Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, Number 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.7083

Comparison of Two Linear Regression Models for Predicting the Literacy Development Index in Indonesia

Iffatu Wardani[1*], Kunto Jiwandono[2], Okta Dyah Pradanti[3], Yuni Wahyu Winjarwati[4], Stevano Aji Aghashie[5]
[1,2,3,4,5]Institut Teknologi dan Bisnis Trenggalek, Indonesia
[1]iffatu.wardani@gmail.com, [2]kuntojiwandono01@gmail.com, [3]dyahokta937@gmail.com, [4]wyuni4197@gmail.com, [5]stevanoaji123@gmail.com

open access

***Corresponding Author**

Article History:
Submitted: 14-10-2025
Accepted: 23-10-2025
Published: 31-10-2025

Keywords:
Correlation; machine learning; multiple linear regression; simple linear regression

ABSTRACT

This study examines four suspected factors that have correlation and influence Community Literacy Development Index (IPLM). The four factors data was taken from each province in Indonesia i.e. the number of accredited libraries, the level of people's reading interest, proportion of population living below 50% of the median income, and high school completion rate. To determine whether these four factors truly affect IPLM, a regression model analysis was conducted. The machine learning models discussed in this study are simple linear regression and multiple linear regression. One multiple linear regression model was used to integrate all four factors together. Four simple linear regression models were applied to assess each factor individually in relation to IPLM. From all these regression models, the adjusted R-squared values were compared. The analysis revealed that the level of people's reading interest factor has a higher adjusted R-squared value in the simple linear regression (0.3828) compared to the multiple linear regression (0.3235). In contrast, the other three factors show lower adjusted R-squared values in their simple linear regressions than in the multiple linear regression. The conclusion is the reading interest factor best used to predict IPLM without involving the other factors. Meanwhile, the remaining three factors should be used collectively when predicting IPLM values.

INTRODUCTION

According to the National Library of the Republic of Indonesia (2024), the Community Literacy Development Index (IPLM) is a tool to measure the progress and development of library services by local governments. IPLM has seven main components. These component are equal access to library services, sufficient collections, adequate library staff, level of community visits to libraries, libraries that meet NSP standards, community involvement in library promotion activities, and the number of library members. Local governments must work to improve community literacy and IPLM. This is important because literacy skills in Indonesia remain low. The level is even lower compared to neighboring ASEAN countries or international standards like UNESCO and PISA (Ministry of Communication and Digital, 2020; Nasrullah & Asmarini, 2024). Indonesia scores 62% in digital literacy. This is far below the ASEAN average of 70%. According to UNESCO, only 0.001% of Indonesians read regularly. The 2022 PISA study shows similar results. Indonesian students rank sixth in reading literacy among other ASEAN countries.

Given the importance of boosting community literacy, we need predictions to identify key factors that affect the community literacy development index. Knowing these factors can help governments to formulate strategic steps and it can be used as a consideration in the process of increasing public literacy. Higher literacy index can brings positive impacts. These include better quality of life, more active participation in economic, social, and political activities, and reduced poverty. It also helps people respond wisely to information (Kalla Institute, 2024). Some suspected causes of low literacy are: limited access to quality reading materials, uneven education access, weak reading culture, and regional disparities. We can analyze these factors to see their influence on IPLM values. Later, they can serve as predictors for IPLM scores.

There are many methods to build prediction. One uses machine learning regression algorithms (Kusuma, 2020). Prediction relies on labeled data to build a regression model between independent variables and the dependent variable (Alkawaz, et.al., 2022). In this study, we check model accuracy by comparing adjusted R-squared values for the variables in the model (Prakash, 2022). The model used should have the largest adjusted R-squared value. The regression models used in this study are simple linear regression and multiple linear regression.

Previous studies have used various machine learning regression models for predictions. Some of them combine models for comparison. For example, researchers applied Decision Tree Regression and Multiple Linear Regression to predict stock prices for Indosat, Telkom, and XL (Thabibi & Supriyanto, 2023). Other researcher predicted Body Mass Index from an asthma patient dataset with Regression (Nuraini, et.al., 2023). In other case, linear regression accuracy

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 5, Number 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.7083

can be compared to Support Vector Regression using Root Mean Square Error (RMSE). The model with the lower RMSE is the more accurate model (Lesnusa, et.al., 2024).

