JOURNAL OF INTERDISCIPLINARY PERSPECTIVES

ISSN Print: 2984-8288, ISSN Online: 2984-8385
Vol. 4 No. 5, pp. 314-323, May 2026

Original Article

Enrollment Trends in Philippine Public and Private Basic Education Schools: A Regional Comparative Trend Analysis

Philip John L. Paja

Author Information:

Graduate School Department, University of Immaculate Conception, Davao City, Philippines

Correspondence:
pajaphilipjohn@gmail.com

Article History:

Date received: February 18, 2026
Date revised: April 22, 2026
Date accepted: May 1, 2026

Recommended citation:

Paja, P.J. (2026). Enrollment trends in Philippine public and private basic education schools: A regional comparative trend analysis. *Journal of Interdisciplinary Perspectives*, 4(5), 314-323.
https://doi.org/10.69569/jip.2026.080

Abstract. This study looks at enrollment trends in Philippine public and private basic education schools across different regions. It examines trends over ten academic years, from SY 2010-2011 to SY 2020-2021. The study aims to find the regions with the highest and lowest enrollment and to estimate the average annual growth rate of student enrollment. It addresses the need for a broad regional analysis that was missing from earlier research and provides data-based insights for educational planning. The dataset comes from the Department of Education (DepEd) and includes three main variables: academic year, educational sector (public or private), and administrative region. To look at trends and predict future enrollment, the study uses Random Forest, Gradient Boosting Machine (GBM), and Linear Regression models. These models were chosen because they can manage complex patterns and give reliable predictions. The Random Forest model did better than both Linear Regression and GBM, achieving an R^2 of 0.98 and a low RMSE for students in basic education. This shows how effective it is at identifying enrollment trends. The findings presented that Region IV-A (CALABARZON) has the highest enrollment, while CAR and BARMM have the lowest. The trends showed changes over the years, identifying our regional differences, and these results provide valuable information for policymakers and education planners. They can use this to make better decisions about resource distribution, program development, and education strategies in various regions. The study shows that data-driven analysis can clearly show enrollment trends and support data-based policy in the Philippines' basic education system.

Keywords: *Data mining techniques; Enrollment; Gradient Boosting model; Linear Regression model; Random Forest model.*

Fluctuating enrollment rates in Philippine basic education reveal persistent gaps in access across regions and between the public and private sectors. This is true even with free education programs and reforms like the K–12 program under Republic Act 10533 (Roman & Villanueva, 2018). Education is essential for national development, economic growth, and social mobility. Human Capital Theory, according to Wuttaphan (2017), sees education as an investment that boosts individual productivity and provides long-term economic benefits. From this view, investing in education improves future earning potential, workforce competitiveness, and national economic stability. In developing countries like the Philippines, education remains a key policy focus due to its direct impact on labor market readiness and socioeconomic progress.

Enrollment trends are important indicators of access, fairness, and the sustainability of institutions within the

education system. According to Tan (2017), changes in enrollment greatly affect government budgets, teacher placements, infrastructure planning, and long-term education strategies. Previous research has raised concerns about workforce readiness and educational competitiveness in the Philippines, despite relatively high participation rates. Additionally, studies (Dela Cruz, Reyes, Santos, & Gomez, 2020; Parveen, Shah, & Mahmood, 2020) on changes in enrollment patterns across different academic fields and regions show that various socioeconomic, demographic, and institutional factors shape student participation. These trends emphasize the need for careful tracking of enrollment behaviors.

With the growing availability of large educational datasets, data mining and machine learning methods have become tools for gaining insights from complex data. Educational Data Mining (EDM) uses statistical, computational, and predictive modeling techniques to analyze educational records and generate actionable knowledge (Rabelo et al., 2024). It was also supported by the study of (Sarker, 2021; Albreiki, Zaki, & Alashwal, 2021) that machine learning algorithms, in particular, have shown in the study of, strong predictive abilities in various educational areas, including performance prediction, dropout detection, and forecasting the outcomes for the institutions.

Many studies (Delima, Sison, & Medina, 2019; Haris, Abdullah, Hasim, & Rahman, 2016) have shown that predictive analytics can effectively model enrollment trends and institutional performance. In contrast, the studies by Zhou (2021) and Sharma & Kumar (2022) found that linear methods may be quite traditional, which is why it has been difficult to capture the nonlinear patterns found in educational datasets. This means that ensemble learning methods, such as Random Forests and Gradient Boosting Machines, can improve predictive accuracy by combining multiple decision models and reducing variance. These methods are especially useful for managing diverse, regionally distributed data.

