Optimizing School Resource Allocation in Lima Puluh Kota Regency through Random Forest Classification of Clustered Student Enrollment Trends

Lestari Margatama[1*], Yopie Hidayat[2], and Indra Riyanto[3]

[1*] Center for Remote Sensing Studies, Universitas Budi Luhur, Jakarta 12260, Indonesia
[2] Faculty of Information Technology, Universitas Budi Luhur, Jakarta 12260, Indonesia
[3] Faculty of Engineering, Universitas Budi Luhur, Jakarta 12260, Indonesia
indra.riyanto@budiluhur.ac.id

Abstract. The pattern of changes in student numbers in public elementary schools is a strategic issue that affects the effective distribution of educational resources, particularly in Lima Puluh Kota Regency. Imbalances in student growth or decline across schools can impact the need for teaching staff, classroom space, and budget allocation. This study aims to classify public elementary schools based on patterns of student number changes, as an effort to support data-driven education policy planning on 359 public elementary schools with student number data from year 2020-2024, as well as supporting attributes. The method used in this study is K-Means clustering and Random Forest classification. Clustering resulted in four clusters: increasing, stable, moderately decreasing, and sharply decreasing student numbers. The classification models were evaluated using K-Fold Cross-Validation with accuracy, precision, recall, and F1-score metrics. The results showed that the pure student-based classification produced an average accuracy of 91.9%, while the trend-based model had an accuracy of 39%. Although trend-based model currently had a lower accuracy, its clustering results were considered more relevant for long-term policy because they were able to illustrate the direction of student change. Descriptive analysis of each cluster also showed a link between the decline in student numbers and low school resource allocation. These findings can be used to support a more targeted and adaptive distribution of educational resources.

Keywords: Clustering, Classification, Student Population, School Change Patterns, K-Means, Random Forest.

1 Introduction

The number of students in elementary schools is often used as an indicator to assess educational development in a region, particularly in the context of meeting the need for equitable basic education. Data from the Ministry of Education, Culture, Research, and Technology shows that there has been a change in the number of students in various elementary schools in Indonesia in recent years [1]. Lima Puluh Kota Regency in West

N. A. Ishak et al. (eds.), *Proceedings of the International Conference on Cross- Disciplinary Academic Research 2025 - Track 1 Advances in Computing, Electronics, Engineering, and Mathematics (ICAR-T1 2025)*, Advances in Engineering Research 296,
https://doi.org/10.2991/978-94-6239-636-4_17

Sumatra, one of the areas with many public elementary schools, has shown varying changes in student numbers across several schools in recent years. This is important to analyze, given that changes in student numbers can impact educational planning, such as the allocation of teachers and other educational resources [2]. If not analyzed properly, these changes can result in disparities in student numbers between schools, overcapacity in certain schools, or mergers with other schools due to student shortages.

A comprehensive understanding of student enrolment changes is crucial to help identify the various challenges facing the education system. By understanding the patterns of student enrolment changes in each school, educational resource allocation planning can be optimized to anticipate potential challenges. Understanding these patterns allows for more precise and sustainable education planning at the regional level. However, to date, analyses have been manual and have not utilized more reliable technology to more accurately understand student enrolment patterns. Data-driven analysis, however, allows local governments to formulate more appropriate policies for managing public services, such as determining intervention strategies for farms experiencing declining trends in the US or allocating resources more efficiently to for safer environment in China [3, 4].

The highly variable student enrolment across elementary schools in Lima Puluh Kota Regency poses a significant challenge to educational planning and resource allocation. Inaccurately detecting student enrolment changes can lead to unequal resource distribution, such as overcapacity in some schools or insufficient facilities in others. If not addressed appropriately, inaccurate educational planning can lead to wasted resources or the neglect of schools that require more attention. Therefore, this study aims to propose a school classification method using the Random Forest method based on K-Means clustering data to identify cluster patterns in student enrolment changes in elementary schools. By focusing on analyzing student enrolment patterns, this study provide a more accurate and comprehensive approach to understanding the dynamics of student enrolment changes, which in turn can serve as a basis for more effective and efficient decision-making for education policymakers.

2 Literature review

The development of data mining and machine learning technology, particularly classification and clustering algorithms, offers new potential in educational data analysis. Algorithms such as Random Forest are widely used in processing complex data due to their accuracy in data classification [5]. On the other hand, the K-Means algorithm is a data mining algorithm that can be used for grouping/clustering data. Using K-Means Clustering, student data can be grouped based on certain patterns [6]. This algorithm is effective in grouping data with diverse characteristics, which will serve as the basis for classifying patterns of changes in the number of students in each school cluster. Random Forest has the advantage of producing relatively good and efficient classification performance on large data sets, but the results are still influenced by the structure and distribution of the data being analyzed [7].

