




Data-Driven Forecasting of Special Education Enrollment: An Explainable Machine Learning Approach

Raul Alberto Garcia Castro^{1,*}, Wildon Rojas Paucar², Elena Miriam Chavez Garces³,
Rubens Houson Pérez-Mamani⁴

^{1,3}Universidad Nacional Jorge Basadre Grohmann, Tacna and 23001, Peru

²Universidad Nacional de Moquegua, Moquegua, 18001, Peru

⁴Universidad Nacional Mayor de San Marcos, Lima, 15001, Peru

(Received: June 1, 2025; Revised: August 5, 2025; Accepted: November 16, 2025; Available online: December 9, 2025)

Abstract

The application of machine learning algorithms in the field of special education remains incipient despite advances achieved in other sectors. This field faces challenges related to inclusion, planning, and resource allocation, especially in contexts where administrative records are often underutilized for analytical purposes. This study proposes an explainable forecasting approach based on 23,905 historical data records to anticipate educational demand in the Special Basic Education (SBE) modality, aiming to develop and validate a Random Forest model applied to a multivariate database of official enrollment records from 2019 to 2024, projecting a slight global contraction from 28,000 to 26,800 enrollments by 2025. The findings reveal nonlinear growth patterns differentiated by region and educational level, mainly in Early SBE (ages 0 to 2), Preschool, and Primary, with a general trend of increasing demand in coastal and highland regions. The models achieved high levels of accuracy ($R^2 > 0.97$), with a Root Mean Squared Error (RMSE) below 190, a Mean Absolute Error (MAE) under 70, and a Mean Absolute Percentage Error (MAPE) below 10%. These results demonstrate the model's utility as a strategic decision-support tool by optimizing resource planning in an education system characterized by territorial heterogeneity. The novelty of this study lies in integrating geospatial analysis and predictive algorithmic interpretability within an explainable artificial intelligence framework, fostering more equitable, transparent, and evidence-based educational planning.

Keywords: Especial Education, Machine Learning; Time Series, Educational Prediction, Random Forest, Territorial Planning, Explainable Analysis

1. Introduction

Special education faces a complex and persistent challenge within traditional educational systems, both in developed and developing countries. Globally, students with disabilities continue to encounter structural barriers that limit their access, participation, and progress in education. These conditions are reflected in lower levels of academic achievement, higher rates of school delay and dropout, and reduced opportunities for social and occupational inclusion in adulthood [1], [2], [3]. One of the underlying causes of these gaps is the limited capacity of school systems to adapt to diversity: regular schools often lack adequate resources, infrastructure, and teacher training to appropriately serve students with special educational needs [4]. As a result, segregated practices and biased diagnostic decisions persist, negatively affecting students' performance and well-being [5].

Within this global context, the educational demand for special education has shown steady growth over the past decade. This increase is driven by progressively inclusive legal frameworks, improvements in the identification of specific needs, and the growing recognition of the educational rights of persons with disabilities. For example, in the United States, students served under the IDEA law represented 13.7% of total enrollment by 2018 [6], while in England, the demand increased by 11% between 2013 and 2018 [7]. In the case of Peru, these global challenges take on particular characteristics. Persons with disabilities receive, on average, fewer years of schooling than those without disabilities and show considerably higher illiteracy rates [8]. Although placement in specialized programs may help improve

*Corresponding author: Raul Alberto Garcia Castro (rgarciac@unjbg.edu.pe)

 DOI: <https://doi.org/10.47738/jads.v7i1.1046>

This is an open access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>).

© Authors retain all copyrights

academic performance and retention [9], the central challenge lies in providing personalized and sustainable support without generating new forms of exclusion.

Similarly, the institutional response in Peru, though increasing, remains limited in relation to the magnitude of educational demand. In 2018, only 1% of students in regular schools had a registered disability diagnosis. Of the 77,496 students identified with disabilities, more than 50,000 attended inclusive modalities and about 18,000 were enrolled in Special Basic Education [8]. By 2022, specialized services reached 25,510 students [10], showing relative progress, though still insufficient compared to the country's actual needs.