Although previous research confirms the usefulness of predictive modeling in educational contexts, much of the literature focuses primarily on higher education institutions or specific academic programs. Few studies provide a broad regional comparison of elementary and secondary enrollment across the public and private sectors in the Philippine context. As a result, few studies, such as Albreiki et al. (2021) and Romero & Ventura (2020), systematically compare multiple machine learning techniques, such as Linear Regression, Random Forest, and Gradient Boosting Machines, to determine the most accurate model for national-level enrollment forecasting. This gap underscores the need for a comparative data analysis approach to evaluate model performance using standardized accuracy metrics. Addressing this research gap is essential for evidence-based policymaking. The accuracy of enrollment forecasting helps the education department to use resources efficiently, predict infrastructure needs, and create specific plans for areas experiencing decline or rapid growth. In resource-limited settings, accurate prediction is especially important for ensuring long-term sustainability and fair access to educational services.

Therefore, this study analyzes enrollment trends among elementary and secondary students in public and private schools across regions in the Philippines. Specifically, it found to: (1) determine the average annual growth rate of student enrollment; (2) identify regions with the highest enrollment concentrations across sectors; (3) compare the performance of Linear Regression, Random Forest, and Gradient Boosting Machine models in generating forecasts; and (4) determine the most efficient predictive technique based on established accuracy metrics. In this study, the results could contribute to the educational data mining literature and provide practical insights for policymakers, administrators, and stakeholders to strengthen strategic educational planning and resource allocation in the Philippines.

Methodology

Research Design

This study employs a quantitative predictive research design, using secondary data analysis and machine learning techniques, to visualize enrollment trends in the Philippine education system. A predictive analytics approach was selected because the study aims to identify patterns and generate forecasts from historical enrollment data, using Jupyter Notebook as the integrated development environment. This design is appropriate for trend analysis and model comparison, particularly when evaluating the performance of different machine learning algorithms in forecasting educational data.

Data Source

The dataset for this study came from publicly available school enrollment data published by the Department of Education and hosted on Kaggle. It covers student enrollment by region, school sector, and academic year from SY 2010 to SY 2021 (Raiblaze, 2022). The data is grouped by academic year, school sector (public or private), and administrative region in the Philippines. Since the study used secondary data, no participants or survey tools were involved. The dataset includes total regional enrollment figures and contains no personally identifiable information. The features were created during preprocessing to improve predictive performance and support precise modeling of enrollment trends.

Table 1. *Philippines School Enrollment Dataset (SY 2010–2011 to SY 2020–2021, Sourced from Kaggle)*

Variable	Description	Value Type	Notes
Academic Year	School year (SY 2010–2011 to SY 2020–2021)	Categorical/Time	Represents the academic period of enrollment data
School Sector	Public or Private	Categorical	Indicates whether enrollment is from public or private schools
Region	Administrative region in the Philippines	Categorical	Used to compare enrollment trends across regions
Enrollment Total	Number of enrolled students	Numerical	Main outcome variable used for trend analysis and forecasting
Year-over-Year Growth	Percentage change in enrollment compared to the previous year	Numerical	Derived feature created to identify trends and growth patterns
Public-to-Private Ratio	Ratio of public enrollment to private enrollment per region	Numerical	Derived feature to support comparative analysis between sectors

Data Preprocessing and Preparation

Data preprocessing was done to ensure data quality and prepare the model. Since the dataset had no missing values, no imputation procedures were necessary. These steps made sure the dataset was structured, consistent, and ready for predictive modeling. Figure 1 below represents the design architecture, and these preprocessing steps include:

Data Cleaning and Validation – Verification of data consistency, removal of duplicate entries (if any), and confirmation of correct data types.

Data Transformation – Categorical variables such as region and school sector were encoded using one-hot encoding to allow inclusion in machine learning models. Derived features (year-over-year growth and public-to-private ratio) were generated to improve model performance.

Data Integration – where applicable, involved standardizing variable names and merging records from DepEd using region and academic year as common identifiers.

Data Splitting – on the other hand, divided the final dataset into training (70%) and testing (30%) subsets. This helped evaluate how well the model would perform on new data and prevent overfitting. Additionally, 10-fold cross-validation was performed as a check, yielding similar predictive performance.

