initially in the form of rice fields, ponds, and vacant land, has now turned into a residential area for housing and apartments. This development and conversion of this land impact property prices in Gununganyar District, which also increase and become the target of investors.

In previous studies, the value of NJOP (Tax Object Sales Value) was used as the primary data to predict land values. Land value prediction is based on NJOP building price, land area, and land price [13]. However, in this study, data collection was carried out by collecting data/samples in the field, using land valuation methods such as market data comparison [7], cost approach, and the income approach. The method used in this research is using a combination of land valuation methods according to the conditions in the field, namely the comparison of market data and the cost approach.

Comparison of market data is carried out by field surveys to obtain physical data and legal aspects of land or property that are the object of careful assessment, as well as to obtain market data for comparison of objects that have physical characteristics [12], legal [10] and similar environmental or that can be compared to the object being assessed. The data from the survey are adjusted to obtain a market value estimation of the object being assessed.

Besides using the market data comparison, the cost approach method is also used to determine property values (land and buildings). The value of the land is estimated using a market data comparison approach and is added to the calculation of the building cost. The cost approach can also be used to estimate the land value of a property, provided that the property value is known. Data regarding the materials and volumes used in building construction are also required to be known. For buildings that are not new, the cost approach considers an estimate of depreciation (depreciation) that includes parts experiencing physical deterioration or functional deterioration.

The data/samples in this study are registered land parcels/indigenous land, which provide information on the transaction price or offer of the land parcel. The sample was selected based on considerations of the characteristics of the village or urban village, proportionally to residential, commercial, and agricultural land use. The selected sample is attempted in the form of an empty plot of land that refers to the base map used as the existing work map.

In searching for samples, the selection of respondents must be appropriate. Respondents are the primary data source who can provide reliable descriptions and information about transaction price information or bid prices for both buying and selling or leasing land [13]. Respondents who can be selected include landowners who have just made a transaction, landowners who intend to sell/lease their land, real estate agents/brokers, developers, and tenants of land or property parcels. Other respondents who can be selected include notaries, village heads, or other officials who are believed to be reliable sources of land market price information if all the required respondents are unavailable.

Result and Discussion

Before analysis with multiple linear deviations, data filtering must be done to clean up missing data. The data is cleaned by eliminating observations indicated by outliers to anticipate the indications of outliers [9][11], namely by looking at the residual value that must be between -2.0 and 2.0 (this value is based on the z score at alpha 5 %, i.e., 1.96 ≈ 2.0).

The estimation results of the regression model and the influence of the independent variable on the dependent variable can be seen in the table 2.

