JMPM: Jurnal Matematika dan Pendidikan Matematika
10 (2), 161-179
https://www.journal.unipdu.ac.id/index.php/jmpm

Print-ISSN : 2502-986X
Online-ISSN : 2502-9878
JMPM Jurnal Matematika dan Pendidikan Matematika
Journals of Mathematics and Mathematics Education

Unveiling Numeracy Competency Domains of High School Students in Indonesia: A Clustering Analysis Approach

Adhi Surya Nugraha[1]*, Marcellinus Andy Rudhito[2]
[1,2] Mathematics Study Program, Faculty of Teacher Training, Sanata Dharma University – Yogyakarta, Indonesia, 55281
*Corresponding author's email: a.s.nugraha@usd.ac.id

Received: 4 October 2025 | Revised: 10 October 2025
Accepted: 16 October 2025 | Published: 30 October 2025

Abstract

This study aims to analyze high school students' numeracy achievement using a data science approach to the 2021–2023 National Assessment data. The analysis covers the domains of Algebra, Geometry, Number, and Data and Uncertainty at three ability levels: knowing, applying, and reasoning. The research methods included secondary data collection, descriptive cleaning and analysis, interdomain correlation, K-Means clustering with Davies–Bouldin Index validation, and result visualization using Python. The results show an increasing trend in national numeracy scores (49,23 to 50,76 to 55,67). The Data & Uncertainty domain reached the highest average (56,79) while having the strongest correlation with the total score. In the cognitive dimension, there was a surge in the ability to applying and reasoning in 2023. The clustering results show that the optimal configuration in 2021-2022 is 7 clusters, while in 2023 it reduces to 3 clusters. These findings suggest prioritizing strengthening data literacy and reasoning skills, along with remediation of number concepts, and targeted interventions based on provincial cluster profiles.

Keywords: numeracy; student competency; national assessment; data science analysis; k-means; clustering

How to cite: Nugraha, A.S, & Rudhito, M. A., (2025). Unveiling numeracy competency domains of high school students in Indonesia: A clustering analysis approach. *JMPM: Jurnal Matematika dan Pendidikan Matematika*, *10*(2), 161-179. https://doi_link_doi_of_article.

INTRODUCTION

In recent years, the emergence of international, regularly conducted assessments in education has consistently generated large databases. These data cover a wide range of variables, such as student achievement, background, and school practices and processes, making them highly useful for educational research. Assessment schemes such as the Programme for International Student Assessment (PISA) of the Organisation for Co-operation and Economic Development (OECD) or the Trends in International Mathematics and Science Study (TIMSS), have had a significant impact on the development of educational research in recent years (Gamazo et al., 2016). However, the relevance of large-scale assessments is not limited to educational research; several authors have also highlighted the significant impact of PISA results on national educational policies and practices (Lingard et al., 2013). However, it is clear that educational policy is often influenced by reports and analyses produced directly by the OECD, as these are the first to be made publicly available after PISA is implemented and because such analyses can be limited by the sheer number of variables PISA offers (Jornet Meliá, 2016). Furthermore, there are certain challenges and responsibilities for educational researchers to delve deeper into the database and find relationships between variables and conclusions that may not be offered by the OECD reports to enrich new knowledge.

Several methodologies can be used for secondary data analysis. The most common is regression analysis, as it allows researchers to simultaneously account for variability at the student and school (Gamazo et al., 2017). Other authors also use different approaches, such as Structural Equation Modeling (Acosta & Hsu, 2014) or ANCOVA (Zhu & Kaiser, 2020). Furthermore, because of the rise of big data, new possibilities for statistical analysis of all types of databases have emerged in recent years. In recent years, data mining has emerged as a powerful analytical approach for large scale educational assessments, such as PISA, enabling deeper insights into factors influencing numeracy (Huang et al., 2024; Kalita et al., 2025; Malik et al., 2025), although it remains a relatively underexplored method in national assessments contexts.

The educational data mining (EDM) approach allows researchers to extract meaningful patterns from large scale educational datasets by building models directly from empirical data, without relying on theoretical distributional assumptions and capable of handling both categorical and numerical variables. In educational context, this technique is used for three main proces: prediction, relationship exploration, and structure identification, including through the application of clustering to map student performance profiles or patterns (Alvarez-Garcia et al., 2024; Barbeiro et al., 2024; Kalita et al., 2025). Recent studies have shown that clustering techniques such as K-means are effective in grouping students based on learning characteristics and assessment results (Sarker et al., 2024; Alvarez-Garcia et al., 2024; Malik et al., 2025; Yel et al., 2024). However, this approach is exploratory in nature, so further studies are needed to confirm and deepen understanding of patterns found.

In the context of Indonesian education, the National Assessment (AN) is an evaluation conducted to measure the overall quality of education in Indonesia. The AN is designed to provide an overview of educational quality based on three main components: the Minimum Competency Assessment (AKM), the Character Survey, and the Learning Environment Survey. The results of the AN are used to provide feedback to schools and the government to continuously improve the quality of education. Furthermore, numeracy in the context of the AN refers to students' ability to understand, use, analyze, and communicate mathematical concepts in various everyday life situations.

Numeracy focuses not only on the ability to calculate, but also on the ability to reason using numbers and data to solve relevant and contextual problems. Numeracy skills in the 21st century encompass the ability to use mathematical knowledge in various life contexts to make informed and responsible decisions. The model in Figure 1 illustrates the conceptual framework for numeracy based on the OECD (2021) which emphasizes the interconnectedness of mathematical knowledge, real-life contexts, digital and physical tools and representations, and critical dispositions and orientations in problem-solving. This approach also aligns with the Independent Curriculum in Indonesia, which views numeracy as part of critical and reflective thinking competencies. Thus, numeracy is not simply a technical skill, but rather a combination of conceptual understanding, a positive attitude toward mathematics, and the ability to use mathematics meaningfully in social, work, and civic life.

Figure 1. Numeracy Models in the 21st Century
(OECD, 2021)

However, the results of the 2021-2023 National Assessment (AN) indicate a significant gap in numeracy competency mastery among high school students in Indonesia. The data reveals striking variations between regions, as well as differences in student achievement (Ministry of Primary and Secondary Education, 2023). This gap not only demonstrates challenges in educational equity but also highlights the need for data-driven interventions to identify key factors influencing numeracy achievement.

In this case, a big data approach is used as the primary method. Python, as a powerful programming language for data analysis, enables efficient processing and analysis of large-scale data from the National Assessment. The research stages involve several main steps, namely: (i) data collection from the official portal data.kemdikbud.go.id, (ii) data cleaning to address issues such as missing values and outliers, (iii) descriptive analysis to understand the distribution of competency scores in each domain, (iv) correlation analysis to identify relationships between competency domains, (v) school clustering based on competency achievement patterns using algorithms such as K-Means, and (vi) data visualization to provide intuitive and easy-to-understand insights (Montshiwa & Botlhoko, 2023).