Predictive Modeling

Three machine learning models were implemented and compared: Linear Regression, Random Forest, and Gradient Boosting Machine (GBM). Linear Regression was included as a baseline statistical model for analysis and prediction. Random Forest and Gradient Boosting Machine were selected for their strong performance in handling nonlinear relationships and structured tabular data. Model training was conducted on the training dataset, and performance was evaluated on the test dataset. Accordingly, following Breiman (2001) and Géron (2019), Random Forest was selected for its strong predictive performance, robustness to overfitting, and effectiveness in modeling nonlinear relationships in structured tabular data. To support the findings of Hastie et al. (2009), which show that ensemble-based approaches consistently demonstrate superior performance to single linear models in complex regression tasks. Model performance was evaluated using RMSE and R^2, which are widely accepted metrics for assessing predictive accuracy and explanatory power in regression and forecasting studies (James et al., 2021; Hyndman & Athanasopoulos, 2021). Hyperparameter Tuning is also applied to the random forest to explore the number of trees, maximum depth, and minimum samples per leaf, using grid search. The GBM was tuned on learning rate, number of boosting rounds, and maximum depth to optimize predictive performance.

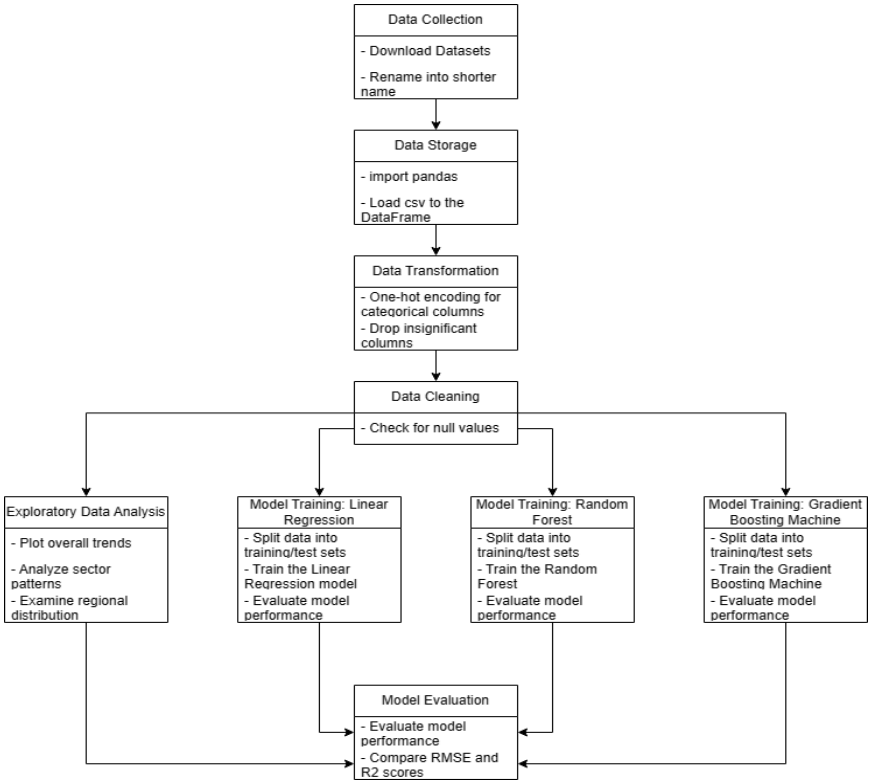


Figure 1. *Design Architecture*

Data Analysis Procedure

Descriptive statistical analysis was first performed to determine enrollment growth rates and identify the regions with the highest enrollment. For predictive evaluation, model performance was assessed using the following metrics: Root Mean Squared Error (RMSE) – measures the magnitude of prediction errors, penalizing large deviations; and R-squared (R^2) – measures the proportion of variance explained, indicating the model's explanatory power. These metrics were selected because they are standard measures in regression-based predictive modeling. The model, based on (James et al., 2021; Hyndman & Athanasopoulos, 2021), with the lowest RMSE and highest R^2 score, was considered the most efficient for predicting enrollment trends. RMSE and R^2 were selected because they are widely recognized as standard evaluation measures in regression and forecasting studies. RMSE is particularly sensitive to forecast accuracy because it penalizes large prediction errors, whereas R^2 provides an interpretable measure of explanatory power for comparative model assessment (Bishop, 2006). The model with the lowest RMSE and highest R^2 on the test dataset was considered the most accurate and efficient for forecasting enrollments.