Understanding the evolution of educational demand in SBE is essential for anticipating needs, allocating resources, and formulating evidence-based inclusive policies [1]. However, across much of Latin America, including the Andean region, there remains an underutilization of educational data for predictive planning, particularly in specialized modalities such as SBE [11]. In this regard, machine learning models applied to multivariate time series provide effective tools for projecting demand and understanding associated factors, under an explainable framework that promotes context-sensitive and equity-oriented decision-making [12], [13].

From an epistemological perspective, Machine Learning (ML) is grounded in computational theories of artificial intelligence, statistical learning, and adaptive systems, which enable the modeling of complex phenomena through algorithms that learn from data [14]. Its current maturity reflects a transformative potential for evidence-based decision-making by identifying hidden patterns in large volumes of information. Machine learning and deep learning technologies are leading innovation processes in sectors such as health, finance, and education, demonstrating effective applicability across multiple contexts [15].

In the educational field, and particularly in special education, these approaches align with the principles of Universal Design for Learning (UDL) and personalized education, contributing to a more flexible and predictive understanding of functional diversity [7], [16]. However, their implementation in middle-income countries such as Peru remains incipient. This study seeks to provide empirical evidence on the prediction of demand in SBE through machine learning algorithms applied to institutional data, thus contributing to the strengthening of inclusive educational planning in the national context.

2. Literature Review

The application of ML in the field of special education has experienced notable growth, gradually displacing traditional descriptive approaches in favor of predictive models with greater analytical power and personalization [17], [18]. In early contributions, Fiore et al. and Nelson et al. conducted institutional assessments of special education services in public and charter schools, without yet incorporating advanced data analysis tools [4], [9]. Years later, the Government Accountability Office (GAO) highlighted persistent structural inequalities, particularly in early needs detection and resource allocation, laying the groundwork for a still-pending methodological transformation [19].

Substantial progress is evident in recent studies that have employed ML for predictive and classification purposes. Broda et al., for example, analyzed data from over 20,000 adults with intellectual disabilities using decision trees and logistic regressions, identifying key predictors such as age, disability level, and service history to model employment and social participation trajectories [20]. For children, Tan et al. (2024) used nonlinear classifiers to predict subjective well-being in students with special educational needs, generating differentiated profiles such as the “socializer” and the “analyst” to personalize interventions based on emotional, social, and academic factors [21].

The field of Autism Spectrum Disorder (ASD) has been one of the most dynamic in integrating artificial intelligence. Wankhede et al. highlighted how ML algorithms detect complex patterns in neuroimaging, genetic data, and observed behaviors, facilitating earlier and more accurate diagnoses [22]. Technologies such as augmentative communication systems, virtual reality, and social robotics have expanded the possibilities for communication and socioemotional development. In a rapid review of 64 studies, Ganggayah et al. confirmed the effectiveness of ML in ASD diagnosis and intervention, while also cautioning about ethical risks related to privacy, equitable access, and algorithmic bias [16].

From a broader perspective, Ghunaim et al. showed how ML has revolutionized diagnosis and educational planning in pediatric genetic syndromes like Cri du Chat or 22q11.2 [23]. Tools like DeepVariant or CNVnator improve genetic precision, while platforms such as IBM Watson Education and Amazon Polly support curricular adaptation. However, despite their promise, these developments have been largely concentrated in medical-clinical settings, limiting their generalization to mainstream educational contexts, particularly at the basic level.

In specifically educational domains, Alkan systematized the main AI technologies applied to students with special needs, including adaptive tutors, expert systems, learning analytics, and educational data mining [14]. These tools can anticipate academic risks, offer automated feedback, and personalize learning experiences. The author also stresses the urgency of developing robust ethical frameworks that safeguard student privacy and promote equitable access to these technologies, especially in vulnerable contexts. In summary, the literature shows that ML offers valuable opportunities to improve special education through individualized analysis, trajectory prediction, and personalized intervention design. However, a significant gap remains in applying ML to the analysis of educational demand in Special Basic Education (EBE), particularly in real school systems and in middle-income countries like Peru. This research aims to address that gap by integrating ML tools with institutional databases to enhance inclusive, evidence-based educational planning.