Numeracy skills are a fundamental competency that plays a crucial role in supporting problem-solving, logical thinking, and data-driven decision-making in the 21st century. However, the results of the 2021-2023 National Assessment (AN) indicate that the numeracy achievement of high school students in Indonesia still varies, with disparities between regions, socioeconomic status, and differences between numeracy competency domains, such as Algebra, Geometry, Number, and Data and Uncertainty. Therefore, this study focuses on identifying key competency domains in numeracy using a data science approach to uncover patterns in student numeracy achievement, analyze factors contributing to the numeracy gap, and generate strategic recommendations that can be used as a basis for improving numeracy learning in schools. Thus, this study not only provides a descriptive mapping of numeracy achievement but also offers data-driven insights to improve the quality of numeracy education in Indonesia in a more targeted and effective manner.

This study offers an innovative approach to analyzing high school students' numeracy achievement by applying data mining and machine learning techniques, which have not been widely used in educational research in Indonesia. Previous studies have generally been limited to simple descriptive and inferential statistical analyses that only describe differences in average scores between regions or types of schools without identifying hidden patterns across numeracy domains or temporal dynamics from year to year. This study uses clustering methods to group students based on their numeracy ability profiles, in-depth correlation analysis to examine the interrelationships between domains such as Algebra, Geometry, Number, and Data and Uncertainty, and predictive modeling to estimate trends in student numeracy achievement over the 2021–2023 timeframe. With this approach, the study is expected to provide a more comprehensive and empirically data-driven picture of the numeracy achievement patterns of high school students in Indonesia.

This approach offers several advantages. First, correlation analysis allows for in-depth mapping of relationships between numeracy competency domains, enabling the identification of domains requiring further intervention. Second, the school-based clustering method helps group educational institutions based on student achievement patterns, allowing strategic recommendations to be tailored specifically to the needs of each group. Third, data visualization provides policymakers with deeper insights and enables evidence-based decision-making. This study also offers novelty in its approach and scope. Most previous studies focused solely on describing AN result without exploring relationships between competency domains or developing data-driven intervention strategies.

Thus, the main objective of this study is to apply data mining techniques, including K-means clustering analysis and correlation analysis, to uncover patterns of relationships between numeracy competency domains and other factors influencing the numeracy achievement of high school students of 2021-2023 National Assessment.

METHODS

This study employed a purely quantitive approach, employing data mining techniques as the primery analytical method. The analysis was conducted in two stages. The first stage employed the K-means clustering algorithm to group students based on achievement patterns in the numeracy competency domain (Liu, 2022; Ouassif et al., 2025), creating clusters that represented the charatceristics of students numeracy performance. The second stage employed correlation analysis to examine the relationships between numeracy domains and their association with contextual factors such as school background and region.

The secondary data analyzed were sourced from the official portal of the Ministry of Primary and Secondary Education. The stages undertaken are: (i) data collection, (ii) data cleaning to address issues such as missing values and outliers, (iii) descriptive analysis to understand the distribution of competency scores in each domain, (iv) correlation analysis to identify relationships between competency domains, (v) school clustering based on competency achievement patterns using algorithms such as K-Means, and (vi) data visualization to provide intuitive and easy-to-understand insights.

In order to explore the relationships among numeracy competency domains and other influencing variables, this study followed a series of analytical steps. The process, described in detail below, employs data mining techniques such as K-means clustering and correlation analysis using data from the 2021–2023 National Assessment of high school students.

(i) Data Collection: The primary data used in this study are the results of National Assessment (AN). This data includes various information, as shown in Table 1. The population of this study included all high school students in Indonesia who took the National Assessment (AN) in 2021, 2022, and 2023. Data were obtained from the official portal of the Ministry of Primary and Secondary Education, which provides AN results in a national aggregate dataset. From this population, this study used a subset of student data that had complete numeracy variables and contextual factors, resulting in a final sample size of 45.304 students (2021), 45.480 students (2022), and 41.221 students (2023). Data selection was conducted using a purposive sampling technique based on the criteria for completeness of the numeracy domain data and contextual variables (see Table 1). Because the data were sourced from nationally standardized AN results, no additional weighting was applied. The sample distribution represents the proportion of students by province and school type as listed in the official AN dataset.

(ii) Data Cleaning and Processing: In this stage, data cleaning and processing were performed to ensure the quality of the dataset prior to analysis. Missing values in numeric variables and contextual factors were removed, while outliers were retained to capture natural variation in student achievement. Next, all numeric variables were standardized using the Z-score normalization method with *Scikit-learn's StandardScaler* to ensure each variable had a comparable scale and was ready for use in K-means clustering and correlation analyses.

(iii) Descriptive Statistical Analysis: This step aims to provide an overview of the distribution of student numeracy achievement. Descriptive analysis is conducted to calculate the mean, median, standard deviation, and distribution pattern. Data visualizations such as histograms, boxplots, and heatmaps are used to illustrate distribution patterns based on the existing variables.

(iv) Correlation Analysis Between Competency Domains: Correlation analysis is conducted to evaluate the relationship between competency domains in numeracy, such as the relationship between Algebra and Geometry, or between Number and Data and Uncertainty. This correlation provides insight into the interactions between domains, which can be used to develop strategies for improving numeracy achievement.

(v) Clustering: This step aims to group schools based on patterns of student numeracy achievement. The clustering stage was conducted using the K-Means clustering algorithm to group schools based on student numeracy achievement patterns. The optimal number of clusters was determined using the Davies–Bouldin Index (DBI) to ensure a balance between homogeneity within clusters and heterogeneity between clusters.

(vi) Data Visualization: Research results are presented in intuitive visualizations, such as graphs or graphical reports. This visualization is designed to facilitate interpretation by policy makers and other stakeholders.

Table 1. Variables in Numeracy Data

Code	Declaration
kd_sekolah	School/Education Unit Code
kd_kokab	District/City Code
wilayah_bagian	Regional Area Based on Time Zone Division
kd_siswa_an	Student Code
NUM	Numeracy Skill Score
NUM_ALJ	Algebra Competency Score
NUM_GEO	Geometry Competency Score
NUM_BIL	Numeracy Score
NUM_DAT	Data and Uncertainty Competency Score
NUM_L1	Knowing Competency Score L1
NUM_L2	Applying Competency Score L2
NUM_L3	Reasoning Competency Score L3
PNUM	Numeracy Learning

RESULTS AND DISCUSSION

Descriptive Analysis

Descriptive analysis was used to obtain an overview of the processed data. Table 2 shows that more than 97% of the processed data for each year was still valid after data cleaning. The initial data obtained from the official portal of the Ministry of Primary and Secondary Education was then cleaned to remove empty datasets and missing values.

