Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 2, 2025
DOI: 10.30829/zero.v9i2.26156
Page: 490-500

Classification of Numeracy Achievement of Junior High School Educational Units Based on National Assessment Data using Random Forest

[1] **Angelin Ica Pramesti**
Department of Mathematics Education, Universitas Sanata Dharma, Yogyakarta, 55281 Indonesia

[2] **Chatarina Enny Murwaningtyas**
Department of Mathematics Education, Universitas Sanata Dharma, Yogyakarta, 55281 Indonesia

Article Info

Article history:

Accepted, 30 October 2025

Keywords:

Learning environment;
Literacy and numeracy;
National assessment;
Random forest;
Student character.

ABSTRACT

This study classifies numeracy achievement in Indonesian junior high schools using 2023 National Assessment data from 11,399 schools. The Random Forest algorithm was applied because it is able to capture nonlinear relationships and complex interactions between heterogeneous predictors, while simultaneously reducing variance through bagging and out-of-bag validation techniques. Two models were developed, one without and one with literacy variables. The addition of literacy increased accuracy from 82.97% to 90.0% and increased the ROC-AUC value from 0.8986 to 0.9609. Based on Gini importance, literacy was the most influential predictor, followed by religiosity, learning experience, gender equality, and class size. Government policies need to integrate literacy and numeracy improvements within a unified curriculum framework and promote gender equality and contextual learning in schools. Furthermore, utilizing data-driven analysis from the National Assessment is crucial for guiding targeted interventions and equitable resource allocation for numeracy improvement.

Corresponding Author:

Chatarina Enny Murwaningtyas,
Department of Mathematics Education,
Universitas Sanata Dharma, Yogyakarta, Indonesia.
Email: enny@usd.ac.id

1. INTRODUCTION

Numeracy skills at the junior high school level can be an important indicator of educational quality. According to the National Assessment (Asesmen Nasional/AN) framework, numeracy, along with literacy, character surveys, and learning environment surveys, is a core component in forming a comprehensive basis for assessing student learning achievement [1]. This aligns with the Programme for International Student Assessment (PISA), which is conducted by the OECD every three years on students aged approximately 15. PISA assesses mathematics, reading, and science, and collects contextual data through student and school questionnaires [2]. With data collected from the AN and PISA, schools and stakeholders can conduct further analysis of student achievement in numeracy.

In practice, student numeracy performance can be grouped into two categories: above the minimum competency level and below the minimum competency level. Furthermore, the minimum competency level itself is divided into three groups: far below, below, and at the minimum level. This grouping helps schools and policymakers design more targeted interventions. Several other factors, apart from test scores, such as teaching practices and class climate, teacher performance, and school leadership, have been shown from a number of

studies to be significant contributors to student performance in mathematics. Multi-level analyses of large-scale international testing programmes, such as PISA, support these findings [3]. At the same time, a number of subjective attributes of students such as motivation, self-efficacy, and attitudes to mathematics have been shown to associate consistently with enhanced outcomes in numeracy [4].

While this study uses Indonesia's National Assessment data, prior international studies provide a robust conceptual rationale for the application of machine learning methods, in particular Random Forest, to educational outcomes. Random Forest has been successfully applied in capturing complex nonlinear patterns across a large number of educational indicators, especially in large-scale assessments such as PISA. These present its high accuracy and interpretability, thereby making it especially suitable for modeling literacy and numeracy achievement.

Random Forest has proven to be an effective machine learning technique for predicting student reading performance. In Ghimire and Mokhtari [5], Random Forest was used to examine how metacognitive reading strategies predict reading achievement; several of these strategies emerged as consistent and important predictors in the model. Similarly, Low, Lim, and Chua [6] compared Random Forest, Naïve Bayes, and k-Nearest Neighbors in forecasting East Asian students' reading proficiency using PISA 2018 data, and found that Random Forest achieved the highest predictive accuracy.

Random Forests have also been used to predict student numeracy performance. Bayirli, Kaygun and Öz [7], in their analysis of 2018 PISA data from various countries in Asia and the Pacific, determined the importance of variables in predicting mathematics achievement using Random Forests. These included parental education level, availability of educational materials, hours per week students spent studying, and school entry age. Bertoletti et al. [8] identified why girls do not perform as well as boys on mathematics tests when they used a multilevel Random Forest model to analyze the influence of family background, school environment, and cognitive ability on gender differences in mathematics performance. Bernardo et al. [9] analyzed low-achieving science students in the Philippines and were able to improve the fit of their model by adding non-cognitive (e.g., motivation and goals) and contextual (e.g., learning experience) variables to variables based solely on cognitive ability when they used machine learning methods including Random Forests. A review conducted by Wang, Perry, Malpique and Ide [3], showed that there is no single "best" set of predictors for academic success, but that student-, family-, and school-based predictors are equally valid, and it is this combination of predictors that best explains what influences academic success.

However, much of the available evidence comes from international studies that use PISA and other large scale datasets. In Indonesia, applications of data mining to the National Assessment remain limited. Only a small number of investigations have reported aggregation at the school level for numeracy using comprehensive National Assessment indicators, and direct comparisons between models that include literacy predictors and those that exclude them are uncommon. This study addresses these gaps by classifying numeracy among junior high schools with Random Forest across three domains (school background, learning environment, and character) and by comparing model specifications without literacy and with literacy to estimate the additional contribution of literacy to model performance. Our objectives are to quantify differences in accuracy and ROC AUC and to identify the most influential predictors for practical intervention. We hypothesize that including literacy yields higher accuracy and ROC AUC and that literacy will rank among the top predictors, and we expect these gains to remain robust across cross validation folds and school strata, providing methodological and practical guidance for schools and policymakers in Indonesia.