Several previous studies have explored the use of data mining and machine learning algorithms for educational data analysis, with interesting and useful results [8-11]. Okoye et al. demonstrated that the Decision Tree algorithm is effective in determining student dropout factors with high accuracy, where the main influencing factors are GPA, faculty, study time, and predicate [12]. Meanwhile, predicted graduation rates for school students using the Random Forest algorithm, demonstrating high accuracy, reaching 100%, and effectively predicting on-time graduation [13]. These models has been proven to be able to identify students who graduate and those who fail, and makes a positive contribution to the development of evaluation and intervention in education.

2.1 K-Means Clustering

Clustering is a process used to group data into several clusters or groups that share certain similarities. In this process, data with similar characteristics will be grouped into one cluster, so that the data within the cluster is expected to have high similarity. Conversely, data placed in different clusters should have low differences or similarities with each other [14]. Clustering results are considered optimal if each cluster shows a high level of similarity between its members, but has a low similarity with members of other clusters. To determine this level of similarity, a numerical calculation is performed between two data objects, which helps evaluate the similarities or differences in characteristics between them [15].

K-Means is a non-hierarchical clustering method that begins by selecting several components from the data population as initial cluster centers. The initial cluster centers are randomly selected from the entire data. The K-Means algorithm then examines each component in the population and assigns it to a specific cluster center based on the minimum distance between that component and each existing cluster. The cluster centers are then updated repeatedly until all data is grouped into their respective clusters, resulting in new, stable cluster center positions [16].

K-Means is an iterative algorithm designed to minimize variation within each cluster. The clustering process involves grouping the data into k mutually exclusive and non-overlapping subgroups, with each subgroup formed based on a cluster center (centroid) calculated from the average position of the cluster members [17]. The measure of similarity or dissimilarity between data in K-Means is generally measured using Euclidean distance, although other distance metrics can be used depending on the needs.

In its implementation, one of the main challenges in the K-Means algorithm is determining the optimal number of clusters (k). Generally, the value of k is determined based on expert knowledge or historical information from the data. However, in many cases, such approaches are not always available, making selecting the appropriate value of k a crucial issue. Some approaches use metrics such as inter-cluster distance and intra-cluster distance to assess the quality of the clustering results. This approach can be used to evaluate various values of k, although the process can require repeated execution of K-Means with a complexity of $O(n)$, which is inefficient for large datasets [18].

The proximity between data and the cluster center is determined by the distance between the objects. This distance is calculated using the Euclidean distance formula, which can be formulated as follows:

$$d_i = \sqrt{\sum_i^p (x_{2i} - x_{1i})^2} \quad (1)$$

where:
d_i = distance between two points
p = dimension of data
x_{2i}= training data
x_{1i}= test data

In general, the K-Means algorithm aims to minimize the sum of the squared distances between each data point and its centroid. This process involves assigning each data point to the nearest centroid, then recalculating the centroid based on the average of all points within that cluster. This process is repeated until there is no significant change in the centroid position, or until the objective function value remains unchanged [19].

Over time, numerous studies have attempted to improve the performance of K-Means, particularly in terms of time efficiency, result accuracy, and distance metric selection. Thus, although K-Means is known as a simple and fast algorithm, the complexity of its practical implementation is significantly influenced by parameter selection, including the appropriate distance metric, outlier handling mechanisms, and centroid initialization. Recent research has shown that with modifications or integration of other methods, the K-Means algorithm can be significantly improved in terms of accuracy and efficiency [20].

2.2 Random Forest Classification

Classification is the process of creating a model that can represent a concept or data category, with the primary goal of predicting the label or class of an unknown object. The resulting model can be an "if-then" rule, a decision tree, a mathematical equation, or even an artificial neural network. The classification process defined by essential components such as *Class*, which is a categorical dependent variable that represents the object's label; and *Predictor* as an independent variable that describes the characteristics or attributes of the data.

Random Forest algorithm is a refinement of the Decision Tree classification method. By constructing multiple decision trees, Random Forest randomly selects data subsets and feature subsets for each tree, creating independence between trees and improving the model's generalization ability. This algorithm was developed from the bagging (bootstrap aggregating) approach and belongs to the group of decision tree-based algorithms.