Table 2. Numeracy Population Data 2021-2023

Year	Initial Data	Clean Data	Percentage
2021	45.304	44.216	97,60
2022	45.480	44.917	98,76
2023	41.221	40.791	98,96

Data cleaning is a crucial step in data analysis because the quality of the data used significantly impacts the final results. In the context of this research, the Numeracy data obtained from the National Assessment (AN) is secondary data that may contain incomplete datasets (missing values), extreme values (outliers), and non-uniform variable scales.

Figure 2 further displays a boxplot of outlier detection from the 2021-2023 data. Outliers are values or observations that deviate significantly from the general pattern of the data. In educational data such as numeracy scores, outliers can arise from measurement errors, incorrect data entry, or simply reflect extremely low or extremely high student performance. Detecting and handling outliers is crucial because these extreme values can affect statistics such as the mean and standard deviation, and impact the performance of algorithms like K-Means Clustering, which is highly sensitive to extreme values due to its reliance on Euclidean distance calculations.

In an educational context, outliers must be analyzed carefully. For example, students with very high numeracy scores may be excellent students, not just erroneous data. Conversely, very low scores may be students with special needs or challenging geographic conditions. Therefore, the decision to remove outliers should not be based solely on statistical analysis but should also consider the substantive context and student metadata. In big data-driven research, outliers can be important signals of systemic inequality.

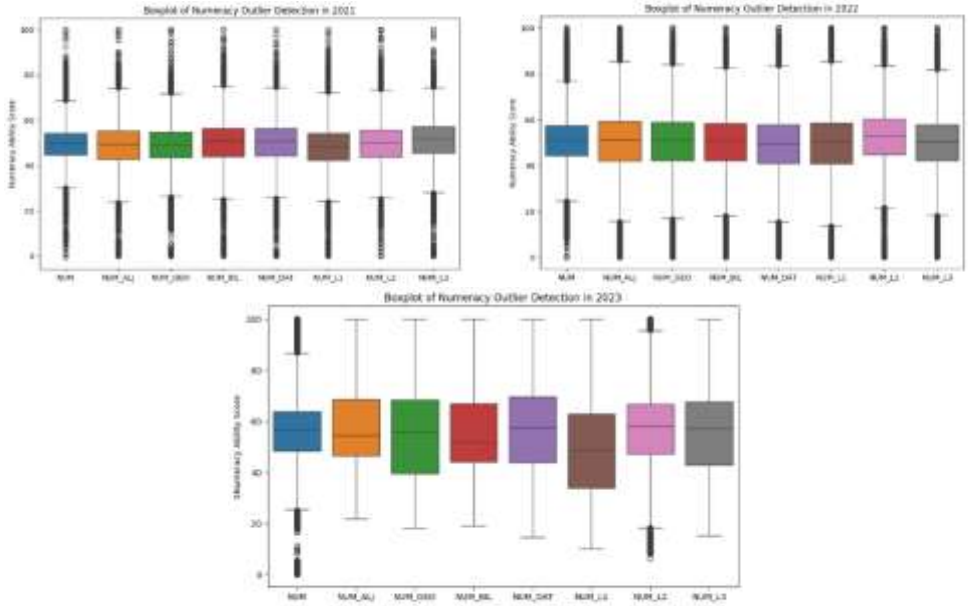


Figure 2. Boxplot of Outlier Detection for 2021-2023 Data

Boxplot visualization results for eight numeracy variables (NUM, NUM_ALJ, NUM_GEO, NUM_BIL, NUM_DAT, NUM_L1, NUM_L2, and NUM_L3) show that the data have a fairly uniform distribution, with median values ranging from 45 to 55. This indicates that the majority of students are at an intermediate level of numeracy ability. However, in each boxplot, there are several points that fall outside the lower and upper limits (whiskers), indicating the presence of outliers. These points represent students with very low scores (less than 20) or very high scores (more than 80), and they appear in almost all numeracy domains.

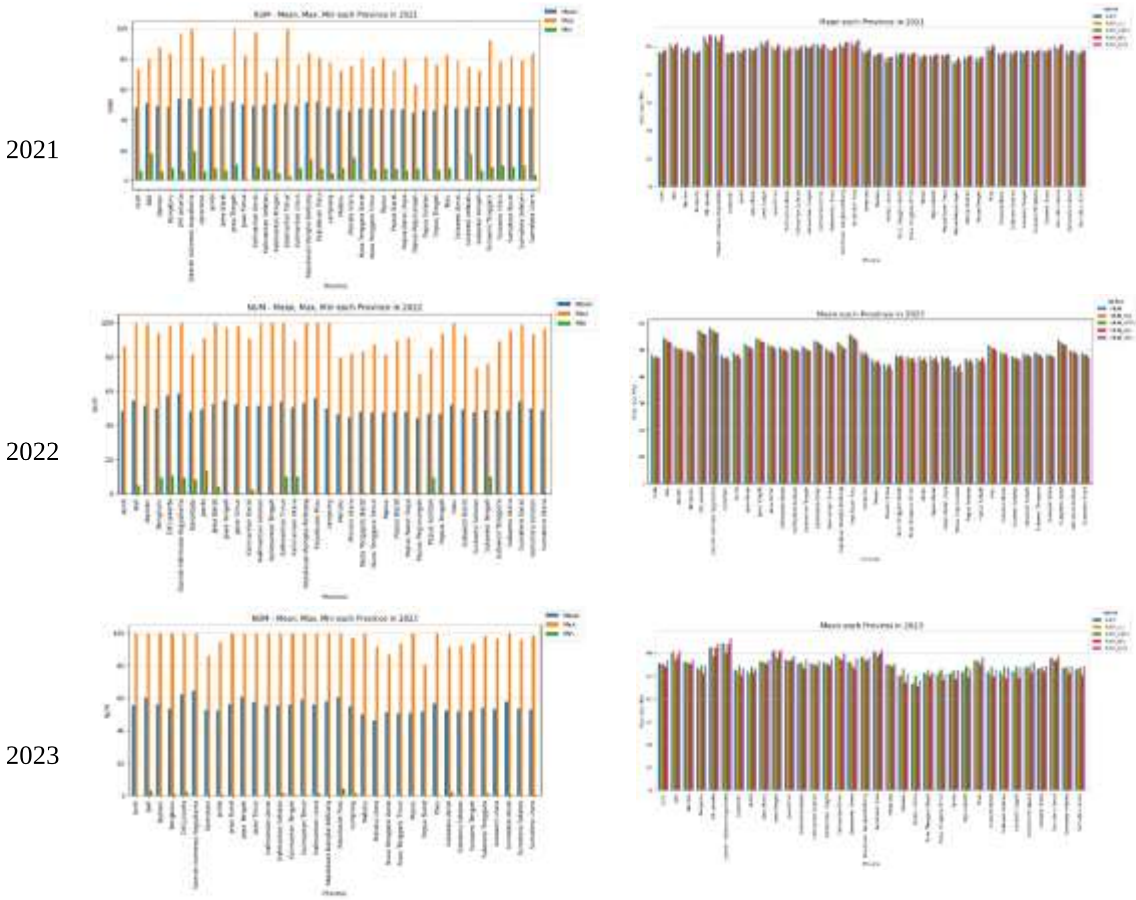
The presence of outliers like these in educational data is normal, as in the context of large assessments like the National Assessment (AN), variation in ability between students is highly likely. According to Osborne (2013), in empirical data, outliers often appear as part of natural variation and not simply as errors. Therefore, outliers need to be understood not simply as "statistical noise," but also as sources of important information—for example, students with special needs, extreme school conditions, or exceptional learning success.