2. RESEARCH METHOD

This research utilized a quantitative approach, employing the Random Forest algorithm to classify numeracy-achievement data of junior high school students derived from the 2023 National Assessment. Python was used for all data analyses. The original database contained 137 variables including background of the educational unit, scores for numeracy and literacy assessments, the Learning Environment Surveys and Character Surveys. To align with the purposes of the research effort and keep the model parsimonious, some questions on the two surveys were aggregated into relevant composite indices based on authority from the Assessment and Learning Center [10], [11]. Thus, a total of 68 working variables were employed. The student data were aggregated at the school level so, subsequently, only school-level data were analyzed.

All categorical indicators in this study were binary and coded 0 or 1. This coding preserves the grammar of meaning of each category without inordinate dimensionality. It is well-suited to two-level factors and computationally economical for tree-based learners [12]. This scheme was used for four indicators: school type (0 = public/state, 1 = private), curriculum (0 = 2013 Curriculum, 1 = Merdeka Curriculum), region type (0 = regency/ district, 1 = city/ municipality), and regional status (0 = urban, 1 = rural). This structure served to enhance the fitness of the model while providing the algorithm opportunity to use the categorical data.

Meanwhile, numerical variables consist of numeracy scores, literacy scores, socioeconomic status, learning environment surveys, and character surveys. The learning environment survey and character survey have a large number of indicators. Therefore, to facilitate analysis, these indicators were grouped according to their category framework and averaged [10], [11], [13]. The learning environment survey categories include: (1) classroom management, (2) affective support, (3) cognitive activation, (4) literacy and numeracy learning, (5) teacher

reflection, (6) school policy, (7) diversity climate, (8) gender health climate, and (9) parental support. Meanwhile, the character survey was grouped into six dimensions, namely: (1) faith & piety , (2) cooperation, (3) creativity, (4) critical thinking, (5) global diversity, and (6) independence.

The target variable is numeracy achievement at the school level. Individual numeracy scores (0–100) were first aggregated to the school level by calculating the average of the sample of students participating in the National Assessment. The average numeracy score was then rescaled from 0–100 to a scale of 1–3, and based on this scale each school was mapped into four achievement categories according to government guidelines [13]. Table 1 summarizes the category labels and their respective cut-off values.

Table 1: Range of numeracy achievement

Category	Indicator	Score Range
Well Below Minimum Competency	Most students have not reached the minimum competency threshold for numeracy.	1.00 to 1.39
Below Minimum Competency	Less than 50% of students have reached the minimum competency for numeracy.	1.40 to 1.79
Reaching Minimum Competency	Most students have reached the minimum competency threshold for numeracy, but more effort is needed to encourage more students to become proficient.	1.80 to 2.09
Above Minimum Competency	Students at the school demonstrate a competent level of numeracy, and a significant number of students are at the proficient level.	2.10 to. 3.00

For binary classification purposes, the four ordinal categories of the target variable were transformed into two: class 0 (at or below minimum competency; n = 3,244 educational units) and class 1 (above minimum competency; n = 7,908 educational units). Next, we built two comparative models to assess the incremental predictive value of the literacy variable: Model A excluded literacy scores, while Model B included them.

Before we start modeling, we applied pre-processing in the form of features cleaning, to increase interpretability and quality of the data. Near zero variance predictors were removed as they had little impact on discrimination. We also used a Pearson correlation method to prune predictors that were highly correlated ($|r| > 0.90$) with each other, thereby reducing redundancy and stabilizing variable-importance estimates [14]. This follows the principle of parsimony in machine learning.

Data were split into training and test, by way of a stratified 70:30 partition. This also has the benefit of having plenty of samples available from the training set to learn the best fitted model, with a large enough hold out for fair evaluation. Stratification preserves the class proportions in both sets reducing the evaluation bias on class imbalance [15]. It is also worth noting that small sample sizes can amplify uncertainty of estimates in validation [16].

We used Random Forests as bagging decorrelates decision trees which controls variance as well as reduces the chance of overfitting and allows for mixed type predictors and non-linear interaction to be accounted for [17], [18]. In our binary target class, 0 (≤ minimum competency) there were 3,244 educational units (29.1%) whereas class 1 (> minimum competency) comprised 7,908 units (70.9%) to give a fairly even unbalance of 2.44: 1 in favour of a majority. To account for this imbalanced distribution whilst leaving the data distribution unchanged, we complexly applied cost sensitive method by way of classweight = "balancedsubsample" to give us class re-weighting independent and in each bagging samples used to grow a tree. For tree b, the weight for class k is given by Equation (1); these weights enter the impurity and split-gain computations, up-weighting the minority class thereby mitigating majority-class bias [19], [20].

$$w_k^b = \frac{n_b}{K \times n_k^b}, \quad k \in \{0,1\} \tag{1}$$

where w_k^b is the class weight for class k in tree b; n_b is the size of the bootstrap sample for tree b; K is the number of classes; and n_k^b is the number of in-bag samples of class k in that bootstrap sample. Gini impurity can be computed as in Equation (2).