Various studies have documented the structural inequalities faced by students with disabilities in terms of access, retention, and educational quality [4], [19], [24]. However, most of these studies rely on descriptive approaches or traditional statistics, which limits their ability to identify complex patterns or differentiated trajectories within this population. In contrast, machine learning has shown high potential for generating predictive models and personalized interventions, especially in areas such as mental health, school well-being, and developmental disorder [21].

However, its application in SBE remains limited, particularly in real school settings and in middle income countries such as Peru. The progressive digitalization of the education system offers an unprecedented opportunity to enhance inclusive planning. Nevertheless, without the use of advanced analytical models, there remains a risk of making poorly contextualized decisions based on aggregate data or on intuitions that fail to capture the true diversity of the student population.

In this context, the present study aims to predict the evolution of educational demand among students with special needs in Peru between 2019 and 2025, through time series modeling using machine learning algorithms. From an explainable perspective, the research seeks to identify temporal patterns, nonlinear relationships, and critical variables that influence demand by region and educational level, generating empirical evidence that contributes to more inclusive, efficient, and data driven educational planning.

However, the review shows that most studies originate from Anglo Saxon contexts or high-income economies, where institutional capacities for educational data management and intelligent technology implementation are considerably greater. In contrast, in Latin America and particularly in Peru, there are still no explainable predictive models applied to special education that integrate sociodemographic, institutional, and pedagogical variables within a territorial framework. This methodological and empirical gap limits the generation of locally grounded evidence for inclusive planning. The present study contributes to filling this gap by adapting and validating an explainable machine learning model aimed at predicting educational demand in the Special Basic Education modality, thereby providing contextualized evidence to support data informed decision making in the Latin American context.

3. Methodology

3.1. Study Design

This study adopts a quantitative retrospective longitudinal design to model the evolution of enrollment in SBE in Peru. The Random Forest algorithm was selected for its ability to capture nonlinear relationships between predictor variables and enrollment, as well as its effectiveness in handling hierarchical data (by region and educational level) and maintaining robustness against outliers—particularly relevant in regions with low student density. This technique has proven effective in educational contexts for analyzing longitudinal data with hierarchical structures [25], [12].

3.2. Data Collection and Processing

Official data were collected from the Peruvian Ministry of Education via the ESCALE portal (<https://escale.minedu.gob.pe>), managed by the Educational Statistics Unit. A total of 23,905 historical enrollment records in EBE for the period 2019–2024 were consolidated. The dataset included variables such as year, region, type of management (public or private), and educational level (Preschool, Primary, and Total Basic), covering all 25 regions of the country. The data were integrated, restructured, and cleaned; no imputation techniques were necessary, as the dataset was sufficiently complete for analysis [26].

Although the database provides full national coverage, potential biases inherent to administrative records are acknowledged, such as the underreporting of students with disabilities and regional discrepancies in the consistency of reporting. These limitations are common in official sources on special education due to differences in diagnostic criteria and institutional recording capacity [10]. Nevertheless, the model incorporated contextual variables such as population, number of institutions, and total enrollment, which partially mitigated these biases and supported a more cautious and contextually grounded interpretation of the projections.

To ensure the interpretability of the Random Forest model, several post hoc explainability techniques were applied. First, global variable importances were computed using the Mean Decrease in Impurity method, allowing the ranking of the most influential predictors in the projections [12]. Additionally, SHAP (SHapley Additive Explanations) values were estimated to quantitatively identify the individual contribution of each variable to regional and temporal predictions [13]. Finally, Partial Dependence Plots (PDP) and Accumulated Local Effects (ALE) plots were generated to visualize the marginal effects of key factors on projected demand [27]. These techniques made it possible to understand the relative importance, direction, and magnitude of predictor influence, facilitating the translation of results into empirical evidence useful for educational management and territorial planning in Special Basic Education in Peru.

3.3. Data Splitting

The database was divided into a training set (80%) and a test set (20%), following methodological recommendations for supervised models applied to time series [28]. This partition strictly preserved the chronological order of the records (2019–2024) to avoid temporal bias. In addition, a walk forward validation scheme was implemented, in which the model is trained cumulatively with previous years and evaluated on the subsequent period, simulating a realistic forecasting scenario and ensuring the validity of the error metrics (RMSE, MAE, MAPE, and R^2) [29].