Although statistically outliers can affect mean scores and analytical models, in this context, outlier removal is not necessary, as these values are viewed as the result of valid measurements. In Exploratory Data Analysis, boxplots are designed to identify outliers, but do not necessarily suggest that they should be removed. Handling outliers must take into

https://dx.doi.org/10.26594/jmpm.v10i2.5972

account the context of the data, the purpose of the analysis, and the type of decisions that will be taken based on the data.

In the context of educational research, it's best to identify outliers and analyze them separately, if necessary. If the goal of the analysis is to build a predictive model (e.g., regression), researchers may consider data transformations (such as logarithmic or robust scaling) or parallel analyses with and without outliers to test model stability. However, if the goal of the analysis is descriptive or to understand the distribution of student abilities, the presence of outliers enriches the narrative and should not be ignored.



a) Min, Max and Mean Numeration Values b) Mean value of each domain

Figure 3. Descriptive Statistics Values each Province

Figure 3 presents the descriptive statistical values for each province, illustrating the mean, median, standard deviation, and overall distribution patterns of students' numeracy achievement across Indonesia. These descriptive results provide an initial overview of how students' numeracy performance varies regionally and serve as a foundation for deeper analysis. Since numeracy skills encompass multiple aspects, they are divided into several key domains such as algebra, geometry, number, and data and uncertainty, which are measured through the cognitive levels of knowing (L1), applying (L2), and reasoning (L3). Understanding how these domains contribute to overall numeracy achievement (NUM) requires both descriptive and correlational analyses. Such analyses help reveal underlying distribution patterns and inter-domain relationships, providing a stronger empirical basis for formulating evidence-based learning recommendations. In this regard, data structure mapping, as noted by Han et al. (2012), represents the initial stage in the data mining process and plays a crucial role in determining the direction of subsequent analyses.

Descriptive analysis includes basic statistical calculations such as mean, median, standard deviation, minimum, and maximum values for each numeracy domain. For example, if the NUM_BIL domain shows a high mean but a large standard deviation, this indicates a gap in students' mastery of number concepts. Visualizations such as histograms or boxplots can be used to determine whether the distribution of each domain is approximately normal or skewed. Variations between domains have a significant impact on overall numeracy outcomes and can inform pedagogical interventions (Montshiwa & Botlhoko, 2023).

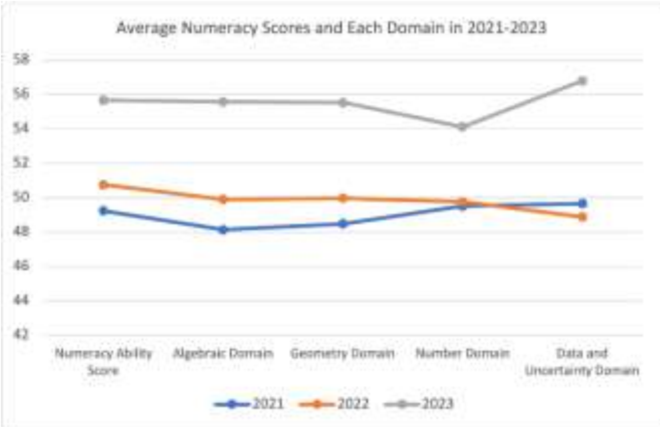


Figure 4. Graph of Average Numeracy Values for 2021-2023

Based on Figure 4 and Table 3, the analysis results show a fairly consistent upward trend in overall numeracy competency achievement. The numeracy proficiency score increased from 49,233 in 2021 to 50,761 in 2022, and reached 55,668 in 2023. This increase reflects improvements in students' mastery of numeracy concepts at the national level and indicates that numeracy learning practices and policies in high schools have improved over the past three years. This indicates that learning interventions, assessments, and curriculum approaches are beginning to have a positive impact on the quality of students' numeracy competency.

Table 3. Average Numeracy Scores and Each Domain in 2021-2023

Aspects observed	2021	2022	2023
Numeracy Ability Score	49,233	50,761	55,668
Algebraic Domain	48,138	49,906	55,573
Geometry Domain	48,485	49,965	55,534
Number Domain	49,517	49,755	54,117
Data Domain and Uncertainty	49,662	48,878	56,794

Furthermore, when analyzed by competency domain, the Data and Uncertainty domain showed the highest average score in 2023, at 56,794. Since 2021, this domain has consistently held the top position or near the top compared to other domains.

This fact supports previous correlation results that this domain has the largest contribution to total numeracy achievement (NUM). This trend aligns with global priorities in 21st-century literacy, which emphasize the importance of reading data, interpreting graphs, and making evidence-based decisions (OECD, 2021). Therefore, contextual, data-based learning should remain a primary focus in high school mathematics instruction.

Meanwhile, the Algebra and Geometry domains showed stable and nearly parallel growth over the three years. Both increased from around 48 in 2021 to over 55 in 2023. This indicates that students are increasingly able to understand symbolic structures and visual representations in mathematics. This increase also reflects the success of the conceptual learning approach that has been implemented more widely in schools. Furthermore, the stability of scores in these two domains can be used as a foundation for strengthening other domains, such as reasoning and modeling in the context of numeracy.

However, it's worth noting that the Numbers domain showed a slightly different trend. After increasing from 49,52 in 2021 to 49,75 in 2022, this domain's score only rose to 54,12 in 2023. While still improving, the growth rate was slower than other domains, and it ranked lowest among them in the last year. This is noteworthy, considering that numbers are the foundation of almost all other mathematical concepts. Lagging behind in this domain can impact students' ability to understand and solve more complex numeracy problems.

Overall, this data trend provides important insights for developing strategies to improve the quality of numeracy education. Significant improvements in the Data and Uncertainty domains need to be maintained and used as a basis for developing cross-domain learning models. Conversely, the Numbers domain requires special attention, both through strengthening basic concepts and using a more contextual and adaptive learning approach. By understanding the dynamics of numeracy achievement in each domain, policymakers and teachers can design interventions that are more targeted, systematic, and have a long-term impact on students' numeracy literacy in Indonesia.

In numeracy assessment, the cognitive level refers to the depth of thinking measured through numeracy questions. Based on the minimum competency assessment framework (AKM), the cognitive level is divided into three, including: 1) Knowing (level 1): related to the ability to recognize information, understand facts, and perform basic calculations, 2) applying (level 2): related to students' ability to use mathematical knowledge in contextual situations and solve routine problems, and 3) reasoning (level 3): related to the ability to think critically, make estimates, construct logical arguments, and solve non-routine problems. These levels reflect the hierarchy of thinking from the most basic to the most complex, and are used to measure the depth of students' understanding and numeracy skills.