$$G(t) = 1 - \sum_{k=1}^{K} p_{k|t}^2 \tag{2}$$

where $G(t)$ is the Gini impurity at node t; $p_{k|t}$ is the proportion of class k among samples reaching node t; and K is the number of classes [21]. The impurity reduction (gain) for candidate can be computed as in Equation (3).

$$\Delta G(s,t) = G(t) - \frac{n_{t_L}}{n_t} G(t_L) - \frac{n_{t_R}}{n_t} G(t_R) \tag{3}$$

where n_{t_L} and n_{t_R} are the (weighted) sample counts in the left and right child nodes; n_t is the (weighted) sample count at node t; $\mathrm{G}(\cdot)$ is defined in Equation (2); and $\Delta G(s,t)$ denotes the impurity decrease from split s at node t. Majority-vote aggregation can be computed as in Equation (4).

$$\hat{y}(x) = \arg\max_{k\in\{0,1\}} \sum_{b=1}^{B} \mathbf{1}\{h_b(x) = k\} \tag{4}$$

where $\hat{y}(x)$ is the predicted class label; h_b denotes the b-th tree; B is the number of trees; and $1\{\cdot\}$ is the indicator function (1 if the condition holds, 0 otherwise).

To determine which features are most important for classification, we use Mean Decrease in Impurity (MDI) or how much on average a particular feature decreases the Gini impurity at each split for every tree in the forest (equation 5). MDI is chosen because it is computationally fast, and it is consistent with how Random Forest uses impurity based decision making in high-dimensional data. In order to help MDI be less sensitive to feature cardinality and inter-feature correlation, we removed highly correlated features via correlation filtering and normalized the reported importances to sum to one.

$$I_{\mathrm{G}}(j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{t\in\mathcal{T}_b:\, v(t)=j} \frac{n_t}{N}\,\Delta G(t) \tag{5}$$

where $I_{\mathrm{G}}(j)$ is the Gini-importance of feature j; $\mathcal{T}_b$ is the set of internal nodes of tree b; $v(t)$ is the splitting variable at node t; n_t is the (weighted) number of samples reaching t; N is the total (weighted) number of training samples; and $\Delta G(t)$ is the impurity decrease achieved at node t.

After generating predictions with the Random Forest, we computed Accuracy, Precision, and Recall at a single decision threshold applied to the predicted probabilities. With the confusion-matrix notation TP (true positives), TN (true negatives), FP (false positives) and FN (false negatives), the metrics are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

In Random Forest, the probability score for the positive class is the average of the tree-wise probabilities, where each tree's probability equals the fraction of positive training samples in the terminal leaf reached by the instance.

To evaluate discrimination across thresholds, we used the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). Definitions:

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

AUC is the area under the ROC curve. In practice, we estimate it numerically using the trapezoidal rule over ordered ROC points (x_i, y_i), where $x_i = FPR_i$ and $y_i = TPR_i$:

$$\widehat{AUC} = \sum_{i}^{k-1} \frac{y_i + y_{i+1}}{2}(x_{i+1} - x_i) \tag{11}$$

The ROC–AUC can also be interpreted as the probability that a randomly chosen positive instance receives a higher score than a randomly chosen negative instance [22]. The random baseline AUC is 0.5. Because our data are highly imbalanced, we also report the Precision–Recall (PR) curve and PR–AUC, which are often more informative for minority-class performance than ROC–AUC [23].

Therefore, the research methodology in this study utilizes not only technical aspects, but also consideration of the quality of data, data processing, and transparent evaluation techniques. The novelty of this research is in examining the role of variables of literacy in classifying the level of implications of numeracy at the level of the school, using the data of the National Assessment, the consideration of the environment of learning and character in it.

3. RESULT AND ANALYSIS

3.1. Descriptive Statistics

A total of 15,842 junior high school (SMP) educational institutions participated in the 2023 National Assessment. The sample size was determined using a government-determined sampling design for system-level monitoring purposes. Educational institutions participating in the National Assessment included Junior High Schools (SMP), Madrasah Tsanawiyah (MTs), Package B (equivalent non-formal education), Christian Theological Junior High Schools (SMPTK), Salafi Wustha Islamic Boarding Schools (PPS Wustha), and Madya WP (Islamic Junior High Schools). This demonstrates that the National Assessment encompasses not only public and private formal schools but also religious and non-formal schools that offer educational programs equivalent to grades 7-9.

The majority of institutions included in the sample are public junior high schools (SMP), with a total of approximately 11,399. This represents approximately 72 percent of all institutions considered in this study. The remaining institutions consist of religious institutions such as Madrasah Tsanawiyah (Islamic junior high schools) and a small number of other non-formal and specialized educational institutions. Although the number of these institutions is small, their inclusion provides valuable depth to the analysis as it illustrates the diversity and complexity of the education system as expressed at the junior high school level in Indonesia. However, the dominance of junior high schools is the primary reason why this study was conducted at this level of analysis. Junior high schools are the most widespread type of school in the country and are therefore considered representative of the general state of numeracy achievement for the formal junior high school system. This focus is policy-relevant: MDI-based feature importance isolates actionable levers at the junior-high level. High-ranking predictors, SES, literacy index, teacher support, and classroom environment, guide resource allocation and the design of targeted professional development. Districts can operationalize these indicators within dashboards to prioritize schools and track intervention impact. The junior-high focus narrows scope but increases actionability for national decision-making and instructional design.