3.4. Predictive Modeling

The Random Forest algorithm was applied in regression mode due to its ability to handle nonlinear relationships and correlated variables, its tolerance to noise, and its suitability for multivariate and heterogeneous data. These properties make it possible to obtain interpretable predictions in multiregional educational contexts [12].

Unlike classical time series models such as ARIMA, which assume stationarity and univariate processes, the administrative records used contain multiple correlated predictors and nonlinear relationships between variables [12], [13]. In this context, Random Forest offers substantial advantages: it handles complex interactions without requiring strict assumptions, is robust to outliers, and integrates categorical and continuous variables without prior normalization [25].

Although neural network models also capture nonlinearities, their training requires large data volumes and intensive hyperparameter tuning, which reduces interpretability and hinders replicability in educational contexts with limited administrative databases [28]. In contrast, Random Forest allows the incorporation of post hoc explainability techniques (feature importance, SHAP values, PDP and ALE plots), promoting transparency and traceability in the projections [13]. These characteristics make it the most suitable option for analyzing educational demand, balancing accuracy, robustness, and interpretability, which are key aspects for territorial planning and public policy design.

To ensure the interpretability of the model, several post hoc explainability techniques were implemented: global feature importances were calculated using the Mean Decrease in Impurity method, SHAP values were estimated to identify the individual contribution of each predictor, and Partial Dependence Plots (PDP) and Accumulated Local Effects (ALE) plots were generated to visualize the marginal effects of the main variables on projected demand.

Furthermore, its implementation in educational studies has shown strong results in predicting enrollment and academic performance [25]. Specific models were developed for each educational level (Preschool, Primary, and Total Basic), as well as one overall national model. Hyperparameters were empirically tuned to optimize model performance. Random Forest was chosen over classical time series models such as ARIMA and recurrent neural network approaches because it requires fewer assumptions about data structure and has demonstrated the ability to deliver accurate results with a low risk of overfitting in scenarios with multiple explanatory variables [28].

3.5. Performance Evaluation

Model performance was evaluated using four commonly adopted metrics in the machine learning literature: the coefficient of determination (R^2), which indicates the proportion of variability explained by the model; the Root Mean Squared Error (RMSE), which penalizes large errors more heavily; the Mean Absolute Error (MAE), representing the average deviation between actual and predicted values; and the Mean Absolute Percentage Error (MAPE), which expresses error in relative percentage terms [30]. These metrics allowed for performance comparisons between models by educational level and the aggregated model, ensuring the consistency and robustness of the comparative analysis.

To improve the performance of the Random Forest model, an empirical tuning of the main hyperparameters was carried out through cross validation and incremental experimentation. The parameters evaluated included the number of trees (`n_estimators`), the maximum tree depth (`max_depth`), the minimum number of samples required to split a node (`min_samples_split`) and per leaf (`min_samples_leaf`), as well as the number of predictors considered at each split (`max_features`). After testing, the optimal model was configured with 300 trees, a maximum depth of 15, `min_samples_split` = 4, `min_samples_leaf` = 2, and `max_features` = 'sqrt', balancing bias and variance while improving predictive performance.

3.6. Computational Environment and Tools

The analysis was conducted using Python 3.10. Data manipulation was performed using the pandas and numpy libraries; model training and evaluation were conducted with scikit-learn; and result visualization was carried out using matplotlib and seaborn. A fixed random seed was set to ensure the experiment's reproducibility [27].

3.7. Research Flow and Model Architecture

Figure 1 presents the research flow and architecture of the explainable machine learning framework proposed in this study, summarizing each methodological stage from data acquisition to the visualization and interpretation of the 2025 forecast results.