Ethical Considerations

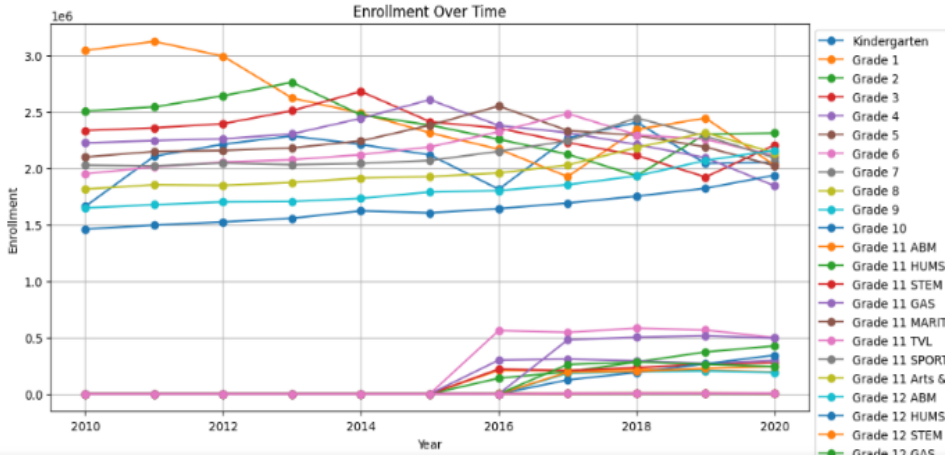
The present study used publicly available secondary data. The data used for this study did not involve any human participants. The dataset used for this study included aggregated enrollment data across various regions. The data set did not include any personal identifiers. The data set was used for academic purposes. The use of the dataset for academic purposes was ethical because the dataset's source was properly cited.

Results and Discussion

Descriptive Analysis of Enrollment Distribution

Figure 2 displays the distribution of student enrollment by region and education level. The histogram shows two clusters, indicating variation in student enrollment across regions and grade levels, including Kindergarten, Grades 1-6, Junior High School, and Senior High School. This implies that some regions have higher student enrollment than others. The differences in student enrollment across regions stem from structural, demographic,

and economic factors. Regions with extensive urban development and economic activity have more students enrolled, while isolated areas have fewer. This is consistent with the demographic characteristics of developing economies, where metropolitan regions experience high demand for education due to population influx (Parveen, Shah, & Mahmood, 2020). Figure 2 provides an overview of the baseline characteristics of the regional disparities, suggesting that predictive modeling techniques, such as temporal and spatial variation, were used.



F

Figure 2. *Distribution Across Enrollment Counts*

Regional Enrollment Comparison

As shown in Figure 3, enrollment levels vary across regions. Region IV-A (CALABARZON) led with about 3.3 million students, which is roughly 16% of the national total. NCR follows with 2.4 million students, making up 12%, and Region III has 2.2 million students, or 11%. At the low end, the Cordillera Administrative Region (CAR) enrolls around 450,000 students, which is 2%, while BARMM has about 1.1 million students, or 5%.