Figure 2. Research site

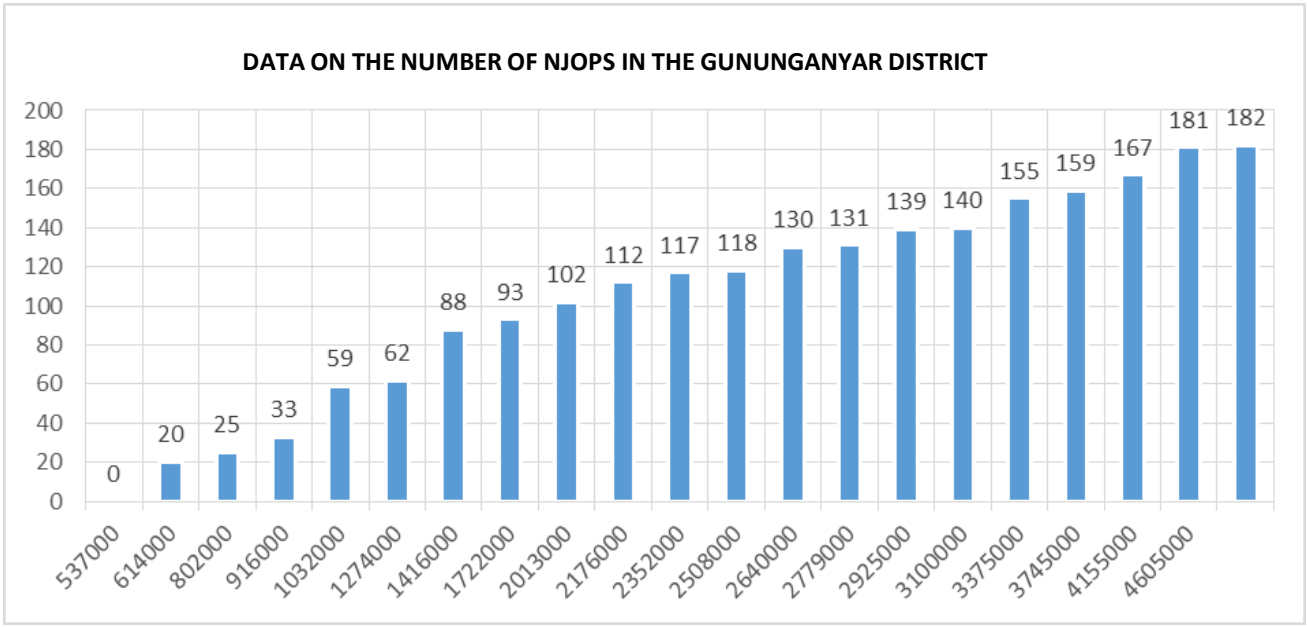

Figure 3. NJOP value at the research site

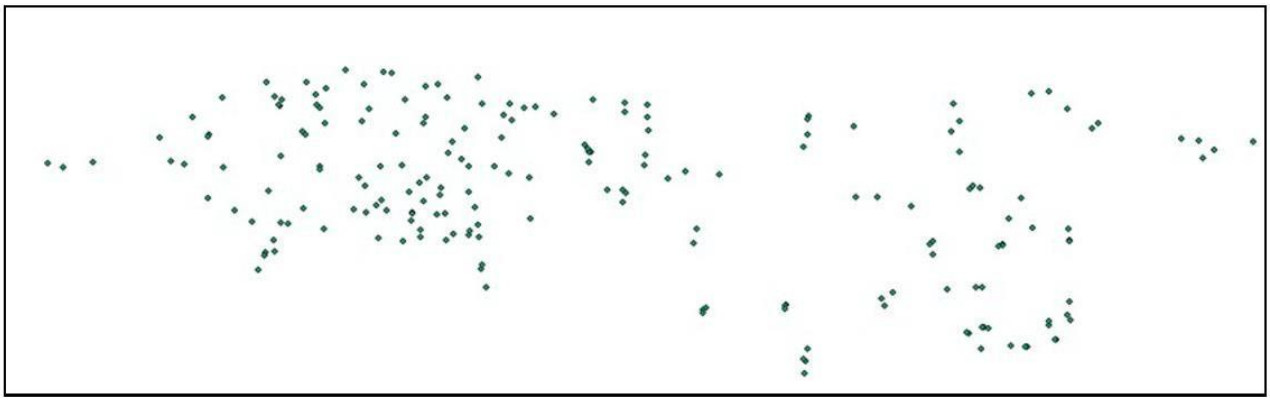

Figure 4. Distribution of the sample in the location area

Table 1. Table of variables making up land values

No	Variable	Description	No	Variable	Description	No	Variable	Description
1	X1	Ownership Status	7	X7	Elevation from street	13	X13	Utility
2	X2	Market Price	8	X8	Land Location	14	X14	Facility
3	X3	Surface Area	9	X9	Road Hierarchy	15	X15	Building Area
4	X4	Front Width	10	X10	Road Widht	16	X16	Number of Floors
5	X5	Back Length	11	X11	Accessibility	17	X17	Building Value
6	X6	Ground Shape	12	X12	Drainage			

Table 2. Regression model estimation

```
> summary(m)

Call:
lm(formula = Y ~ . - `No Urut Entry`, data = clean_df)

Residuals:
     Min       1Q   Median       3Q      Max
-2731017  -654046   -61113   713239  1797565

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.430e+05  1.684e+06  -0.560  0.57635
X1           4.121e+05  2.270e+05   1.815  0.07160 .
X2           3.867e-03  2.272e-04  17.021  < 2e-16 ***
X3          -3.764e+04  2.305e+03 -16.333  < 2e-16 ***
X4           2.026e+05  4.632e+04   4.375 2.34e-05 ***
X5           9.255e+04  1.681e+04   5.506 1.67e-07 ***
X6           6.982e+05  2.885e+05   2.420  0.01677 *
X7          -4.815e+05  2.483e+05  -1.939  0.05448 .
X8           6.183e+04  1.169e+05   0.529  0.59763
X9          -6.059e+04  3.301e+05  -0.184  0.85464
X10          3.371e+04  9.479e+04   0.356  0.72260
X11         -9.575e+05  2.183e+05  -4.386 2.24e-05 ***
X12          1.072e+06  3.332e+05   3.217  0.00161 **
X13          2.005e+06  2.781e+05   7.209 3.06e-11 ***
X14         -7.428e+05  1.388e+05  -5.352 3.40e-07 ***
X15          3.461e+04  3.124e+03  11.079  < 2e-16 ***
X16          1.034e+06  2.427e+05   4.261 3.69e-05 ***
X17         -1.431e-02  1.137e-03 -12.587  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 933300 on 142 degrees of freedom
Multiple R-squared:  0.8256,    Adjusted R-squared:  0.8047
F-statistic: 39.55 on 17 and 142 DF,  p-value: < 2.2e-16
```