In Fig. 1, Random Forest creates multiple trees, collectively referred to as a "forest." A larger number of trees generally results in higher classification accuracy. In the process of predicting new data, the data is fed into each of the formed decision trees. Each

tree produces its classification results, and the final class of the data is determined based on a majority vote from all trees. This process is repeated for all trees in the Random Forest, and the final vote results become the class prediction for that sample [21].

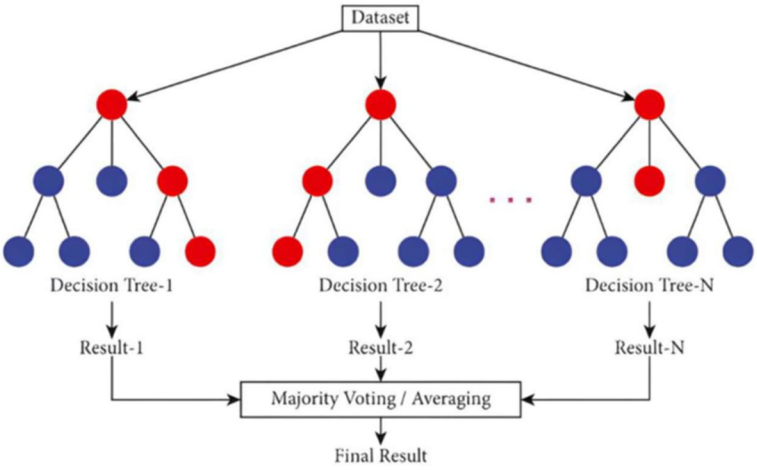

Fig. 1. Random Forest illustration

Villegas-Mier et al. identified three key parameters to consider in the tuning process: Number of trees (*n_estimators*), Maximum number of features per split (*max_features*), and Maximum tree depth (*maximum_depth*) [22]. More trees typically result in better accuracy, as the variation in results between trees can balance the overall predictions. However, too many trees will increase computation time and can reduce model efficiency. Therefore, it is necessary to find the critical point between the optimal number of trees between accuracy and training time.

Max Feature parameter determines how many features are considered for each node split. Selecting different feature subsets between trees minimizes correlation between them and allows for a more accurate evaluation of the contribution of each feature. Too large a value can lead to overly complex models, while too small a value can lead to underfitting. Trees with very high depths are prone to overfitting because they overfit the training data. Conversely, trees with too shallow a depth may not capture complex patterns in the data.

With proper tuning, Random Forest performance can be significantly improved, both in terms of accuracy and computational efficiency, compared to using default parameters from commonly used programming libraries [22]. Beyond parameter tuning, Random Forest has other characteristics that make it superior in various classification and regression tasks. Each tree in a Random Forest is built independently based on a randomly selected subset of data and features (random subspace method). These are then combined using majority voting for classification or averaging for regression, producing final predictions that are more stable and robust against overfitting [23].

This ensemble process can utilize parallel training, making it computationally efficient, with a time complexity linear to the number of trees and samples (O(n)). Random Forest also has internal mechanisms that support feature importance measurement, handling missing values, and validating model performance without external validation data using the Out-Of-Bag Error method [22, 24].

Random Forest's effectiveness is further enhanced by its ability to automatically and incrementally select important features. Features that don't significantly contribute to accuracy can be eliminated from the training process to improve time efficiency and reduce training bias [21].

3 Proposed Model

The main objective of this stage is to evaluate the extent to which internal school attributes, such as the number of teachers, classes, study groups, budget, and classroom conditions, can explain the pattern of changes in student numbers that have been established through the clustering process. This study also applies the K-Fold Cross-Validation method to improve model accuracy and reliability by enabling systematic model testing on multiple data subsets [25]. The model will be trained and validated using the K-Fold Cross Validation technique to measure performance based on evaluation metrics:

- n_estimators =100
- max_depth determined automatically based on the data structure.
- min_samples_split and min_samples_leaf allowing flexibility in the formation of nodes and leaves in the decision tree.
- criterion uses 'gini' as an impurity measure to determine the quality of the split at each node.
- random_state to maintain consistency and reproducibility of the experimental results.

This study uses the data on the number of students in each public elementary school in Lima Puluh Kota Regency, obtained from the Ministry of Education and Culture's Education Data [1]. This data covers the number of students from Grade 1 to Grade 6 for each school year from 2020 to 2024. Overall, this dataset covers 362 public elementary schools spread across 13 sub-districts in Lima Puluh Kota Regency. Data collection was carried out by accessing and downloading data from the *Dapodik* platform, which provides information on the number of students enrolled in each public elementary school. Data on the number of public elementary schools in Lima Puluh Kota Regency can be seen in Table 1.