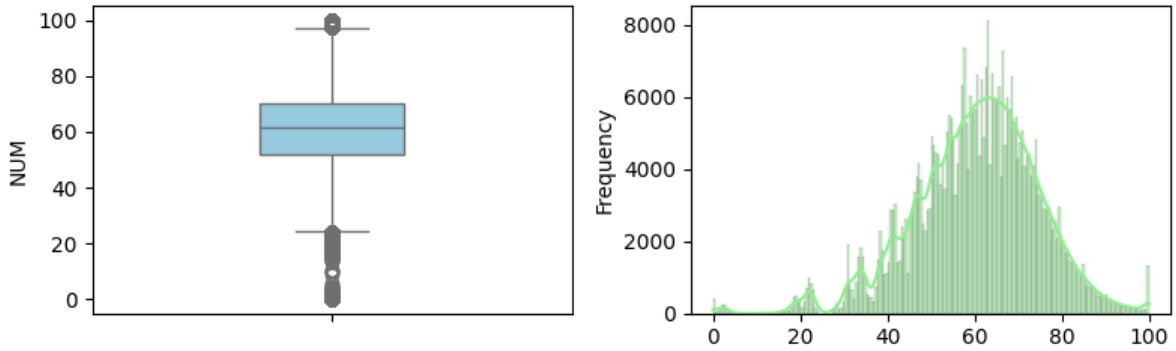

Figure 1. Numeracy Score Distribution

A descriptive summary of junior high school students' numeracy achievement is shown in Figure 1. The mean achievement was 60.44 with a standard deviation of 14.30. The lowest recorded score was 0.00 and the highest was 100.00. This description is further clarified by using quartiles for the distribution, namely Q1 = 51.67, Q2 = 61.44, and Q3 = 69.93. The interquartile range of 18.26 indicates that 50 percent of the group fell within the achievement range between 51.67 and 69.93. Thus, the distribution of numeracy scores tends to be concentrated in the group of schools with moderate to relatively high achievement. The distribution of numeracy scores as seen in Figure 1 is nearly symmetrical, as the difference between the median and mean is relatively small. Therefore, the data are not extremely skewed. However, there are several outliers on the right and left sides of the distribution. Some schools scored very low, even close to 0. Some schools scored very high, or even close to the upper limit of 100.

The wide spread of numeracy results reflects differences in learning environments across regions in Indonesia. This means that the resources students have access to likely vary significantly across schools. Furthermore, the quality of teaching and other factors, such as the learning environment, can also influence differences in student achievement. This finding aligns with previous research by Wulandari et al. [24], Yerizon et al. [25], and Arwi & Lestari [26], which showed that student achievement is influenced by various factors, including structural aspects and the context of their educational experiences. Some schools exhibit very low numeracy levels, indicating significant obstacles to achieving equal educational opportunities for all students in Indonesia. Conversely, the presence of schools with high numeracy achievements indicates the presence of various supporting factors that can support the learning process and improve student outcomes.

Overall, the descriptive analysis presented provides a clear and in-depth picture of the state of numeracy achievement at the junior high school level in Indonesia. The distribution of scores generally follows a normal pattern, although there are extreme values at both ends of the distribution. In the classification modeling process,

outlier values were retained to reflect significant differences in achievement between schools. This emphasizes that efforts to improve educational quality must be tailored to the specific needs of each school. Therefore, it is crucial to design targeted support programs for schools that have not yet achieved high achievement standards to narrow the gap in education quality between educational units. Meanwhile, successful schools can serve as examples of good practices that can be replicated more widely across Indonesia.

The variables used in this study encompass three main dimensions: background of the educational unit, learning environment, and a student characteristics survey. The educational background dimension consists of school type, curriculum, region type, region status, number of students, number of computers, number of libraries, aid recipients, socioeconomic status, and literacy scores. This dimension serves as a basic indicator representing the structural conditions and resources of schools, thus providing a starting point for explaining differences in numeracy achievement.

The learning environment dimension encompasses various aspects reflecting instructional practices and school climate, including learning enhancement, classroom management, literacy learning, numeracy learning, gender equality, affective support, cognitive activation, student experiences, conceptions & efficacy, and programs & policies. These variables represent contextual factors that directly and indirectly influence student learning processes and outcomes, making them crucial in analyzing the determinants of numeracy achievement [24].

The dimensions of student character encompass values, attitudes, and non-cognitive competencies, including faith & piety, mutual cooperation, creativity, critical thinking, global diversity, independence, inclusive climate, and external support. This dimension is important because numeracy competency development is influenced not only by academic factors but also by affective and sociocultural aspects that shape students' learning attitudes and resilience [25].