Based on the estimation results of the multiple regression model above, it shows that there are variables that are not statistically significant to the value (Y), such as the variable X7 (Elevation from the Road), X8 (Soil Location), X9 (Road Class), and X10 (Road Width). It can be seen from the p-value of each of these variables > the level of significance (α=5%). Therefore, these variables need to be eliminated to get the best model for predicting the value of Y.

So that the best regression equation from the estimation results of multiple linear regression analysis is:

$$\hat{Y} = -834301.61 + 442102.73\ \mathbf{X1} + 0.001\ \mathbf{X2} - 38141.22\ \mathbf{X3} + 205356.48\ \mathbf{X4} + 93900.73\ \mathbf{X5} + 706920.59\ \mathbf{X6} - 492437.85\ \mathbf{X7} - 909345.72\ \mathbf{X11} + 1104338.55\ \mathbf{X12} + 1997689.05\ \mathbf{X13} - 739678.60\ \mathbf{X14} + 35008.85\ \mathbf{X15} + 1027414.79\ \mathbf{X16} - 0.01\ \mathbf{X17}$$

Coefficient of Determination

The magnitude of the contribution of the influence of the variables making up the Land Value can be known through the coefficient of determination (R2), which is 0.825 or 82.5%, while the remaining 17.5% is the contribution of other factors/variables not discussed in this study. This means that the Land Value variable can be explained by (a) the Bid Price variable and the physical variables of the land (soil shape, front width, land area, elevation from the road and back length), (b) environmental variables (accessibility, drainage, utilities, facilities, area), and (c) land ownership status. Besides the physical, environmental, and legal aspects above, the variables that must be considered are important building-forming variables (building area, number of floors, and building value).

Hypothesis test

Simultaneous Test (F Test)

The test criteria state that if the p value ≤ level of significance (α=5%) then reject H0, meaning that there is a significant effect simultaneously (together). On other hand, if p value > level of significance ((α=0.05) then accept H0, meaning that there is no significant effect simultaneously (together) Independent Variable (X) to Value (Y).

H0: there is no significant effect simultaneously (together) Independent Variable (X) on Value (Y) H1: there is a significant effect simultaneously (together) the independent variable (X) on the Value (Y).

The results of simultaneous hypothesis testing resulted in a calculated F value of 48.83 with a p value of 2.2e-16 (0.000). The test results show p value (0.000) < level of significance (α= 0.05) then reject H0, this means that there is a simultaneous (together) significant effect of the independent variable (X) on the value (Y).

Partial Test (t Test)

Partial hypothesis testing (t test) is used to determine whether there is a partial (individual) effect of the variables making up the land value on the land value. The test criteria state if the value of tcount > ttable or p value < level of significance (α=0.05) then there is a significant effect partially (individually). On other hand, if value of tcount < ttable or p value > level of significance (α=0.05) then there is no partial significant effect of variable X on the value (Y). Partial test results are described as table 3.

Normality test

The normality assumption test aims to test whether the residuals in the regression analysis model are normally distributed or not. In the regression analysis, the residuals are expected to be normally distributed.

It can be detected through a Probability Plot to test whether the residuals are normally distributed or not. The test criteria state that if the residual points spread around the diagonal line, the residuals are declared to be normally distributed. The following are the results of detecting normality assumptions through the Probability Plot (figure 5).

Based on the probability plot above, it can be seen that the residual points spread around the diagonal line. This means that the residuals are declared to be normally distributed. Thus, the assumption of normality is met.