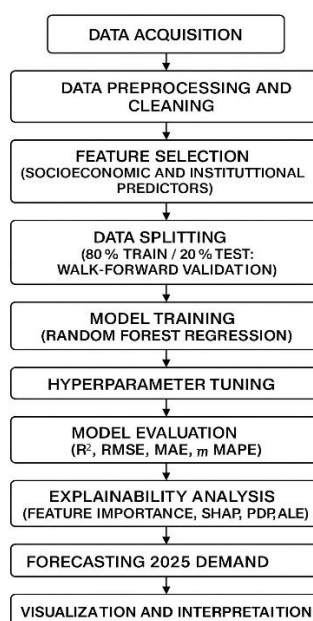


Figure 1. Research Flow and Architecture of the Explainable Machine Learning Framework

The research flow of the proposed model is presented in Figure 1. It summarizes the sequence of procedures adopted in this study, from data acquisition to the generation of explainable forecasts for 2025. The process begins with the collection of 23,905 administrative enrollment records from the ESCALE platform (2019–2024), followed by data preprocessing, cleaning, and the selection of relevant features such as region, educational level, and management type. The dataset was split into training (80%) and testing (20%) subsets under a walk-forward validation scheme to preserve temporal order. The Random Forest regression model was trained and optimized through hyperparameter tuning, and its performance was evaluated using R^2 , RMSE, MAE, and MAPE metrics. Explainability techniques, including Feature Importance, SHAP values, Partial Dependence Plots (PDP), and Accumulated Local Effects (ALE), were applied to identify the contribution and direction of influence of each predictor variable. Finally, the model forecasted the national enrollment demand for 2025, integrating the results into visual interpretations that support data-driven educational planning.

Equations Used for Model Evaluation

The following equations were used to evaluate the performance of the Random Forest regression model. The predicted value of enrollment (\hat{y}) was estimated as the average output of all trees in the ensemble. Model accuracy and error were quantified using four common metrics: the coefficient of determination (R^2), RMSE, MAE, and MAPE, as shown below.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

$$MAPE = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

These indicators were applied to both the training and testing datasets to ensure consistency and robustness across all predictive models developed for each educational level and the overall national model.

Pseudocode of the Random Forest Forecasting Framework

Input: Dataset D (2019–2024), features X, target y (enrollment)

Output: Forecasted enrollment for 2025

1. Load dataset D from ESCALE
2. Clean and preprocess data
3. Split data: 80% training, 20% testing
4. For each iteration (walk-forward validation):
 - a. Train Random Forest model
 - b. Predict \hat{y} on test set
 - c. Compute R^2 , RMSE, MAE, MAPE
5. Tune hyperparameters ($n_estimators$, max_depth , etc.)
6. Apply explainability methods (Feature Importance, SHAP, PDP, ALE)
7. Forecast enrollment for 2025
8. Visualize and interpret results

Experimental Setup

All experiments were conducted using Python 3.10 on a workstation equipped with an Intel Core i7-12700 processor, 32 GB of RAM, and Windows 11 OS. The main libraries employed were *pandas* and *numpy* for data manipulation, *scikit-learn* for model training and evaluation, and *matplotlib* and *seaborn* for visualization. A fixed random seed was used to ensure reproducibility across all experiments.

4. Results and Discussion

The following section presents the main findings of the predictive modeling of special education demand in Peru for the 2019–2025 period, using machine learning algorithms applied to time series data. The results include territorial comparisons, educational level analyses, and model performance evaluations.

For trends by educational level (Preschool, Primary, and Total Basic), only the most representative graphs are presented, while detailed tables are available in a supplementary dataset. The emphasis is placed on visualizing general patterns and the accuracy achieved by the predictive models. In this context, [figure 2](#) illustrates the projected evolution of Special Basic Education (EBE) demand by department between 2019 and 2024, highlighting regional differences in enrollment growth across the country.