Simple linear regression has appeared in earlier research. For instance, it predicted red onion harvests in Brebes Regency, Central Java. The independent variable was the land area for harvest (Maulana, et.al., 2024). Multiple linear regression also has been used in several past studies. One of them is predicted exam results using 33 attributes as independent variables (Soleh, et.al., 2023) and tehe other is predicted student performance using multi-label classification and hybrid regression (Alshanqiti & Namoun, 2020). Another research is predicted cases of malnourished toddlers based on several factors like vitamin A provision, healthy homes, exclusive breastfeeding, active health centers, and more. Model evaluation that used in this research are R-squared and RMSE (Aprihartha, et.al., 2025). Other research predicted physical performance at the Big Hall of National Street Implementation (BBPJN). Independent variables used are physical planning and realization factors (Prasmono & Ahdika, 2023). Further work predicted house prices based on number of bedrooms, land size, building size, and others. In this study, multiple linear regression was evaluated with R-squared, RMSE, and Mean Absolute Error (MAE) (Nuris & Nuzuliarini, 2024; Zao,et.al., 2023).

Adjusted R-squared is an evaluation metric for regression model accuracy. The metric value works based on the number of independent variables and sample size so it can be used to see the performance of independent variables in a regression model. Adjusted R-squared can stand alone if there are no outliers (Tatachar, 2021; Karch, 2020).

METHOD

In this research, the authors carried out five research stages (Lesnusa, et.al., 2024) as shown in Figure 1. These stages are literature review, data collection, data cleaning, modeling, and the last is the model evaluation stage.

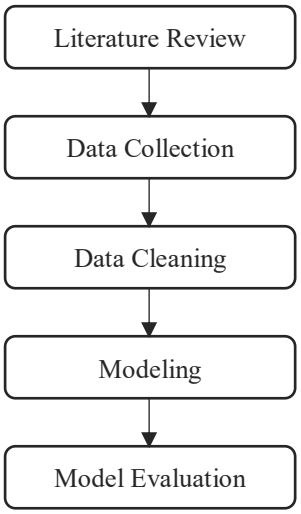

Fig 1. Research Methodology

In the literature review stage, the authors identified the problem by reviewing library materials as references. Information from various sources was collected and selected to support the research. Selected scientific articles used as references included Indonesian public literacy, machine learning, simple linear regression, multiple linear regression, correlation analysis, and other topics relevant to the research topic. A full explanation of this step is provided in the introduction.

The second stage is data collection. In this study, the authors used data taken from the official website of the Central Statistics Agency (BPS). The data used includes calculated numbers based on BPS data from 2021 to 2024. The variable data are as follows: community literacy development index for each province, the number of accredited libraries in each province, the level of reading interest of people in each province, proportion of population living below 50% of the median income in each province, and high school completion rate. To simplify variable naming during analysis, the first variable is referred to as IPLM. This IPLM serves as the dependent variable for the entire regression analysis process. The second variable is referred to as JPA, while the third variable is called MB. The forth variable is called PDBM, and the last data is called SSMA. The variables JPA, MB, PDBM, and SSMA serve as independent variables and are analyzed to determine whether they can influence the IPLM value.

Before analysis can be performed, the data obtained must be cleaned on the third stage. The first step in data cleaning is checking for empty or null values. Several provinces, such as Southwest Papua, South Papua, Central Papua, and Highland Papua, have null values in several variables. This is because these provinces are young, newly established provinces. Due to these null values, the aforementioned provinces were removed and not included in the analysis process. Thus, the total number of provinces analyzed for this study is 34.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.7083

The next step after removing null values is feature selection. In this step, independent variables with low correlations with the dependent variable can be removed. This is intended to increase model accuracy. Before correlation analysis, a normality test is necessary for the dependent variable. If the dependent variable is normally distributed, the Pearson correlation analysis can be performed. If the dependent variable is not normally distributed, the Spearman or Kendall correlation analysis can be performed. The results of the normality test indicate that the IPLM variable is normally distributed, as it has a p-value of 0.1576 > 0.05. Therefore, the Pearson correlation analysis can be performed. The Pearson correlation analysis can be visualized using a correlation heatmap, as shown in Figure 2. Figure 2 shows the correlation values between the independent and dependent variables. Variables with positive (>0.1) and negative (<-0.1) correlation values can be included in the analysis process. From the correlation analysis, it can be concluded that all independent variables can be included in the regression analysis process.