Table 1. Dataset main attribute.

Attribute	Type	Description
Nama_Sekolah	String	The name of a public elementary school as a unique identity
Jumlah_Siswa_2020	Numeric	Number of students from grades 1 to 6 in 2020
Jumlah_Siswa_2021	Numeric	Number of students from grades 1 to 6 in 2021
Jumlah_Siswa_2022	Numeric	Number of students from grades 1 to 6 in 2022
Jumlah_Siswa_2023	Numeric	Number of students from grades 1 to 6 in 2023
Jumlah_Siswa_2024	Numeric	Number of students from grades 1 to 6 in 2024

The risk of overfitting can still be minimized by applying cross validation, furthermore, to enrich the analysis and improve the model's accuracy in identifying patterns of student enrolment change, several additional internal school attributes were added to the dataset. These additional attributes include number of teachers, number of classes, and number of study groups (*rombel*), annual school budget, and class feasibility. These additional attributes aim to provide a more comprehensive perspective on school conditions and characteristics, which may influence patterns of student enrolment change over time. The structure of the additional internal school dataset used in this study can be seen in Table 2.

Table 2. Additional attributes.

Attribute	Type	Description
Jumlah_Guru	Numeric	Number of teachers in school
Jumlah_Kelas	Numeric	Number of classrooms in the school
Jumlah_Rombel	Numeric	Number of active study groups
Kondisi_Kelayakan_Kelas	Numeric	Percentage of suitable classroom

4 Results

This study conducted clustering with an approach focused on analysing the trend patterns of changes in student numbers over the past five years. Prior to clustering, all student data from 2020 to 2024 was normalized using the Z-Score Standardization method. After normalization, the trend value, or slope, was calculated using a linear regression function for each school. The Davies-Bouldin Index (DBI) value provides an indication of how well clusters are separated and how compact they are. The lower the DBI value, the better the clustering results. The results of this evaluation can be seen in detail in Table 3.

Table 3. Davis-Bouldin Index.

Cluster	Davies-Bouldin Index (DBI)
2	0.7751
3	0.5981
4	***0.5464***
5	0.5487

The lowest DBI value occurred at k = 4 with a value of 0.5464, which shows that at this number of clusters, the model produces the clearest separation between clusters and the highest internal consistency.

A descriptive analysis was conducted to determine the characteristics of school resources in each cluster. This analysis covered four main attributes: number of teachers, number of classes, number of study groups, and percentage of classroom occupancy. To clarify the characteristics of each cluster, the analysis results are presented in a table of averages per cluster as shown in Table 4.

Table 4. Dataset main attribute.

Cluster	Teachers	Classrooms	Classes	Suitability (%)
0	10.5	7,1	7,2	96,7
1	9,0	6,6	6,3	95,0
2	9,3	6,8	6,3	97,8
3	9,8	8,1	6,7	96,4

Cluster 3 shows the highest average teacher ratio, namely 9.8 teachers, reflecting the condition of schools experiencing a decrease in the number of students and the potential for a reduction in the number of teachers. Cluster 3 has the largest number of classes (8.1), indicating that schools in this cluster generally have a larger classroom capacity. This is consistent with schools experiencing a decline in student enrolment, thus requiring fewer active classrooms. Conversely, Cluster 1 has the smallest number of classrooms, at 6.6, likely because these schools are able to adapt to the declining student population. The classification result with the first nine of the schools are shown in Table 5.

Table 5. Samples of school classification result

School Name	No. of Teachers	No. of Classrooms	No. of Classes	Classification
SD Negeri 01 Batu Balang	11	8	6	Severe Decline
SD Negeri 01 Bukik Limbuku	9	6	6	Increasing
SD Negeri 01 Gurun	9	6	6	Moderate Decline
SD Negeri 01 Harau	10	6	6	Normal
SD Negeri 01 Koto Tuo	14	13	11	Moderate Decline
SD Negeri 01 Lubuak Batingkok	8	7	6	Stable
SD Negeri 01 Pilubang	13	7	7	Stable

SD Negeri 01 Sarilamak	19	11	13	Increasing
SD Negeri 01 Solok Bio-Bio	9	6	6	Severe Decline

The distribution of the number of schools in each cluster shows that of the total 359 schools, 33 schools saw an increase in student enrollment numbers; 158 schools in a relatively stable number of students over the years; 128 schools experienced a moderate decline; and 40 of the schools showing a severe decline.