Table 2. Correlation Between Independent Variables and Dependent Variables

Variabel	Numeracy Score	Variabel	Numeracy score	Variabel	Numeracy score
School type	0.0046	learning improvement	0.1524	belief in diversity	0.4246
curriculum	0.2367	classroom management	0.2531	inclusive climate	0.3960
region type	0.1829	literacy learning	0.1845	external support	0.2660
area status	-0.2616	numeracy learning	0.1203	faith & piety	0.4706
total students	0.3028	gender equality	0.4419	mutual cooperation	0.4593
total computers	0.0440	affective support	0.2810	creativity	0.3017
total libraries	0.0778	cognitive activation	0.2217	critical thinking	0.3892
recipients of assistance	0.2128	student experience	0.4730	global diversity	0.2930
socioeconomic	0.2544	conception & efficacy	0.3110	independence	0.3684
literacy score	0.7115	programs & policies	0.3793		

Based on the correlation analysis shown in Table 2, it can be seen that the strength of the relationship between the independent variables and numeracy achievement varies. Some of these variables, such as student experience ($r = 0.4730$), faith & piety ($r = 0.4706$), and literacy score ($r = 0.7115$), are correlated with numeracy achievement, so these variables need attention to improve numeracy. However, some of these variables, such as school type ($r = 0.0046$), total computers ($r = 0.0440$), and total libraries ($r = 0.0778$), have very low correlations ($r < 0.1$), indicating that these variables need to be considered for removal during the feature cleaning stage to minimize information gain.

Although Random Forest is generally tolerant of multicollinearity, there are situations where multicollinearity (highly correlated predictor pairs, e.g., $|r| > 0.90$) can lead to decreased model stability, less accurate feature importance estimates, and redundant information between features. Therefore, to address these issues and improve model reliability and effectiveness, correlation-based filtering methods are applied to remove predictors with high correlations with each other. Predictors with high correlations were removed to eliminate multicollinearity between variables and improve model efficiency and accuracy, leaving only those predictors that significantly contribute to predicting numeracy achievement. This approach not only improves the technical rigor of the predictive model but also provides the context needed to formulate effective, data-driven education policies.

3.2. Classification Model Analysis

The data collected to compare the performance of the two models (without literacy and with literacy) clearly shows significant differences. For the baseline model, the model without literacy, accuracy reached 82.97% and precision reached 0.8838. Recall and F1 scores reached 0.8748 and 0.8793, respectively. When the model used literacy scores, we found that each performance measure experienced changes in the same direction and at the same rate; accuracy increased to 90.00% and precision increased to 0.9292, recall increased to 0.9298, and the F1 score increased to 0.9295. Meanwhile, the ROC AUC changed from 0.8986 to 0.9609. The increase in the ROC AUC value indicates that the model is better able to distinguish educational entities that are above and below the minimum level of numeracy competency. The results of the comparative study can be seen in Table 3.

Table 3. Comparison of Model Performance

Model	Accuracy	Precision	Recall	F1-score	ROC AUC
Without literacy scores	0.8297	0.8838	0.8748	0.8793	0.8986
With literacy scores	0.9000	0.9292	0.9298	0.9295	0.9609

Although the data distribution indicates improved performance, the national average numeracy score of 60.44 remains low compared to international standards. This may reflect differences in the difficulty of assessment instruments. National assessments, which measure achievement of minimum competencies required by the national curriculum, tend to use simple items. International assessment instruments such as the Programme for International Student Assessment (PISA), assess numeracy in complex, cross-cultural, real-life situations that require higher-order thinking skills [25], [27].This clearly indicates that although a significant number of schools in Indonesia have passed national standards, graduates' preparedness for international numeracy situations remains somewhat limited. Therefore, policies to improve numeracy should not only address national standards but also focus on the application of numeracy to real-life problems, as reflected in the PISA framework.

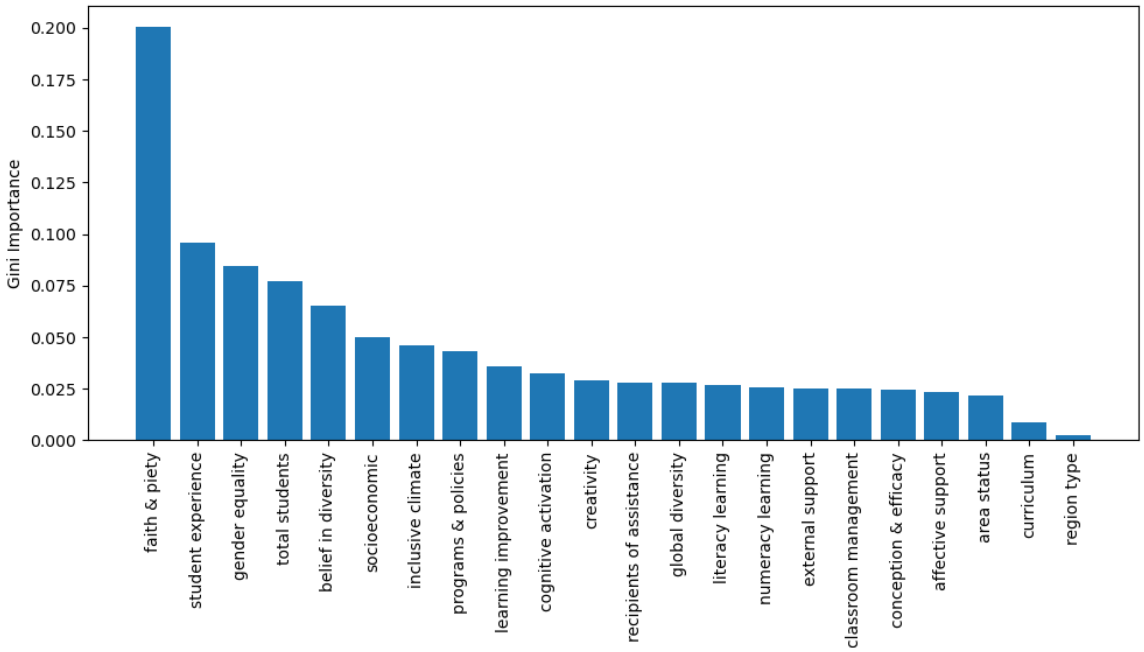

Figure 2. Important Features of the Baseline Model (Without Literacy)