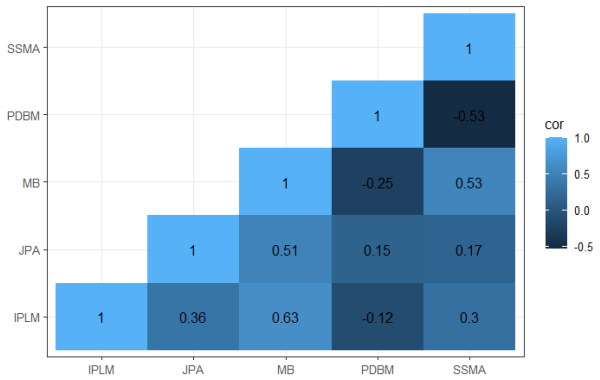

Fig 2. Heatmap correlation

Before modeling stage, descriptive analysis is carried out on the existing data. Descriptive analysis was used to determine the characteristics of the independent variable data. The results of the descriptive analysis from RStudio are shown in Table 1. The JPA variable has the largest value because it calculates the number of accredited libraries for each province. However, there is a significant gap in the JPA values. One province has only 15 accredited libraries, while another province has 3,129. Looking at the average for all variables, JPA has an average of 418.71, SSMA has an average of 65.81, MB has an average of 69.41, and PDBM has an average of 9.07. The average for JPA is greater than the standard deviation value. Meanwhile, SSMA, MB, and PDBM have averages greater than the standard deviation, so it can be concluded that for these three variables, the data deviation is small and the distribution of values is spread evenly.

Table 1. Descriptive Analysis

Variable	Mean	Median	StD	Min	Max
JPA	418.71	204.50	655.88	15.00	3129.00
SSMA	65.81	67.02	10.69	39.50	89.69
MB	69.41	69.78	5.70	50.86	79.99
PDBM	9.07	7.65	7.01	0.06	23.61

The fourth stage of this study is modeling. In multiple linear regression modeling, all independent variables will be included in the analysis process. From the modeling results in RStudio, a regression model is obtained as shown (1). After obtaining the regression model, residual normality and homoscedasticity tests were performed. The residual normality test was performed using the Kolmogorov-Smirnov method. The normality test yielded a residual p-value of 0.2956 > 0.05, and from this value, it can be concluded that the residuals are normally distributed and the regression model (1) can be used to predict IPLM. The residual boxplot also shows that the residuals are normally distributed and no outlier in it, as shown in Figure 3.

$$\text{IPLM} = 0.0006247\,\text{JPA} + 0.8899710\text{MB} + 0.0016548\text{PDBM} - 0.0336325\text{SSMA} + 10.8316739 \quad (1)$$

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.7083

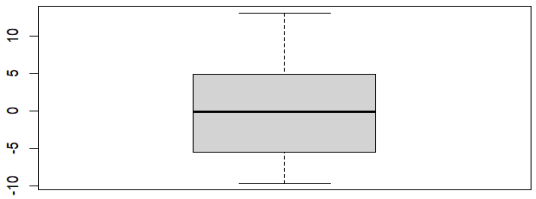

Fig 3. Boxplot Residual

The next step is to test for homoscedasticity. From the test results using RStudio, it can be concluded that the assumption of residual homoscedasticity is met. This can be seen in Figure 4, where the q-q residuals follow a straight line pattern and the fitted values are randomly distributed. The points on the graph are evenly and randomly distributed around the zero line without forming a specific pattern. This means that the residual variance is constant at all levels of the predicted value or independent variable. From the results of the homoscedasticity test that has been conducted, equation (1) can be used to predict the IPLM value based on the variables JPA, MB, PDBM, and SSMA with an adjusted R squared value for this regression model of 0.3235.

The next step is to build a simple linear regression model for each variable. Simple linear regression analysis was conducted to determine the effect of a single independent variable on the dependent variable. In this study, each of the variables JPA, MB, PDBM, and SSMA was also examined using simple linear regression. The analysis results obtained a regression model equation as shown in (2), (3), (4), and (5). Equation (2) is the model for the JPA variable, equation (3) is the model for the MB variable, and equation (4) is the model for the PDBM variable. Meanwhile, the model for the SSMA variable is explained in (5). The simple linear regression equation model sequentially has adjusted R squared values of 0.1058, 0.3828, -0.01586, and 0.06148.