5 Conclusion

The classification results of public elementary schools in Lima Puluh Kota Regency based on changes in student enrollment trends over the past five years (2020–2024) identifies 168 schools are declining in the number of students, of which, 40 schools are severely losing significant number of students over the period. The output from this model can be used to provide a more comprehensive teacher resources optimalization between schools.

Acknowledgments. The authors would like to express thanks to Universitas Budi Luhur through Directorate of Research and Community Service for the support in the publication of this article.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. *Indonesian Education Statistics*. 2022, Ministry of Education and Culture.
2. *Education Statistics of Lima Puluh Kota Regency*. 2024.
3. Ghosh, R., *Data-driven governance and performances of accountability: critical reflections from US agri-environmental policy.* Science as Culture, 2024. **33**(1): p. 70-96.
4. Liu, J., et al., *Building safer and more resilient cities in China: A novel approach using a dynamic nonhomogeneous Gray model for data-driven decision-making.* PLoS One, 2024. **19**(12).
5. Han, J., J. Pei, and H. Tong, *Data mining: concepts and techniques*. 2022: Morgan kaufmann.
6. Jayasree, R. and S. Selvakumari, *Design of a Prediction Model to Predict Students' Performance Using Educational Data Mining and Machine Learning.* Engineering Proceedings, 2023. **59**(1): p. 25.
7. Sivakumar, S. and S. Venkataraman, *Evaluating Machine Learning Approaches: A Comparative Study of Random Forest and Neural Networks in Grade Classification.* Indonesian Journal of Data and Science, 2025. **6**(1): p. 73-80.
8. Tarik, A., H. Aissa, and F. Yousef, *Artificial Intelligence and Machine Learning to Predict Student Performance during the COVID-19.* Procedia Computer Science, 2021. **184**: p. 835-840.

9. Retno, D.M., et al. *Comparative Analysis of Machine Learning Techniques Student Final Grade Classification Accuracy*. in *2024 International Conference on Smart Computing, IoT and Machine Learning (SIML)*. 2024.
10. Çetinkaya, A., Ö.K. Baykan, and H. Kırgız, *Analysis of Machine Learning Classification Approaches for Predicting Students' Programming Aptitude.* Sustainability, 2023. **15**(17): p. 12917.
11. Pallathadka, H., et al., *Classification and prediction of student performance data using various machine learning algorithms.* Materials Today: Proceedings, 2023. **80**: p. 3782-3785.
12. Okoye, K., et al., *Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education.* Computers and Education: Artificial Intelligence, 2024. **6**: p. 100205.
13. Gul, M.N., et al., *Data driven decisions in education using a comprehensive machine learning framework for student performance prediction.* Discover Computing, 2025. **28**(1): p. 153.
14. Tibay, J.P., S.C. Ambat, and A.C. Lagman. *School-Based Management Performance Efficiency Modeling and Profiling using Data Envelopment Analysis and K-Means Clustering Algorithm*. 2019. IEEE.
15. Liu, R., *Data Analysis of Educational Evaluation Using K-Means Clustering Method.* Computational intelligence and neuroscience, 2022. **2022**: p. 1-10.
16. Fan, X., et al. *Cluster Analysis of Effect of Social Support Using K-Means Algorithm on Learning Motivation of Primary and Secondary School Students*. 2024. IEEE.
17. Miraftabzadeh, S.M., et al., *K-means and Alternative Clustering Methods in Modern Power Systems.* IEEE access, 2023. **11**: p. 1-1.
18. Punhani, A., et al., *Binning-based Silhouette Approach to Find the Optimal Cluster using K-Means.* IEEE access, 2022. **10**: p. 1-1.
19. Chong, B., *K-means clustering algorithm: a brief review.* Academic Journal of Computing & Information Science, 2021. **4**(5): p. 37-40.
20. Ghazal, T.M., et al., *Performances of k-means clustering algorithm with different distance metrics (With Full Text).* Intelligent Automation & Soft Computing, 2021. **30**(2): p. 735-742.
21. Paul, A., et al., *Improved Random Forest for Classification.* IEEE transactions on image processing, 2018. **27**(8): p. 4012-4024.
22. Villegas-Mier, C., et al., *Optimized Random Forest for Solar Radiation Prediction Using Sunshine Hours.* Micromachines (Basel), 2022. **13**(9): p. 1406.
23. Wang, X., et al., *Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier.* BMC medical informatics and decision making, 2021. **21**(1): p. 105-14.
24. Salman, H.A., A. Kalakech, and A. Steiti, *Random forest algorithm overview.* Babylonian Journal of Machine Learning, 2024. **2024**: p. 69-79.
25. Brownlee, J., *Machine Learning Mastery with Python*. 6 ed. 2021.