Based on Table 4 and Figure 5, there is a significant increase in the ability to apply and reason in 2023 compared to previous years. These results indicate that high school students are increasingly developing higher-order thinking skills (HOTS), especially in applying and reasoning mathematical concepts in real-world contexts. This is a positive achievement because the reasoning domain is an indicator of student readiness to face the challenges of the 21st century.

Table 4. Average Cognitive Numeracy Level Score 2021-2023

Numeracy Cognitive Level	2021	2022	2023
Ability to knowing	47,655	49,128	49,957
Ability to applying	49,263	52,055	55,976
Ability to reasoning	50,557	49,452	55,282

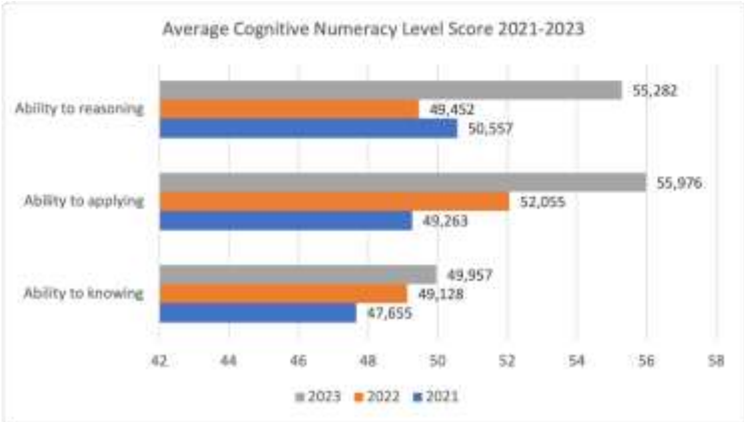


Figure 5. Average Cognitive Numeracy Level Score 2021-2023

Improved scores on the ability to apply and reason also indicate that current numeracy learning is moving beyond memorization and basic calculations and toward contextual, functional, and analytical approaches. This could be due to various factors such as problem-based learning, numeracy projects, realistic, and contextual learning. Conversely, stagnation at the knowledge level may indicate that basic skills such as symbol recognition or basic operations are still not fully developed across students.

Correlation Between Domains

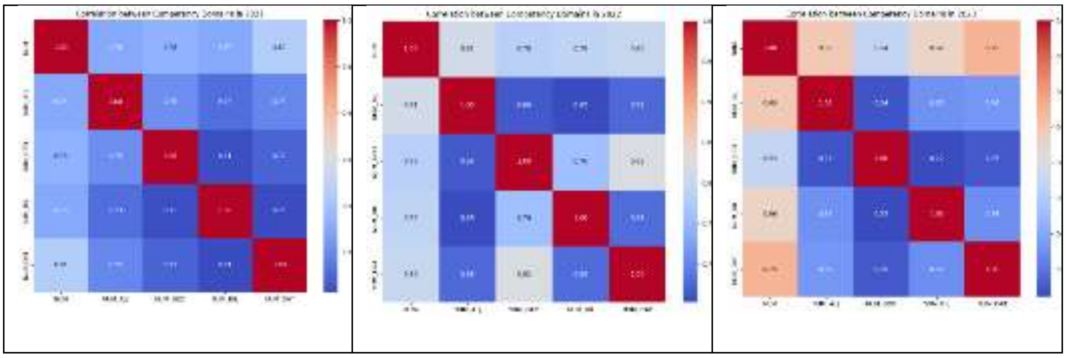


Figure 6. Correlation Between Competency Domains in Each Year (2021-2023)

Figure 6 illustrates the correlation between numeracy competency domains for high school students from 2021 to 2023. Based on the heatmap visualization of correlations in 2021, it appears that all domains have a positive and fairly strong relationship with each other, with correlation values ranging from 0.71 to 0.81. The highest correlation occurred between the total numeracy score (NUM) and the Data and Uncertainty domain (NUM_DAT) at 0,81, indicating that students' ability to understand and interpret data has a significant impact on overall numeracy achievement. This aligns with the 21st-century numeracy literacy framework, which emphasizes the importance of reading graphs and tables, and making data-based inferences in everyday life (OECD, 2021).

In addition to NUM_DAT, other domains such as Algebra (NUM_ALJ), Geometry (NUM_GEO), and Number (NUM_BIL) also showed a fairly high correlation with NUM, at 0,78. This indicates that mastery of various basic mathematics domains remains important, although their contribution to the total score is not as significant as that of the data-based domain. The relatively balanced correlations between these domains reflect that numeracy learning approaches cannot be focused on a single area but must be developed comprehensively. Students who excel in one domain tend to excel in other domains as well.

However, the lowest correlation was recorded between Geometry (NUM_GEO) and Number (NUM_BIL), at 0,71. While still strong, this can be interpreted as meaning that the spatial and visual skills honed in geometry are more loosely related to the numeracy or number manipulation abilities found in the number domain. The differences in cognitive characteristics required by each of these domains may explain why the relationship between them is not as strong as in other domains. This is also in line with findings showing that connectivity between numeracy domains can differ depending on the cognitive context and type of problem (Montshiwa & Botlhoko, 2023).

Overall, these findings provide important insights into developing numeracy learning strategies. The fact that NUM_DAT has the highest correlation with the total numeracy score suggests that learning interventions focused on the context of data and uncertainty can be an effective strategy for improving students' numeracy literacy. Furthermore, strengthening integration between domains through contextual and cross-skill approaches can also enhance the interconnectedness of student competencies. By understanding these correlation patterns, teachers and policymakers can develop more targeted interventions based on the most strategic domains for improvement.

The correlation analysis between numeracy domains in 2022 shows a slightly different pattern of relationships compared to the previous year. The highest correlation remains for the total numeracy score (NUM) with the Data and Uncertainty domain (NUM_DAT), at 0,85, even higher than in 2021 (0,81). This reaffirms the importance of mastering data literacy in contributing to students' overall numeracy scores. Data literacy, which includes the ability to read graphs, tables and interpret contextual quantitative information, is increasingly becoming a key indicator in measuring 21st century numeracy competence (OECD, 2021)

Interestingly, in 2022, the correlation between NUM and NUM_GEO was recorded as low compared to other domains, at only 0,76, suggesting that the geometry problems presented likely did not contribute significantly to the total numeracy score. In fact, the relationship between NUM_GEO and other domains was relatively weaker, for example, only 0,65 against NUM_ALJ and 0,72 against NUM_DAT. This indicates that spatial or visual abilities measured in geometry tend to be more independent than other, more numerical or data-based domains. This should be taken into account in curriculum designs that integrate visual and numerical approaches more harmoniously.