$$\text{IPLM} = 0.004461\,\text{JPA} + 68.802469 \qquad (2)$$

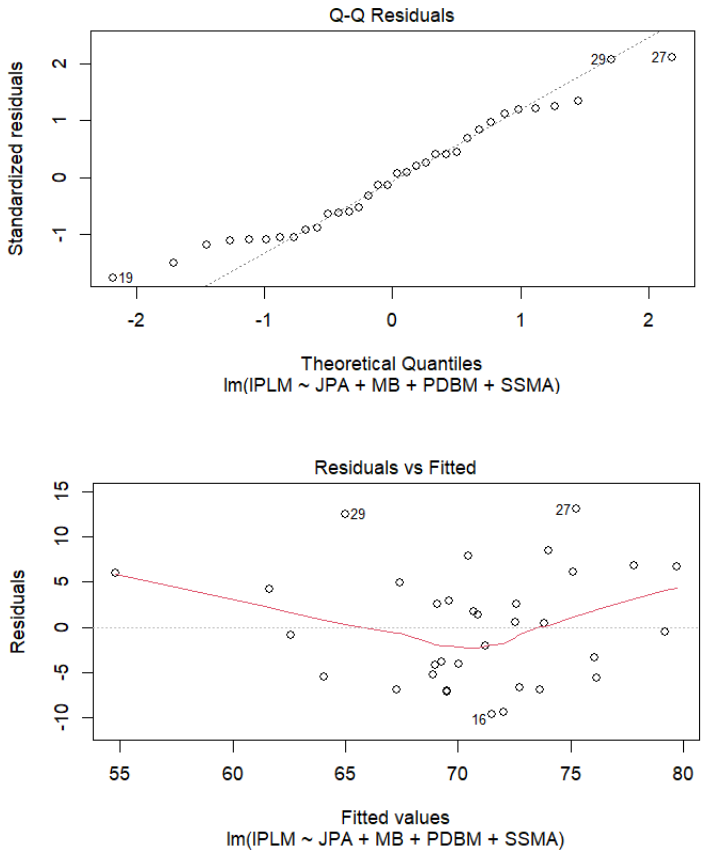

Fig 4. Plot Residual

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.7083

$$\text{IPLM} = 0.8923\,\text{MB} + 8.7317 \quad (3)$$

$$\text{IPLM} = -0.140\,\text{PDBM} + 71.940 \quad (4)$$

$$\text{IPLM} = 0.2252\,\text{SSMA} + 55.8523 \quad (5)$$

The las stage of this study is model evaluation. The stage is performed by comparing the adjusted R-squared of all regression equations. The adjusted R-squared of each variable is examined. If the value in simple linear regression is greater than the adjusted R-squared value in multiple linear regression, then the variable should be used to predict IPLM using simple linear regression. If the adjusted R-squared value is greater in multiple linear regression, then the variable should be used to predict IPLM using multiple linear regression.

RESULT

The previous discussion explained that correlation analysis was performed for feature selection. The correlation analysis was performed on all independent variables. The correlation value with the dependent variable was observed and used to eliminate variables with relatively low correlation. The correlation analysis yielded the correlation values shown in Table 2. Based on the aforementioned criteria, the four variables were included in the analysis process.

Table 2. The correlation value of the independent variable against the dependent variable

Variable	Correlation Value
JPA	0.36
MB	0.63
PDBM	-0.12
SSMA	0.3

Regression modeling was performed using two methods i.e. multiple linear regression and simple linear regression. Regression analysis was used to predict the value of the dependent variable IPLM based on the independent variables JPA, MB, PDBM, and SSMA. The adjusted R-squared values for each variable were compared in both regressions. The adjusted R-squared comparison is shown in Table 3. Table 3 shows that the adjusted R-squared value for the multiple linear regression was 0.3235. Meanwhile, the simple linear regression analysis yielded adjusted R-squared values of 0.1058, 0.3828, -0.01586, and 0.06148, respectively, for the variables JPA, MB, PDBM, and SSMA

Table 3. Adjusted R-squared value

	Adjusted R-squared	
Variable	Simple Linear Regression	Multiple Linear Regression
JPA	0.1058	0.3235
MB	0.3828	
PDBM	-0.01586	
SSMA	0.06148	

DISCUSSION

For the study, data was taken from the official website of the Central Statistics Agency (BPS). Data from each province in Indonesia were collected to determine the number of accredited libraries, reading interest, the number of residents with below-median income, and the proportion of residents who completed high school. These data were selected because they are suspected of influencing the Community Literacy Development Index (IPLM). This study aims to compare the accuracy of regression models when analyzing using one variable and multiple variables. Each variable's performance is evaluated using adjusted r-squared.

CONCLUSION

Based on the correlation analysis, the factor with the highest influence on the IPLM is the level of reading interest, with a correlation value of 0.63. Meanwhile, a comparison of the adjusted R-squared (R-squared) showed that the reading interest factor also had a higher adjusted R-squared value for simple linear regression (0.3828) than the adjusted R-squared for multiple linear regression (0.3235). From this, it can be concluded that the reading interest factor should be used separately and not analyzed together with other factors when predicting the IPLM value. Meanwhile, the other three factors had smaller adjusted R-squared values for simple linear regression compared to the adjusted R-squared values for multiple linear regression. Therefore, it can be concluded that these three factors should be analyzed together with other factors to more accurately predict the IPLM value.