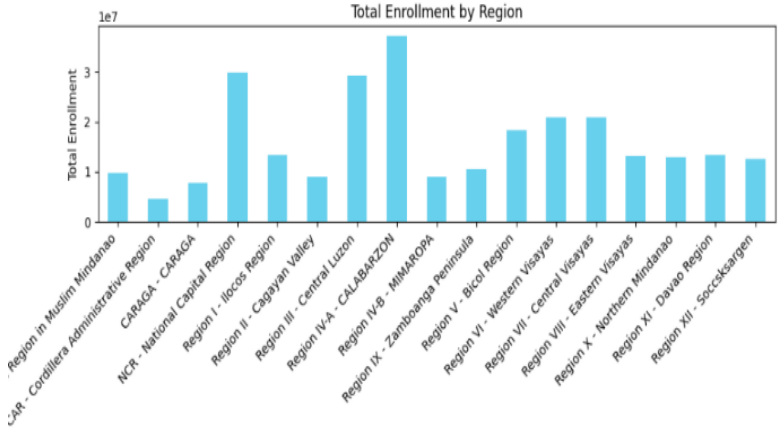


Figure 3. *Total Enrollment per Region*

The prominence of CALABARZON, NCR, and Central Luzon reflects wider demographic and economic factors. Highly urbanized areas tend to attract larger populations and exhibit higher educational participation. This trend mirrors the earlier findings by Tan (2017), which connect urbanization to increased demand for education. The distribution of students also shows an imbalance: public schools enroll nearly 1.75 times as many students as private institutions. This indicates a heavy reliance on public education within the system. Figure 4 shows that most enrollments happened in the public sector; this imbalance affects predictive modeling. While class imbalance can skew machine learning models toward the majority classes, no resampling was done here to maintain the real-world distribution. This choice supports the validity of the findings, though it is recognized as a methodological limitation.

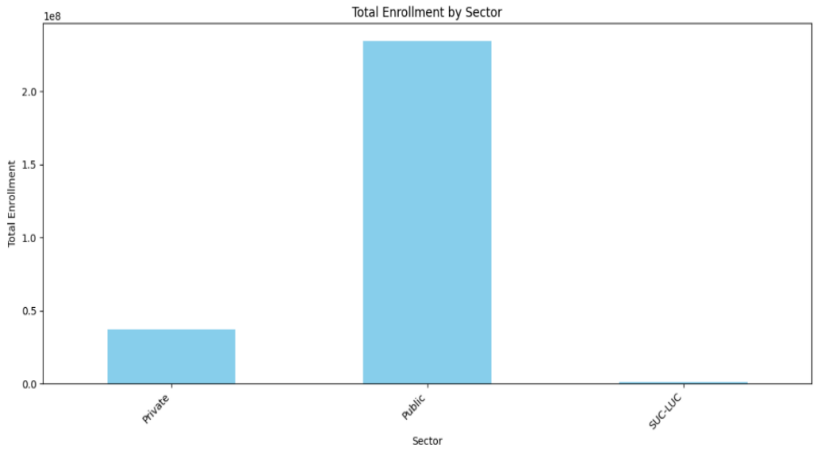


Figure 4. *Total Enrollment per Sector*

Predictive Model Performance

To evaluate forecast precision, Linear Regression, Random Forest, and Gradient Boosting Machine were used. The models' performance was evaluated using two key parameters: R-squared, which measures predictive power, and root mean squared error, which quantifies prediction error.

Linear Regression

Figure 5, predicted and actual enrollment using Linear Regression. The model explains a reasonable amount of the variance, but it consistently underestimates enrollment in areas with higher student density. The RMSE indicates a moderate level of error, while the R^2 suggests fair overall explanatory power. This pattern of underestimation shows a common limitation of Linear Regression when dealing with nonlinear relationships, especially in complex datasets (Sarker, 2021). While the method is easy to interpret, its predictive flexibility is limited when applied to different educational data.

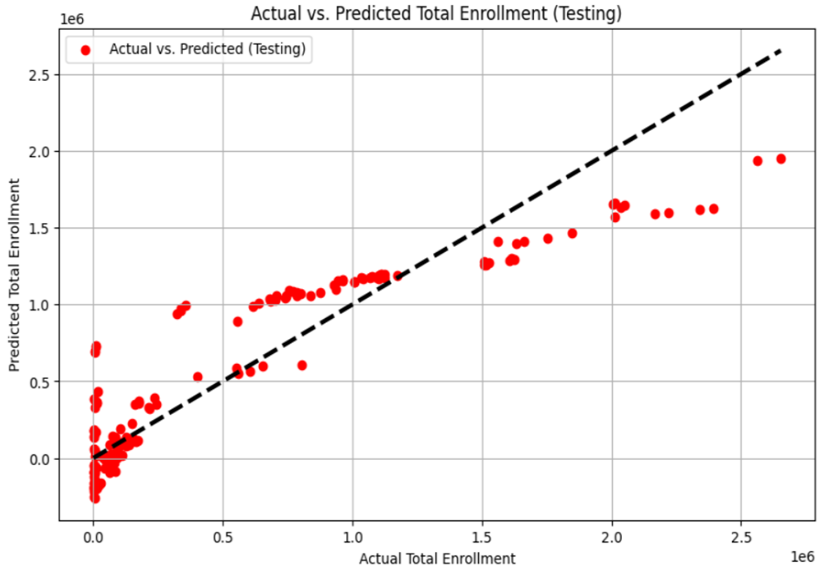


Figure 5. *Linear Regression Output*

Random Forest

As shown in Figure 6, the Random Forest predictions align closely with the ideal fit, with most points falling exactly where they should. Compared with Linear Regression, RF produced a higher R^2 and a lower RMSE,