Meanwhile, the relationship between NUM_ALJ and NUM remained high (0,81), as did NUM_BIL with NUM (0,79). This suggests that algebra and number remain important pillars in students' numeracy structure, but their position is beginning to be displaced by the role of data literacy. The correlation between NUM_ALJ and NUM_DAT was also high (0,80), indicating synergy between symbolic-formal (algebraic) understanding and data-based understanding. This finding supports a numeracy learning approach that connects symbolic thinking skills to real-life, data-based contexts, such as social and economic issues.

In general, correlations between domains in 2022 remained positive and quite strong, although more varied than in the previous year. The weakest correlations were recorded between NUM_ALJ and NUM_GEO (0,65), and between NUM_GEO and NUM_BIL (0,71). This indicates that geometry is the domain with the lowest connectivity to other numeracy domains, and a more integrative learning approach is needed to prevent this domain from developing in isolation. It is crucial for numeracy learning to connect various domains contextually and transdisciplinary, rather than being taught in isolation.

These results suggest that strengthening data literacy remains a priority for numeracy education interventions, given its consistently high contribution to the total numeracy score. Furthermore, the geometry domain requires special attention, both in terms of question content in national assessments and in classroom learning approaches. Project-based modules or visual data modeling could be alternative strategies to bridge the gap between geometry and other numeracy domains. Thus, this correlation analysis not only provides an overview of the relationships between domains but can also be used to develop more comprehensive and evidence-based strategies for improving numeracy learning.

The correlation analysis of the numeracy domains of high school students in 2023 shows a significant decrease in the strength of the relationships between domains compared to the previous two years. The highest correlation between the total numeracy score (NUM) and the NUM_DAT (Data and Uncertainty) domain was only 0.75, lower than in 2022 (0.85) and 2021 (0.81). This decrease indicates that the contributions of each domain to the total numeracy score have become more separate or unintegrated. This could reflect changes in test design, a shift in curriculum focus, or possibly fragmented learning in the field.

Furthermore, the correlation between NUM_ALJ (Algebra) and other domains decreased drastically, reaching only 0.24 with NUM_GEO and 0.36 with NUM_DAT. In fact, the correlation between NUM_GEO and NUM_BIL was only 0.22, indicating that visual and spatial competencies in geometry appear to develop quite separately from basic numerical and data-based skills. This phenomenon also reflects findings Montshiwa & Botlhoko (2023), that emphasize that separation between domains can reduce the effectiveness of predictive models in educational data mining and complicate efforts to identify key domains in educational assessment.

This low correlation also indicates that students may perform highly in one domain but very poorly in another, without a strong pattern of consistency. This situation certainly poses a serious challenge for curriculum designers and mathematics teachers because it reflects a lack of integration in learning approaches. However, according to the OECD (2021) assessment framework, 21st-century numeracy literacy demands cross-domain skills, where competencies such as algebraic reasoning, spatial representation, and data analysis are interconnected to comprehensively solve real-world problems.

This situation also requires attention in policy analysis. Strong correlations between domains indicate internal coherence between assessment and learning systems, while weak correlations indicate disconnects between learning topics. To bridge this gap, a transdisciplinary pedagogical approach is needed that connects mathematical contexts with the real world and across domains. For example, statistics learning can be integrated with algebraic and geometric models through spatial data-based projects.

From a research and assessment development perspective, the 2023 results demonstrate the importance of rethinking the design of numeracy instruments. Too sharp a separation between domains in assessments can reduce the model's ability to identify holistic patterns of student proficiency. In other words, the decline in correlation is not simply a statistical figure, but a reflection of the disconnect between teaching approaches and numeracy measurement. Further research is needed to explore whether this is due to assessment policies, technical factors, or weak integration within the school curriculum.

Clustering

Clustering is used to group schools/students into homogeneous sets based on numeracy achievement patterns, thus maximizing the distinction between sets. The chosen approach is K-Means because it is efficient for similarly scaled numerical data and works by minimizing the sum of the squares of the distances to the cluster centers (centroids). Recent studies highlight its effectiveness in educational data analysis and performance profiling (Jain, 2010). In practice, determining the optimal number of clusters (k) is not an a priori decision but is evaluated using validity indices such as the Davies–Bouldin or silhouette score (Xu & Tian, 2015).

The index we use is the Davies–Bouldin Index (DBI), which is the average “worst case similarity ratio” between clusters, combining compactness within clusters (the dispersion of members relative to the centroid) and separation between centroids. The smaller the DBI, the better the clustering quality, as the clusters are more compact and more distant.

Table 5 shows that 2021 and 2022 both achieved the minimum values at k=7, 0.536 (2021) and 0.395 (2022), respectively; the lowest in the entire table). This means that the data structure for these two years is best mapped into seven segments: each segment is relatively compact, while the distance between segments is quite wide. For 2022 in particular, the very low DBI indicates a very clear separation between groups—an ideal situation for designing differentiated learning interventions per segment (e.g., a segment strong on data-uncertainty but weak on numbers, and vice versa).

In contrast, 2023 showed a minimum DBI at k=3 (DBI= 0,534). This finding implies that the 2023 numeracy achievement landscape is structured into three large clusters, which adequately represent population variation. Substantively, this can occur when the distribution of scores across domains is more "clustered" into three general patterns—although the linear relationship between domains may weaken. Consequently, the 2023 mapping strategy is more appropriate starting at the macro level (three large profiles). It can be further deepened into sub-segments based on additional analyses (e.g., Silhouette, centroid inspection, or domain-specific analysis).

https://dx.doi.org/10.26594/jmpm.v10i2.5972

Conceptually, differences in the optimal number of clusters between years confirm that the structure of population heterogeneity changes over time, possibly due to changes in assessment design, curriculum implementation, or the dynamics of learning practices. Based on this, for 2021–2022, seven segments were used as the basis for clustering. For 2023, three segments were used as the basis for clustering. This approach is consistent with the principles of evidence-based mapping in educational data mining and learning analytics (Liu, 2022; Yel et al., 2024).

Table 5. DBI Result

Cluster	DBI		
	2021	2022	2023
2	0,643	0,5751	0,613
3	0,595	0,613	0,534
4	0,6721	0,582	0,665
5	0,594	0,503	0,689
6	0,576	0,444	0,658
7	0,536	0,395	0,633
8	0,559	0,426	0,581

2021 Clustering Results

Based on the 2021 DBI values in Table 5, the 2021 K-Means Clustering Map is divided into seven clusters. Figure 7 shows the clustering based on similarity in numeracy profiles (a combination of domain scores and cognitive levels). Visually, a color gradation is visible, from blue (C1–C3) to gray (C4–C5) to orange/red (C6–C7). This clustering can be seen as a sequence from relatively lower to relatively higher. The resulting pattern is similar to the eastern Indonesian cluster (Papua–Maluku–Nusa Tenggara) where many cluster in the blue cluster, while some parts of the west/south (parts of Sumatra, Java–Bali, and Kalimantan) are more frequently found in the neutral to warm cluster (gray–orange/red). This pattern shows a regional gradient in numeracy achievement that is consistent with variations in access and learning ecosystems (teacher availability, facilities, connectivity) that are often discussed in the literature on educational equity.