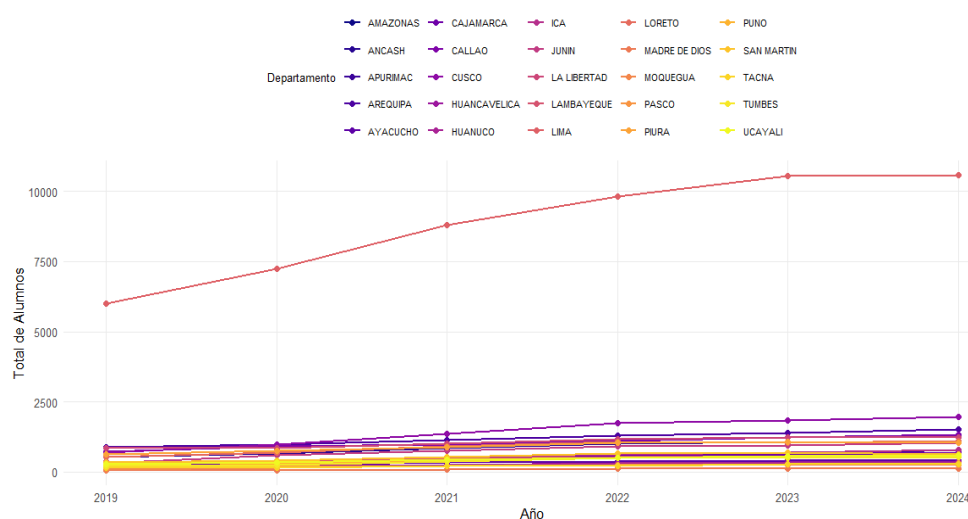


Figure 2. Evolution of EBE Demand by Department (2019–2024)

[Figure 2](#) shows the projected evolution of the demand for students with special educational needs by department between 2019 and 2024. A nationwide increase is observed, although significant regional differences appear both in absolute volumes and growth patterns. Lima, located on the central coast, stands out with a demand exceeding 10,000 projected students by 2024, far above the rest of the country. It is followed by departments such as Junín, Cusco, Arequipa, and La Libertad, in the central and south-central highlands and coast, which present moderate but consistent growth. In contrast, regions like Madre de Dios, Ucayali, Tumbes, Moquegua, and Pasco maintain low and stable levels.

Beyond the absolute figures, a clear geographical pattern emerges: departments in the central and southern coast and highlands concentrate the highest demand, likely due to greater population density, better educational infrastructure, and stronger institutional presence. Conversely, most rainforest regions (e.g., Loreto, Ucayali, and Madre de Dios) show lower volumes, possibly associated with access barriers, territorial dispersion, and limited availability of specialized services. Although relative indicators (such as rates by child population) are required to adequately assess coverage, the observed distribution suggests that geographic location strongly shapes the concentration and projection of special education demand, underscoring the importance of differentiated territorial planning.

In this context, [figure 3](#) shows the projected evolution of Special Basic Education (EBE) demand by educational level between 2019 and 2024, distinguishing enrollment trends for ages 0–2, Preschool, and Primary.

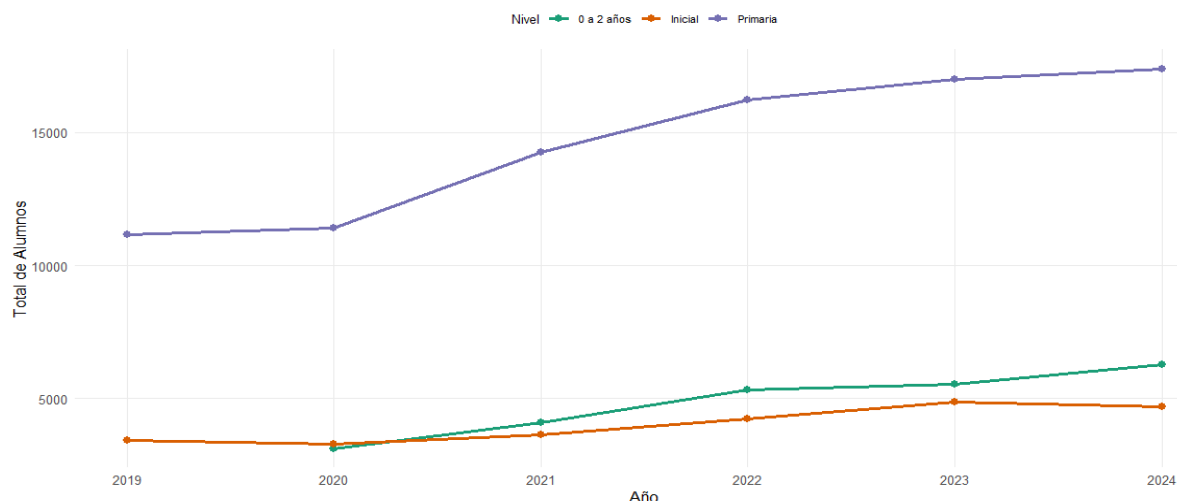


Figure 3. Evolution of EBE Demand by Educational Level (2019–2024)

Figure 3 shows the projected evolution of special education demand, broken down by educational level: ages 0 to 2 (PRITE programs), Preschool, and Primary. The Primary level consistently maintains the highest volume throughout the entire period, with steady growth surpassing 17,000 students by 2024. This trend is expected, as Primary comprises six school grades (first through sixth) and covers a larger share of the school-age population.