showing that it handles enrollment differences across regions more effectively. By limiting variance, as in the studies by Zhou (2021) and Sharma & Kumar (2022), and thereby avoiding overfitting, the model captured nonlinear patterns that LR missed. Nearly 90% of the predictions clustered near the perfect line, a result that highlights the stability of Random Forest and its value for structured educational data

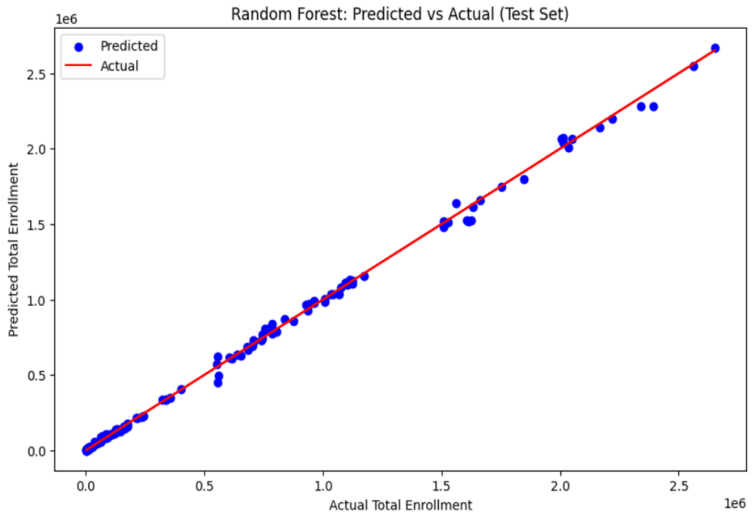


Figure 6. *Random Forest Output*

Gradient Boosting Machine (GBM)

Figure 7 demonstrates the similarity between real-world enrollment patterns and the predictions of the Gradient Boosting Machine (GBM). Compared with the Linear Regression method, the GBM showed slightly better performance and was on par with Random Forest. The step-by-step learning process of GBM helps reduce residual prediction errors. The good performance reported by Rabelo et al. (2024) for GBM demonstrates the high precision of boosting algorithms in nonlinear prediction.

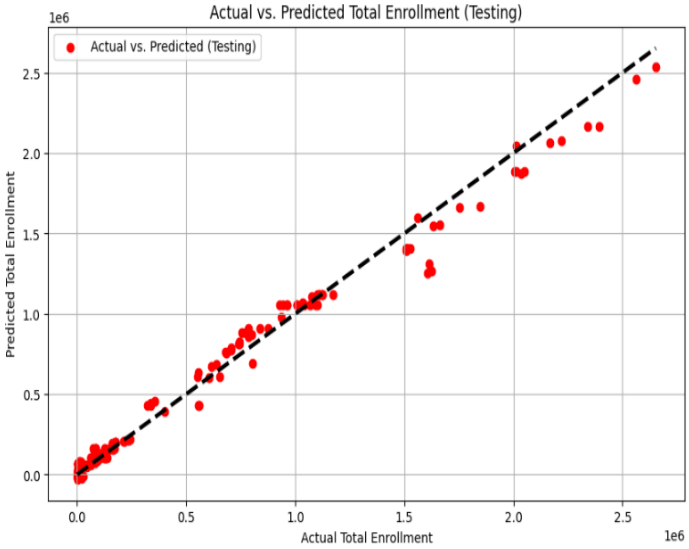


Figure 7. *Gradient Boosting Regression Output*

Comparative Evaluation of Models

When comparing the three models in Figures 8 and 9 using R2 and RMSE, Random Forest and Gradient Boosting outperformed Linear Regression. They showed better predictive accuracy and reduced bias. The importance of these methods in educational data would be apparent, especially where enrollment is non-linear and varies across

geographic regions.