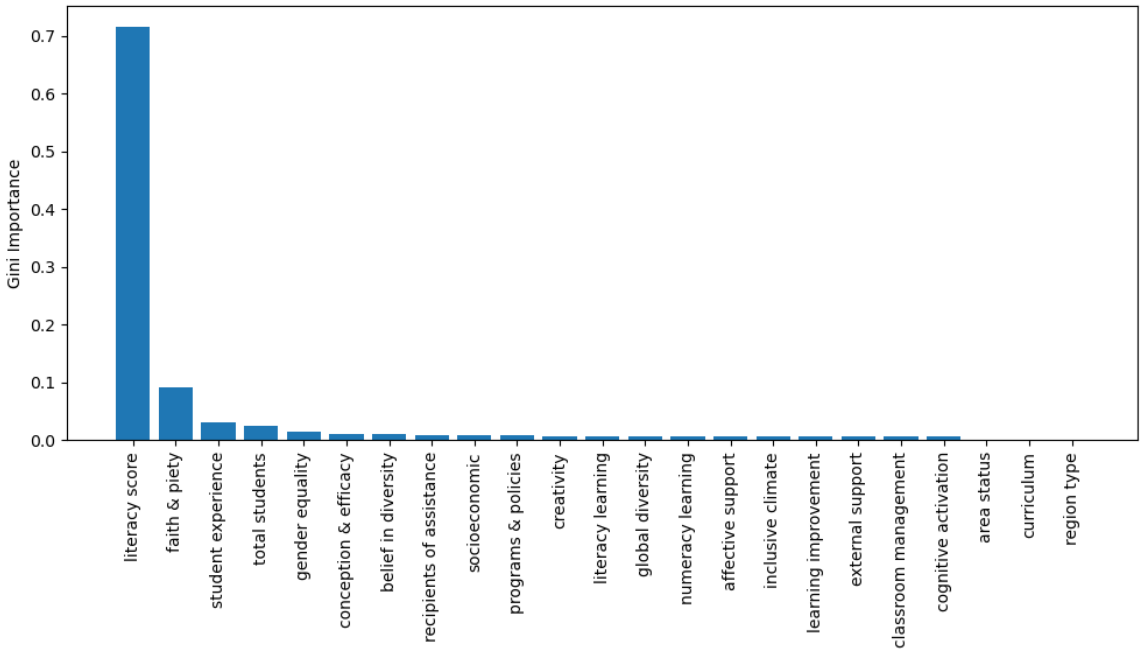

Figure 3. Important Features of the Literacy-augmented Model

Identifying the most influential variables in explaining outcomes is crucial, as it allows for the design of more specific interventions. The key features of the model without literacy are shown in Figure 2, while the key features of the model with literacy are shown in Figure 3. Findings related to the key features in Figures 2 and 3 indicate that the order of dominance of the variables exhibits a different pattern between the two models. In the case of the model using literacy, literacy scores were identified as the most dominant predictor, with a significantly greater weight than other variables. This is consistent with the previously drawn conclusion that literacy serves as the basis for classifying educational units as above or below a minimum level of numeracy competency. Internationally, the close relationship between literacy and numeracy has been well documented, showing that literacy competency, reading confidence, and reading learning conditions are significant contributors to performance in mathematics and science education [28]. Therefore, it is crucial that programs designed to address numeracy complement programs that address the literacy strengthening dimension in both instructional and curriculum integration.

In addition to literacy, a number of non-academic variables are also ever-present as important predictors in both models. The variables faith & piety come in second after literacy as predictors in the model with literacy, and even become the chief predictors in the model without literacy. The connection of religiosity to academic achievement is explained by more discipline, intrinsic motivation, and responsibility for learning. That is, the education of character and moral values are viewed as laying a foundation for consistent and responsible modes of learning behavior [29].

The variable of student experience emerges, too, as an important variable. Student involvement, active learning, hands-on opportunities for practice and opportunities for problem-solving are closely linked to numeracy achievement. Experiential learning theory posits that authentic learning experiences allow students to tie abstract mathematical concepts to real-world experience and enhance their understanding of problem-solving skills they already have. Other studies indicate that this experiential learning enhances concept retention and critical thinking [30].

Additionally, gender equity has become an important predictor variable, although of somewhat lesser weight. Its consistency of emergence indicates that gender equity in the learning process is influential in numeracy achievement. The OECD report indicates that gender bias in access and classroom experiences can impact PISA results, and gender gap studies indicate that when access and opportunities for learning become equal, the gender gap in mathematics practically disappears [31].

Lastly, total students are another important variable. Too many students create situations where the teacher-student interaction is impaired, and not enough individual attention is given. The size of classes can have an effect on the participation from students and teacher-student interaction, while small classes create an environment where individual attention can be improved and teaching activities can be more productive [32].