Table 3. T-test on the variable land value

No	Variables		Value t count	*p value*	(α=0.05)	Description	Regression Coefficient B1	Influence to Land Value
1	X1	Ownership Status	2.002	0.047112	*p value* < α	Significant	442102.73	Positive
2	X2	Bid Price	20.564	2e-16	*p value* < α	Significant	0.001	Positive
3	X3	Surface Area	-19.826	2e-16	*p value* < α	Significant	-38141.22	Negative
4	X4	Front Width	4.595	9.35e-06	*p value* < α	Significant	205356.48	Positive
5	X5	Long Back	5.878	2.74e-08	*p value* < α	Significant	93900.73	Positive
6	X6	Ground Shape	2.495	0.013733	*p value* < α	Significant	706920.59	Positive
7	X7	Elevation from Street	-2.047	0.042463	*p value* < α	Significant	-492437.85	Negative
8	X11	Accessibility	-4.749	4.86e-06	*p value* < α	Significant	-909345.72	Negative
9	X12	Drainage	3.665	0.000347	*p value* < α	Significant	1242371.661	Positive
10	X13	Utility	7.293	1.82e-11	*p value* < α	Significant	1997689.05	Positive
11	X14	Facility	-5.477	1.86e-07	*p value* < α	Significant	-739678.60	Negative
12	X15	Building Area	11.604	2e-16	*p value* < α	Significant	35008.85	Positive
13	X16	Number of floors	4.367	2.38e-05	*p value* < α	Significant	1027414.79	Positive
14	X17	Building Value	-13.290	2e-16	*p value* < α	Significant	-0.01	Negative

It can be detected through a Probability Plot to test whether the residuals are normally distributed or not. The test criteria state that if the residual points spread around the diagonal line, the residuals are declared to be normally distributed. The following are the results of detecting normality assumptions through the Probability Plot (figure 5).

Based on the probability plot above, it can be seen that the residual points spread around the diagonal line. This means that the residuals are declared to be normally distributed. Thus, the assumption of normality is met.

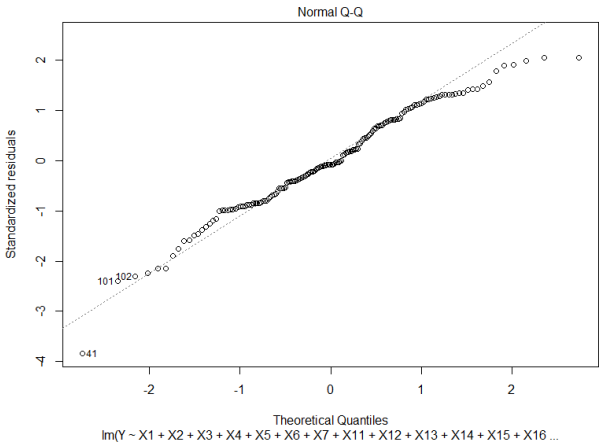

Figure 5. Normality test through probability plot

Heteroscedasticity Test

The assumption of heteroscedasticity is used to determine whether the residuals have a homogeneous variance or not. In testing the assumption of heteroscedasticity, it is expected that the residuals have a homogeneous variance. Testing the assumption of heteroscedasticity can be seen based on the scatter plot. Residuals are said to have a homogeneous variance if the residual points on the scatter plot spread randomly.

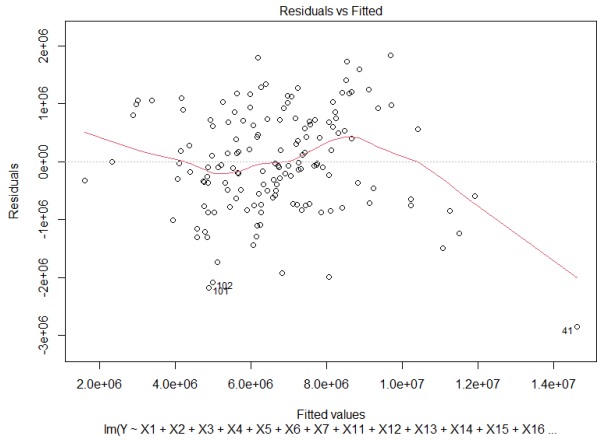

Figure 6. Detection of heteroscedasticity through scatter plot

```
> bptest(m2, studentize=F)

        Breusch-Pagan test

data:  m2
BP = 27.506, df = 14, p-value = 0.01653
```

Figure 7. Pagan's Breusch Test

Based on the scatter plot above, the residual points spread randomly based on the X-axis and Y-axis observations. Thus, it can be concluded that the residuals generated by the regression model have a homogeneous variance [8], so the non-heteroscedasticity assumption is stated to be fulfilled.