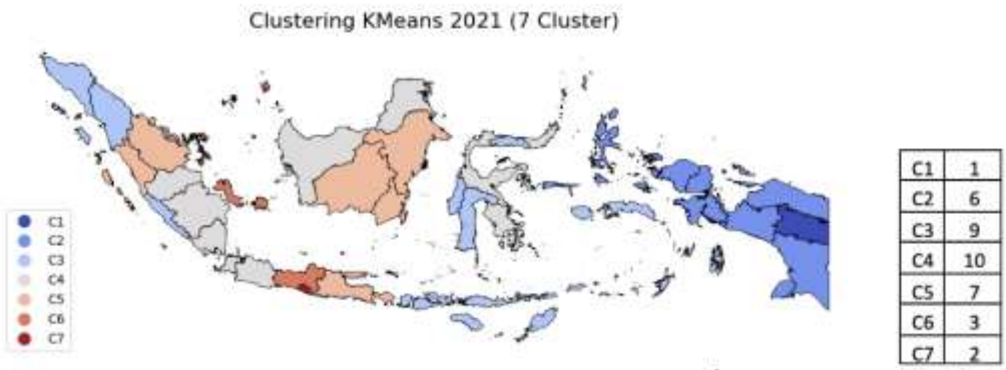


Figure 7. Clustering Results in 2021

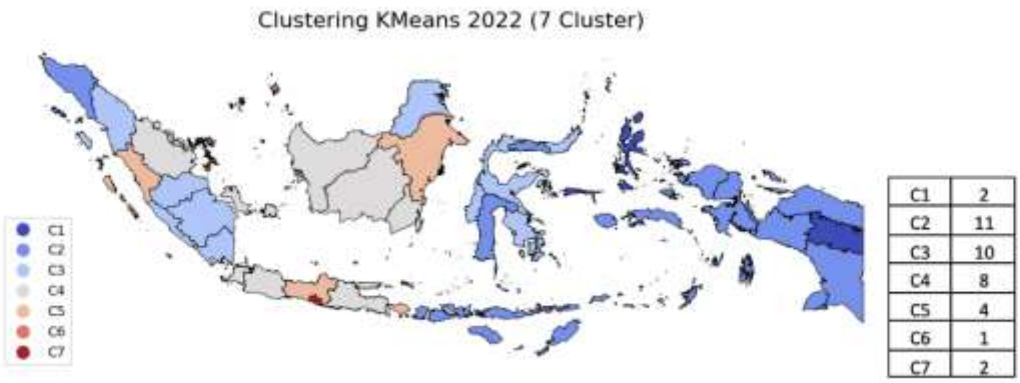
The cluster frequency distribution is also informative. C4 dominates (10 provinces), followed by C3 (9) and C5 (7); while C1 (1), C7 (2), and C6 (3) have fewer provinces. This means that most provinces fall within the middle (C3–C5) range, with

https://dx.doi.org/10.26594/jmpm.v10i2.5972

similar profiles, while very few provinces are truly extreme—either lagging behind (C1) or advancing (C7). For policymaking, this means that a national numeracy improvement strategy is most effective if it targets the large middle segment (C3–C5) with generic interventions (e.g., strengthening data literacy and reasoning), while extreme provinces require specific intervention packages: for the low cluster (C1/C2), focus on strengthening numeracy prerequisites and supporting infrastructure/teacher training; for the high cluster (C6/C7), encourage HOTS enrichment, cross-domain projects, and inter-regional benchmarking.

2022 Clustering Results

In 2022, seven clusters were used for clustering. This year's selection was based on the lowest DBI value across all scenarios (≈0.395). This indicates the clearest separation between clusters. Therefore, this seven-cluster configuration represents the most informative structure of variation in numeracy achievement between regions for formulating interventions.



C1	2
C2	11
C3	10
C4	8
C5	4
C6	1
C7	2

Figure 8. Clustering Results in 2022

Spatially, the 2022 pattern still shows a similar regional gradient to 2021. Eastern Indonesia (Papua–Maluku–Nusa Tenggara) is predominantly clustered in the blue clusters (C1–C3), while parts of the western/central regions are more frequently found in gray to orange/red (C4–C7). When read as a gradation from relatively lower (C1) to higher (C7), the map indicates that spatial disparities in numeracy achievement remain, but the separation in 2022 is more pronounced (in line with the very low DBI). However, the low-high determination needs to be confirmed by reading the centroid of each cluster (mean domain and cognitive level) for accurate interpretation.

In terms of membership distribution, the majority of provinces are in C2 (11 provinces) and C3 (10 provinces), followed by C4 (8 provinces). The extreme clusters have few members: C1 (2), C6 (1), C7 (2), while C5 (4) acts as a transitional group. This means that the 2022 landscape is concentrated in the middle segment (C2–C4) with similar characteristics, while only a few provinces are very lagging or very advanced. This composition is ideal for a broad-scale improvement strategy in the middle segment, while preparing specific intervention packages for the extreme segment (intensive remedial for

C1; enrichment/HOTS and cross-domain practices for C7).

To ensure accuracy, the 2022 results need to be cross-validated with other internal indicators (e.g., Silhouette) and contextual factors (SES, school district status, teacher ratio, connectivity). This validation ensures that differences between clusters are not merely technical artifacts, but rather reflect real differences in numeracy competency profiles that are relevant to be followed up at the policy and learning practice levels.

2023 Clustering Results

Based on Figure 9, three clusters were used in 2023. The selection of k = 3 in 2023 is consistent with the previous DBI evaluation. This indicates that the structure of variation in numeracy achievement between regions in 2023 is good, with only three large clusters. Compositionally, C1 comprises 16 provinces, C2 comprises 13 provinces, and C3 comprises only five provinces. The dominance of C1–C2 indicates that the majority of provinces are clustered in two relatively similar main profiles, while C3 represents pockets with the most divergent characteristics.

Figure 9. Clustering Results in 2023

Spatially (based on the coloring of the map), there is visible consolidation of the area. One cluster (C1) is spread across many provinces, one intermediate cluster (C2) fills the gaps between regions, and a small number of provinces are clustered in a contrasting cluster (C3). This suggests a narrowing of the extreme gap. More provinces are moving toward the intermediate profile, while only a few are in the most divergent profile (C3). The shift from seven clusters (2021–2022) to three clusters (2023) suggests a simplification of the variability landscape. This could be due to a more homogeneous score distribution, changes in the weighting of cognitive domains/levels in the instrument, or more uniform learning practices in the field. Overall, the 2023 clustering results suggest that macro-segmentation (three profiles) is sufficient to map national needs.