.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.7083

REFERENCES

Alkawaz, Ali Najem et al. 2022. Day-Ahead Electricity Price Forecasting Based on Hybrid Regression Model. *IEEE Access* 10(September): 108021–33.

Alshanqiti, A., & Namoun, A. (2020). Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification. *IEEE Access* 8: 203827–44.

Aprihartha, M.A., Azzahro, S.P., & Aziza, R. (2025). Pemilihan Model Regresi Linear Berganda Terbaik Untuk Menentukan Faktor-Faktor Penyebab Kasus Balita Gizi Buruk Di Jawa Tengah. *Jurnal EurekaMatika* 13(1): 35–46.

Kalla Institute (2024). Rendahnya Minat Literasi di Indonesia. Accesssed: Aug 30, 2025. https://kallainstitute.ac.id/rendahnya-minat-literasi-di-indonesia/.

Karch, Julian. (2020). Improving on Adjusted R-Squared. *Collabra: Psychology* 6(1): 1–11.

Kusuma, P. D., (2020). Machine Learning Teori, Program, Dan Studi Kasus. Yogyakarta: Deepublish.

Lesnusa, G.N., Angreni, D.S., & Ardiansyah, R. (2024). Perbandingan Akurasi Linear Regression Dan Support Vector Regression Dalam Prediksi Suhu Rata-Rata. *The Indonesian Journal of Computer Science* 13(4): 6112–18.

Maulana, A., Martanto, M., & Ali, I. (2024). Prediksi Hasil Produksi Panen Bawang Merah Menggunakan Metode Regresi Linier Sederhana. *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 4, pp. 2884–2888

Ministry of Communication and Digital (2020). Teknologi Masyarakat Indonesia: Malas Baca tapi Cerewet di Medsos. Accessed: Aug 30, 2025. https://www.komdigi.go.id/berita/sorotan-media/detail/teknologi-masyarakat-indonesia-malas-baca-tapi-cerewet-di-medsos

Nasrullah, R., & Asmarini, P., (2024). Meningkatkan Literasi Indonesia Melalui Optimalisasi," *Badan Pengemb. dan Pembin. Bhs. Risal. Kebijak.*, no. 4, pp. 1–16.

National Library of the Republic of Indonesia (2024). IPLM 2024 Catat Rekor Tinggi, Literasi Nasional semakin Meningkat. Accessed: Aug 30, 2025. https://www.perpusnas.go.id/berita/iplm-2024-catat-rekor-tinggi-literasi-nasional-semakin-meningkat

Nuraini, A.T., Setiawan, A., & Susanto, B. (2023), Perbandingan Kinerja Regresi Decision Tree dan Regresi Linear Berganda untuk Prediksi BMI pada Dataset Asthma, *Jurnal Sains dan Edukasi Sains*, vol. 6, no. 1, pp. 34–43.

Nuris, Nuzuliarini. (2024). Analisis Prediksi Harga Rumah Pada Machine Learning Metode Regresi Linear. *Explore* 14(2): 108–12.

Prakash, Kolla Bhanu. (2022). *Data Science Handbook: A Practical Approach*. Wiley AI.

Prasmono, A.S.P., & Ahdika, A. (2023). Analisis Regresi Berganda Pada Faktor-Faktor Yang Mempengaruhi Kinerja Fisik Preservasi Jalan Dan Jembatan Di Provinsi Sumatera Selatan. *Emerging Statistics and Data Science Journal* 1(1): 47–56.

Soleh, M., Nurnawati, Kumalasari, E., & Uning, L. (2023). Penerapan Data Mining Dengan Metode Regresi Linear Untuk Memprediksi Data Nilai Hasil Ujian Menggunakan RapidMiner. *JISKA (Jurnal Informatika Sunan Kalijaga)* Vol. 8, No(ISSN:2527–5836 (print) | 2528–0074 (online)): Pp. 10 – 21.

Tatachar, Abhishek V. (2021). Comparative Assessment of Regression Models Based On Model Evaluation Metrics. *International Research Journal of Engineering and Technology* 8(9): 853–60. www.irjet.net.

Thabibi, A., & Supriyanto, R. (2023). Perbandingan Model Multiple Linear Regression Dan Decision Tree Regression (Studi Kasus: Prediksi Harga Saham Telkom, Indosat, Dan Xl), *Jurnal Ilmu Teknologi dan Rekayasa*, vol. 28, no. 1, pp. 78–92.

Zhao, Guangcai et al. 2023. State-of-Health Estimation With Anomalous Aging Indicator Detection of Lithium-Ion Batteries Using Regression Generative Adversarial Network. *IEEE Transactions on Industrial Electronics* 70(3): 2685–95.