In addition, the heteroscedasticity assumption test can be seen with the Breusch Pagan test. The residual is said to have a homogeneous variance if the resulting p value > the level of significance (α = 0.05).

Based on the results above, the resulting p value is 0.01653, where the p value > alpha (0.01), with a 1% significance level, it can be concluded that the heteroscedasticity assumption is fulfilled.

Conclusion

The land price model in the study area consists of 14 independent variables, with influencing factors in the form of Bid Price, Soil Physical Factors (soil shape, front width, land area, elevation from the road and back length), Environmental Factors (accessibility, drainage, utilities), facilities) and Legal Factors (land ownership status). In addition to the bid price variable, physical variable, environmental variable, and legal aspect variable above, the variables that must be considered are the building-forming variables (building area, number of floors, building value).

The study used 197 survey sample points in Gununganyar District and obtained R2 = 82.5%. According to the results of statistical tests, the variables of Road Width, Land Location (Normal, Hook, Skewer), and Road Class (Collector Road, Local Road) have no effect in determining the Land Value.

The mass appraisal helps the state management agency responsible for land offer prices that are close to market prices based on the factors that affect land prices for each area or specific area. Multiple regression analysis answers the problem of predicting land values in certain areas with complex data.

References

[1] Suen, I-Shian. The impact of compact and mixed development on land value: A case study of Richmond, Virginia, *Urban Scie Journal*. (2018).

[2] McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: A comparison of modern approaches, *Journal of PropertyResearch*. **30(4)** (2013) 239– 265.

[3] Teodoro Semeraro, Benedetta Radicchio, Pietro Medagli, Stefano Arzeni, Alessio Turco and Davide Geneletti. Integration of ecosystem services in strategic environmental assessmentof a peri-urban development plan, *Sustainability Journal*. (2021).

[4] Nick Gallent & Dave Shaw. Spatial planning,area action plans and the rural- urban fringe, *Journal of Environmental Planning and Management.* (2013).

[5] Erna Lópeza, Gerardo Boccoa, Manuel Mendozaa, Emilio Duhaub. 2001. Predicting land-cover and land-use change in the urban fringe: A case in Morelia city, Mexico, *Landscape and Urban Planning*. **Volume 55, Issue 4, 10 August** (2001) Pages 271-285.

[6] Monk, S., Pearce, B. J., & Whitehead, C. M. E. Land-use planning, land supply, and house prices, *Environment and Planning A*. **28(3)** (1996) 495-511. DOI: 10.1068/a280495

[7] Pham Thi Ha, Nguyen Tran Tuan, Nguyen Van Quan y, Nguyen Van Trung. Land price regression model and land value region map to support residential land price management: A Study in Nghe an Province, Vietnam. (2022) 71-83.

[8] Broomhall, D. (1995). Urban encroachment, economic growth, and land values in the urban fringe. *Growth and Change*. **26(2)** (1995) 191–203.

[9] Mariano Cordoba, Juan Pablo Carranza, Mario Piumetto, Federico Monzani, Monica Balzarini.A spatially based quantile regression forest model for mapping rural land values. *Journal of Environmental Management*. **289** (2021) 112509.

[10] Jannet C. Bencure, Nitin K. Tripathi, Hiroyuki Miyazaki, Sarawut Ninsawat and Sohee Minsun Kim. Development of an Innovative Land Valuation Model (iLVM) for mass appraisal application in sub-urban areas using AHP: An integration of theoretical and practical approaches. (2019).

[11] J. Avanija, Gurram Sunitha, K. Reddy Madhavi, Padmavathi Kora, and R. Hitesh Sai Vitta. 2021. Research article prediction of house price using XGBoost regression algorithm, *Turkish Journal of Computer and Mathematics Education*. **Vol. 12 No.2** (2021) 2151–21552151.

[12] N T Sugito, I Soemarto, S Hendriatiningsih, B E Leksono. Modified estimation of land values with spatial weight in Bandung city, *International Geography Seminar.* 2019.

[13] Priya P, Arul Kumaran M, Dhinesh Kumar K, Nivas Singha S, Rajkumar K. 2021. Prediction of property price and possibility prediction using machine learning. *Annals of R.S.C.B*. **Vol. 25, Issue 4** (2021) Pages. 3870–3882.