In addition to these predominant variables, variables such as programs & policies, learning climate, diversity climate, safety environment and teacher reflection also appear, but with lesser weight. Effective leadership in implementing programs and policies in schools has a positive impact on the learning process. while the inclusive classroom learning climate can generate improved student participation [33]. It is seen that although the impact is of lesser weight than literacy or religiosity, these variables are still significant in forming a learning eco-system that is favorable for numeracy.

The results of the importance of these variables in this section show that numeracy achievement is impacted by a combination of academic (literacy score) and non-academic (faith & piety, student experience, gender equity, total students) variables. Background variables such as leadership, school climate and teacher reflection, although of lesser weight, are nevertheless significant in forming a holistic and effective learning environment for numeracy.

4. CONCLUSION

This study demonstrates the effectiveness of the Random Forest Algorithm in determining the numeracy performance of junior high school students in Indonesia, based on data from the 2023 National Assessment involving 11,399 schools. Furthermore, two different models were constructed to compare the effect of using literacy as a variable in determining the model's classification accuracy. The results indicate that including literacy as a variable increases the overall classification accuracy from 82.97 percent to 90.00 percent, and the ROC-AUC value increases from 0.8986 to 0.9609. Through feature importance analysis, it was found that the strongest predictors of student numeracy include literacy, religiosity, learning experience, gender equality, and class size.

In other words, the findings of this study indicate that students' numeracy abilities are influenced by various cognitive and non-cognitive factors derived from the learning environment and social dynamics within each school. Furthermore, the methodological approach in this study demonstrates that integrating variables from various domains within a data-driven framework enables more accurate predictions and supports the development of more targeted and evidence-based education policies.

Based on the research findings and literature review, we propose three main recommendations. First, the government should prioritize the development of an integrated curriculum that supports literacy and numeracy development. Second, education policies should emphasize efforts to promote gender equality and enrich students' learning experiences through contextual approaches aligned with the social realities of the school environment. Third, data collected through the National Assessment process should continue to be utilized to design data-driven interventions in schools and to distribute educational resources fairly and equitably. However, several other considerations need to be considered when interpreting the results of this study. For example, because this study used cross-sectoral data at the school level, it does not provide information on student progress from year to year or the long-term impact of school-based interventions. Furthermore, this study did not conduct longitudinal validation of the model, so its stability over time has not been tested. Therefore, future research is recommended to use a longitudinal approach to assess student numeracy growth and the sustained impact of data-driven policies on learning outcomes in schools. Finally, this study contributes to the development of educational policies that utilize data as a basis for decision-making processes, as well as to the development of policies that fairly integrate literacy and numeracy according to the social context in which the policies are implemented.

5. REFERENCES

[1] M. Mellyzar, N. Novita, M. Muliani, M. Marhami, and S. R. Retnowulan, 'The Literacy and Numeracy Ability Profile Which are Viewed From Minimum Assessment Components (AKM)', *LJ*, vol. 11, no. 2, p. 168, Dec. 2023, doi: 10.22373/lj.v11i2.19866.

[2] OECD, 'PISA 2022 Results Factsheets Indonesia', 2023. [Online]. Available: https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/indonesia_c2e1ae0e-en.html

[3] X. S. Wang, L. B. Perry, A. Malpique, and T. Ide, 'Factors predicting mathematics achievement in PISA: a systematic review', *Large-scale Assess Educ*, vol. 11, no. 1, p. 24, June 2023, doi: 10.1186/s40536-023-00174-8.

[4] K. E. S. Street, L.-E. Malmberg, and S. Schukajlow, 'Students' mathematics self-efficacy: a scoping review', *ZDM Mathematics Education*, vol. 56, no. 2, pp. 265–280, May 2024, doi: 10.1007/s11858-024-01548-0.

[5] N. Ghimire and K. Mokhtari, 'Evaluating the predictive power of metacognitive reading strategies across diverse educational contexts', *Large-scale Assess Educ*, vol. 13, no. 1, p. 4, Feb. 2025, doi: 10.1186/s40536-025-00240-3.

[6] A. H.-M. Low, A. H.-L. Lim, and F.-F. Chua, 'Predicting Factors that Affect East Asian Students' Reading Proficiency in PISA', *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 3–2, pp. 2065–2074, 2023, doi: 10.30630/joiv.7.3-2.2341.

[7] E. G. Bayirli, A. Kaygun, and E. Öz, 'An Analysis of PISA 2018 Mathematics Assessment for Asia-Pacific Countries Using Educational Data Mining', *Mathematics*, vol. 11, no. 6, p. 1318, Mar. 2023, doi: 10.3390/math11061318.

[8] A. Bertoletti *et al.*, 'The Determinants of Mathematics Achievement: A Gender Perspective Using Multilevel Random Forest', *Economies*, vol. 11, no. 2, p. 32, Jan. 2023, doi: 10.3390/economies11020032.

[9] A. B. I. Bernardo, M. O. Cordel, M. O. Calleja, J. M. M. Teves, S. A. Yap, and U. C. Chua, 'Profiling low-proficiency science students in the Philippines using machine learning', *Humanit Soc Sci Commun*, vol. 10, no. 1, p. 192, May 2023, doi: 10.1057/s41599-023-01705-y.