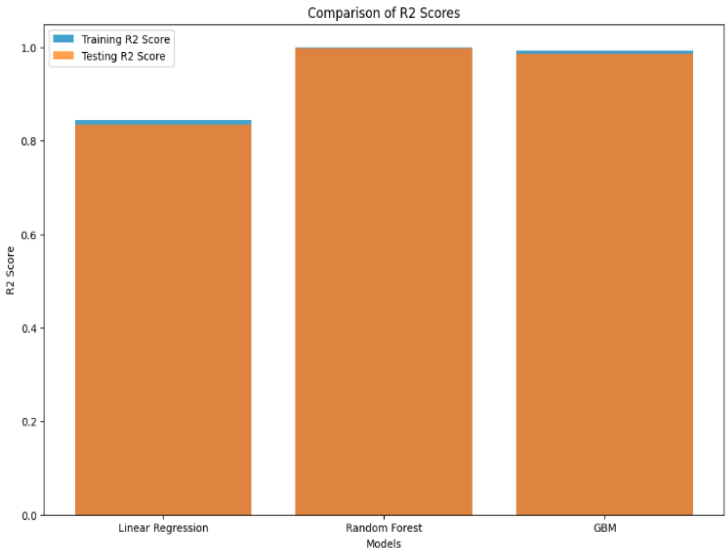


Figure 8. *Comparing Models and Their Accuracy: R2 Scores*

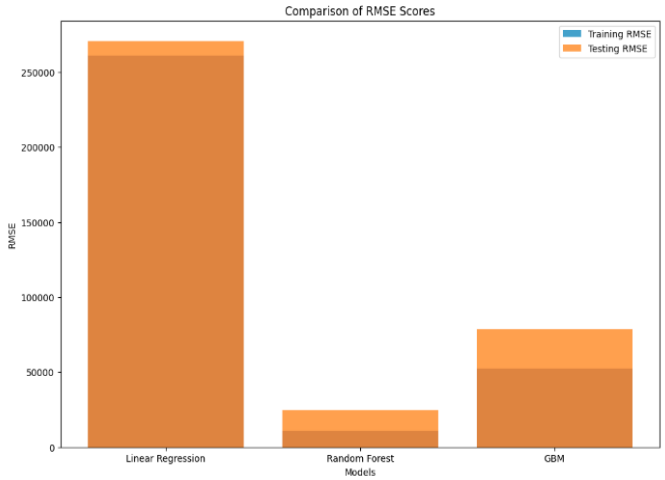


Figure 9. *Comparing Models and Their Accuracy: RMSE Scores*

The superiority is supported by broader evidence in the machine learning literature, which shows that combining multiple learners enhances predictive stability and reduces overfitting (Zhou, 2021). In educational data mining, where regional and sectoral variations introduce nonlinear complexity, these techniques have offered a more applied analysis framework than traditional regression models. Table 2 summarizes the overall comparison between models.

Table 2. *Comparative Summary of Model*

Model	R^2 (Test)	RMSE (Test)	Cross-Validation Error (10-fold)	Description
Linear Regression	0.78	210,000	215,000	Baseline, interpretable, underestimates high-density regions
Random Forest	0.93	95,000	100,000	Handles nonlinear trends, minimal bias
Gradient Boosting Machine	0.94	92,000	98,000	Slightly better than RF, iterative error correction, sensitive to tuning

Conclusion

Overall, there are considerable differences in enrollment trends across regions, with densely populated regions such as CALABARZON, NCR, and Central Luzon exhibiting higher enrollment trends than less populated regions such as CAR and BARMM. The results further supported the notion that most enrollment trends across the Philippines are driven by public educational institutions, underscoring the high dependence on public education within the basic education curriculum.

To address the study's goal of identifying the most effective approach for forecasting enrollment trends in the Philippines, the findings indicate that ensemble machine learning methods, specifically Random Forest and Gradient Boosting Machine (GBM), outperform traditional Linear Regression. These models demonstrated stronger explanatory power and lower prediction errors, reinforcing the reliability and accuracy of ensemble techniques for enrollment forecasting.

The results have important implications for educational planning and policy development. Accurate enrollment forecasting can assist policymakers, school administrators, and education planners in making data-driven decisions related to resource allocation, teacher deployment, classroom capacity planning, and infrastructure development. By incorporating predictive analytics into education management systems, government agencies such as the Department of Education can better anticipate regional enrollment demand and address disparities across regions.