CONCLUSION AND SUGGESTIONS

This study maps the numeracy achievement of high school students in Indonesia using a data mining approach through the K-Means clustering algorithm and correlation analysis on the 2021–2023 National Assessment data. The results reveal three main findings: (1) there is a consistent increase in national numeracy scores, with the Data and

https://dx.doi.org/10.26594/jmpm.v10i2.5972

Uncertainty domain being the strongest contributor to the total score, underscoring the importance of data literacy in the context of 21st-century numeracy; (2) significant improvements in the cognitive levels of application and reasoning indicate a shift in learning practices toward a contextual problem-solving approach; and (3) changes in the structure of heterogeneity across regions indicate a consolidation of achievement patterns from seven clusters (2021–2022) to three main clusters (2023) based on the DBI index.

These findings not only broaden the empirical understanding of variations in numeracy achievement in Indonesia but also provide a conceptual contribution to the educational data mining literature by demonstrating that clustering techniques can be used to identify spatial and cognitive patterns in student numeracy performance. Furthermore, the results of this study confirm the relevance of applying large-scale assessment analytics to designing adaptive and differentiated education policies. Practically, this research emphasizes the need for learning strategies that balance strengthening data literacy and developing higher-order reasoning skills according to regional cluster profiles, to support sustainable and equitable numeracy improvement.

ACKNOWLEDGMENT.

The author expresses his deepest appreciation and gratitude to the Institute for Research and Community Service (LPPM) of Sanata Dharma University for funding support through the 2025 Internal Research Grant. This support is very meaningful in the implementation and completion of this research.

REFERENCES

Acosta, S. T., & Hsu, H. Y. (2014). Negotiating diversity: an empirical investigation into family, school and student factors influencing New Zealand adolescents' science literacy. *Educational Studies*, *40*(1), 98–115. https://doi.org/10.1080/03055698.2013.830243.

Alvarez-Garcia, M., Arenas-Parra, M., & Ibar-Alonso, R. (2024). Uncovering student profiles. An explainable cluster analysis approach to PISA 2022. *Computers and Education*, *223*, 105166. https://doi.org/10.1016/j.compedu.2024.105166.

Barbeiro, L., Gomes, A., Correia, F. B., & Bernardino, J. (2024). A review of educational data mining trends. *Procedia Computer Science*, *237*(4), 88–95. https://doi.org/10.1016/j.procs.2024.05.083.

Gamazo, A., Martínez-Abad, F., Olmos-Migueláñez, S., & Rodríguez-Conde, M. J. (2017). Evaluación de factores relacionados con la eficacia escolar en PISA 2015: Un análisis multinivel. *Revista de Educacion*, 379, 56–78. https://doi.org/10.4438/1988-592X-RE-2017-379-369.

Gamazo, A., Olmos-Migueláñez, S., & Martínez-Abad, F. (2016). Multilevel models for the assessment of school effectiveness using PISA scores. *TEEM '16: Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp.1161–1166). Association for Computing Machinery. https://doi.org/10.1145/3012430.3012663.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques (3rd Edition)*. Morgan Kaufmann Publishers/Elsevier.

Huang, Y., Zhou, Y., Chen, J., & Wu, D. (2024). Applying machine learning and SHAP method to identify key influences on middle-school students' mathematics literacy performance. *Journal of Intelligence*, *12*(10), 93.

https://doi.org/10.3390/jintelligence12100093.
Jain, A. K. (2010). Data clustering: 50 years beyond K-Means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.
Jornet Meliá, J. M. (2016). Methodological analysis of the PISA project as international assessment. *RELIEVE - Revista Electronica de Investigacion y Evaluacion Educativa*, *22*(1), 1-25. https://doi.org/10.7203/relieve22.1.8293.
Kalita, E., Oyelere, S. S., Gaftandzhieva, S., Rajesh, K. N. V. P. S., Jagatheesaperumal, S. K., Mohamed, A., Elbarawy, Y. M., Desuky, A. S., Hussain, S., Cifci, M. A., Theodorou, P., Hilčenko, S., Hazarika, J., & Ali, T. (2025). Educational data mining: a 10-year review. *Discover Computing*, *28,* Article 81. Springer Science and Business Media B.V. https://doi.org/10.1007/s10791-025-09589-z.
Lingard, B., Martino, W., & Rezai-Rashti, G. (2013). Testing regimes, accountabilities and education policy: Commensurate global and national developments. *Journal of Education Policy*, *28*(5), 539–556. https://doi.org/10.1080/02680939.2013.820042.
Liu, R. (2022). Data analysis of educational evaluation using K-means clustering method. *Computational Intelligence and Neuroscience, 2022,* 3762431. https://doi.org/10.1155/2022/3762431
Malik, S., Patro, S. G. K., Mahanty, C., Hegde, R., Naveed, Q. N., Lasisi, A., Buradi, A., Emma, A. F., & Kraiem, N. (2025). Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. *Scientific Reports*, *15*(1), 8738. https://doi.org/10.1038/s41598-025-92324-x.
Ministry of Primary and Secondary Education. (2023). *Portal data pendidikan: Kementerian pendidikan dasar dan menengah*. https://data.kemdikdasmen.go.id.
Montshiwa, T. V., & Botlhoko, T. (2023). Modelling and predicting learners' numeracy test results using some regression and machine learning classifiers. *Journal of Statistics Applications and Probability*, *12*(3), 1345–1363. https://doi.org/10.18576/jsap/120337.
OECD. (2021). *The assessment frameworks for cycle 2 of the programme for the international assessment of adult competencies*. OECD Publishing. https://doi.org/10.1787/4bc2342d-en.
Osborne, J.W. (2013). *Best practices in data cleaning: a complete guide to everything you need to do before and after collecting your data*. SAGE Publications.
Ouassif, K., Ziani, B., Herrera-Tapia, J., & Kerrache, C. A. (2025). Empowering education: leveraging clustering and recommendations for enhanced student insights. *Education Sciences*, *15*(7), 819. https://doi.org/10.3390/educsci15070819.
Sarker, S., Paul, M. K., Thasin, S. T. H., & Hasan, M. A. M. (2024). Analyzing students' academic performance using educational data mining. *Computers and Education: Artificial Intelligence*, *7,* 100263. https://doi.org/10.1016/j.caeai.2024.100263.
Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1.
Yel, M.B., Efendi, S., Zarlis, M., & Suwilo, S. (2024). Clustering diagnostic assessment of students with the k-means algorithm based on talents and interests. *Proceedings of the 3rd International Conference on Advanced Information Scientific Development (ICAISD 2023)* (295–301). SCITEPRESS – Science and Technology Publications, Lda. https://doi.org/10.5220/0012457200003848.
Zhu, Y., & Kaiser, G. (2020). Do east asian migrant students perform equally well in mathematics?. *International Journal of Science and Mathematics Education*, *18*, 1127–1147. https://doi.org/10.1007/s10763-019-10014-3.