[10] Pusmendik, *Framework Instrumen Survei Karakter*. Jakarta: Pusat Asesmen dan Pembelajaran, Badan Penelitian,Pengembangan dan Perbukuan, 2021. Accessed: Oct. 01, 2025. [Online]. Available: https://pusmendik.kemdikbud.go.id/pdf/file-146

[11] Pusmendik, *Framework Survei Lingkungan Belajar*. Jakarta: Pusat Asesmen dan Pembelajaran, Badan Penelitian,Pengembangan dan Perbukuan, 2021. Accessed: Oct. 01, 2025. [Online]. Available: https://pusmendik.kemdikbud.go.id/an/page/survei_lingkungan_belajar

[12] A. Zheng and A. Casari, *Feature Engineering for Machine Learning*. Accessed: Oct. 19, 2025. [Online]. Available: https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/

[13] Pusmendik, *Framework Asesmen Kompetensi Minimum*. Pusat Asesmen dan Pembelajaran, Badan Penelitian, Pengembangan dan Perbukuan, Kementerian Pendidikan dan Kebudayaan, 2021. Accessed: Oct. 01, 2025. [Online]. Available: https://repositori.kemendikdasmen.go.id/25488/

[14] C. F. Dormann *et al.*, 'Collinearity: a review of methods to deal with it and a simulation study evaluating their performance', *Ecography*, vol. 36, no. 1, pp. 27–46, Jan. 2013, doi: 10.1111/j.1600-0587.2012.07348.x.

[15] R. Kohavi, 'A study of cross-validation and bootstrap for accuracy estimation and model selection', in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, in IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Aug. 1995, pp. 1137–1143.

[16] G. Varoquaux, 'Cross-validation failure: Small sample sizes lead to large error bars', *NeuroImage*, vol. 180, pp. 68–77, Oct. 2018, doi: 10.1016/j.neuroimage.2017.06.061.

[17] L. Breiman, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[18] G. Biau and E. Scornet, 'A random forest guided tour', *TEST*, vol. 25, no. 2, pp. 197–227, June 2016, doi: 10.1007/s11749-016-0481-7.

[19] H. He and E. A. Garcia, 'Learning from Imbalanced Data', *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[20] F. Pedregosa *et al.*, 'Scikit-learn: Machine learning in Python', *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[21] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Chapman and Hall/CRC, 2017.

[22] T. Fawcett, 'An introduction to ROC analysis', *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, June 2006, doi: 10.1016/j.patrec.2005.10.010.

[23] J. Davis and M. Goadrich, 'The relationship between Precision-Recall and ROC curves', in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, ACM Press, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[24] R. I. Wulandari, R. F. Maulana, A. R. Imtiyaz, A. S. Felisa, A. D. Ramadhani, and A. Wulandari, 'Pengaruh Lingkungan Belajar Terhadap Hasil Belajar Peserta Didik di SMP Negeri 8 Gresik', *jipipi*, vol. 1, no. 3, pp. 123–132, Nov. 2024, doi: 10.31004/b4tdaf34.
[25] Y. Yerizon, A. Arnellis, and A. Cesaria, 'Deskripsi Kemampuan Literasi Numerasi Siswa SMP Ditinjau dari Gaya Belajar. Studi Kasus di Kota Padang', *AJPM*, vol. 12, no. 3, p. 2862, Sept. 2023, doi: 10.24127/ajpm.v12i3.8393.
[26] N. I. F. Arwi and M. Lestari, 'Pengaruh Lingkungan Belajar dan Aktivitas Belajar Terhadap Hasil Belajar Siswa pada Mata Pelajaran Fiqih', *GURU: Jurnal Cendekia Profesi*, vol. 1, no. 2, pp. 194–200, June 2024.
[27] K. Stacey, 'The PISA View of Mathematical Literacy in Indonesia', *Journal. Math. Edu.*, vol. 2, no. 2, pp. 95–126, July 2011, doi: 10.22342/jme.2.2.746.95-126.
[28] J. Marôco, 'What makes a good reader? Worldwide insights from PIRLS 2016', *Read Writ*, vol. 34, no. 1, pp. 231–272, Jan. 2021, doi: 10.1007/s11145-020-10068-8.
[29] T. Lickona, *Mendidik Untuk Membentuk Karakter*. Bumi Aksara, 2022.
[30] Y.-M. Huang, P.-S. Chiu, T.-C. Liu, and T.-S. Chen, 'The design and implementation of a meaningful learning-based evaluation method for ubiquitous learning', *Computers & Education*, vol. 57, no. 4, pp. 2291–2302, Dec. 2011, doi: 10.1016/j.compedu.2011.05.023.
[31] OECD, *PISA 2018 Results (Volume I): What Students Know and Can Do*. OECD Publishing, 2019.
[32] P. Blatchford, P. Bassett, and P. Brown, 'Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools', *Learning and Instruction*, vol. 21, no. 6, pp. 715–730, Dec. 2011, doi: 10.1016/j.learninstruc.2011.04.001.
[33] M.-T. Wang and J. L. Degol, 'School Climate: a Review of the Construct, Measurement, and Impact on Student Outcomes', *Educ Psychol Rev*, vol. 28, no. 2, pp. 315–352, June 2016, doi: 10.1007/s10648-015-9319-1.