Despite these contributions, this study has certain limitations. It must be noted that the dataset used in this study included only aggregate enrollment levels and excluded socioeconomic factors, migration levels, or institutional quality that could affect enrollment. It must also be noted that the imbalance between public and private school enrollment could affect the study's levels. In future studies, additional variables could be utilized, including more complex sampling methods, and more complex models, such as neural networks or deep learning techniques, could be applied.

This study has demonstrated that machine learning-based predictive models, including Random Forest and Gradient Boosting Machine, can serve as an effective analytical framework for forecasting enrollment levels in the Philippine education system. The high levels of predictability demonstrated by the study are valuable in assisting decision-makers in strategic decision-making.

Contributions of Authors

Author 1: conceptualization, data gathering, data analysis, writing – original and revised draft

Funding

None

Conflict of Interests

No conflict of interest.

Acknowledgment

The researchers would like to express their sincere gratitude to everyone who contributed to the successful completion of this data mining research paper. First, the researcher co-contributors, Ms. Nelma Mae Loja, MIS and Mr. Rafael Dela Peña for their help and contribution, and to extend their deepest appreciation to the Graduate School Department of the esteemed institution, the University of Immaculate Conception, and to our research advisers, Dr. Rhodessa Cascaro and Dr. Glen Gara, for their invaluable guidance, insightful feedback, and continuous support throughout this coursework project. Additionally, the researchers are grateful to the Kaggle Repository research team for the dataset, whose collaborative spirit and expertise have significantly enhanced the quality of the work. Lastly, to acknowledge the support from their families and friends, whose encouragement and understanding have been a source of strength during this research endeavor, and to God Almighty, who has been the source of all knowledge and wisdom. Thank you all for your contributions and support.

References

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. Education Sciences, 11(9), 552. https://doi.org/10.3390/educsci11090552

Bishop, C. (2006). Pattern recognition and machine learning. New York, NY: Springer. https://link.springer.com/book/9780387310732

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Dela Cruz, A., Reyes, M.L., Santos, J.P., & Gomez, R.T. (2020). Higher education institution enrollment forecasting using data mining techniques. International Journal of Advanced Trends in Computer Science and Engineering, 9(2), 2060–2064. https://doi.org/10.30534/ijatcse/2020/179922020

Delima, A.J., Sison, A., & Medina, R. (2019). Variable reduction-based prediction through a modified genetic algorithm. International Journal of Advanced Computer Science and Applications, 10(5), 356–363. https://doi.org/10.14569/IJACSA.2019.0100544

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

Haris, N.A., Abdullah, M., Hasim, N., & Rahman, F.A. (2016). A study on students' enrollment prediction using data mining. In D. Taniar, M. H. Böhlen, & J. W. Rahayu (Eds.), Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication (pp. 1–6). Danang, Vietnam: ACM. https://doi.org/10.1145/2857546.2857633

Hyndman, R., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). Melbourne, Australia: OTexts. https://otexts.com/fpp3
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-0716-1418-1
Parveen, K., Shah, N.H., & Mahmood, Z. (2020). Evaluation of enrollment trends in technological subjects at the secondary level in Punjab. Global Social Sciences Review, 5(1), 538–550. https://doi.org/10.31703/gssr.2020(V-I).55
Rabelo, A., Rodrigues, M., Nobre, C., Isotani, S., & Zárate, L. (2024). Educational data mining and learning analytics: A review of educational management in e-learning. Information Discovery and Delivery, 52(2), 149–163. https://doi.org/10.1108/IDD-10-2022-0099
Raiblaze. (2022). Philippines school enrollment data [Dataset]. Kaggle. https://tinyurl.com/4nt24sdp
Roman, A.G., & Villanueva, R.U. (2018). Enrolment trend analysis among transition periods of a university for management intervention. International Journal of Science and Research, **7**(10), 604–608.
Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355
Sarker, I. (2021). Machine learning: Algorithms, real-world applications, and research directions. SN Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x
Sharma, R., & Kumar, P. (2022). Forecasting student enrollment using ensemble learning techniques. Expert Systems with Applications, 195, 116595. https://doi.org/10.1016/j.eswa.2022.116595
Tan, E. (2017). Quality, inequality, and recent education reform. Philippine Review of Economics, 54(2), 110–137. https://tinyurl.com/yzab626w
Wuttaphan, N. (2017). Human capital theory: The theory of human resource development, implications, and future. Social Sciences, 18(2), 240–253
Zhou, Z.-H. (2021). Machine learning. Springer Singapore.