In second place, the 0 to 2 years level shows a substantial increase between 2020 and 2024, doubling its initial volume. This growth may be related to the expansion of early intervention and inclusive education programs in early childhood. Finally, the Preschool level shows a more stable trend with a smaller relative increase, which may reflect more limited coverage or lower identification of special educational needs at this stage.

Overall, the figure highlights that demand growth is not uniform across levels, and that early childhood services (ages 0 to 2) require special attention in future planning, given their accelerated growth and the critical importance of early intervention. Figure 4 illustrates the projected evolution of Special Basic Education (EBE) demand by type of management between 2019 and 2024, highlighting the clear predominance of public institutions over private ones throughout the period.

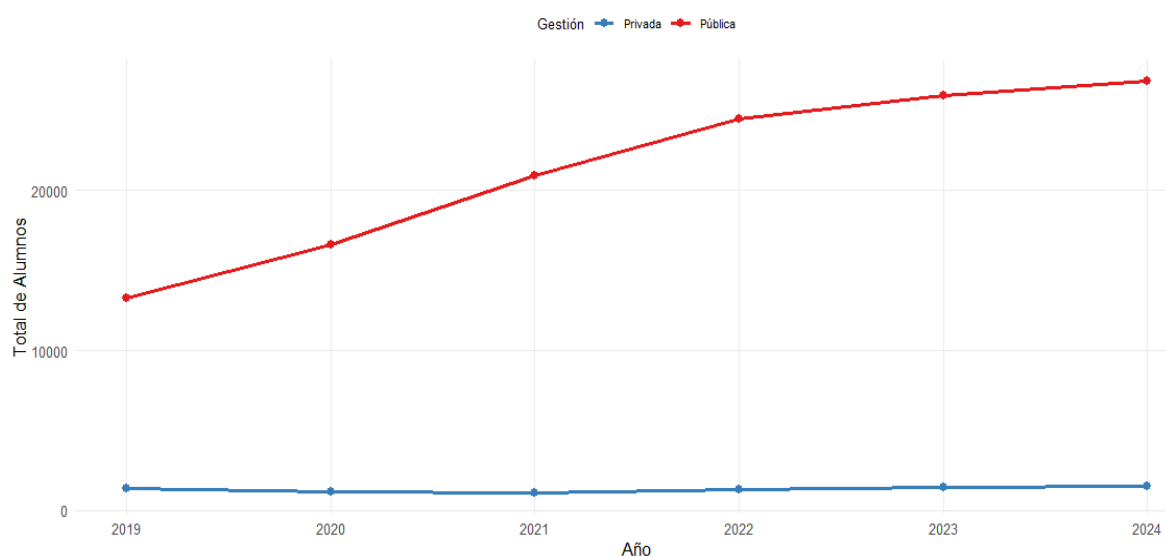


Figure 4. Evolution of EBE Educational Demand by Type of Management.

The figure 4 shows the evolution of special education demand by type of management: public and private. A clear predominance of the public sector is observed, with demand steadily increasing from approximately 13,000 to over

23,000 students between 2019 and 2024. In contrast, demand in privately managed institutions remains static and at very low levels (below 2,000 students throughout the entire period). This disparity reflects the central role of the state in providing special education services in Peru, particularly in contexts of social vulnerability. It may also be influenced by the limited availability of specialized private offerings and the higher costs associated with these services compared to regular education.

The graph also suggests that the growing demand is being absorbed primarily by the public system, which could lead to additional pressure on infrastructure, specialized personnel, and state funding if not accompanied by a proportional increase in resources. Figure 5 presents the projected trend of enrollment demand in Special Basic Education (EBE) by department for the 2019–2025 period, showing a sustained increase until 2023 followed by a slight decline in 2025.

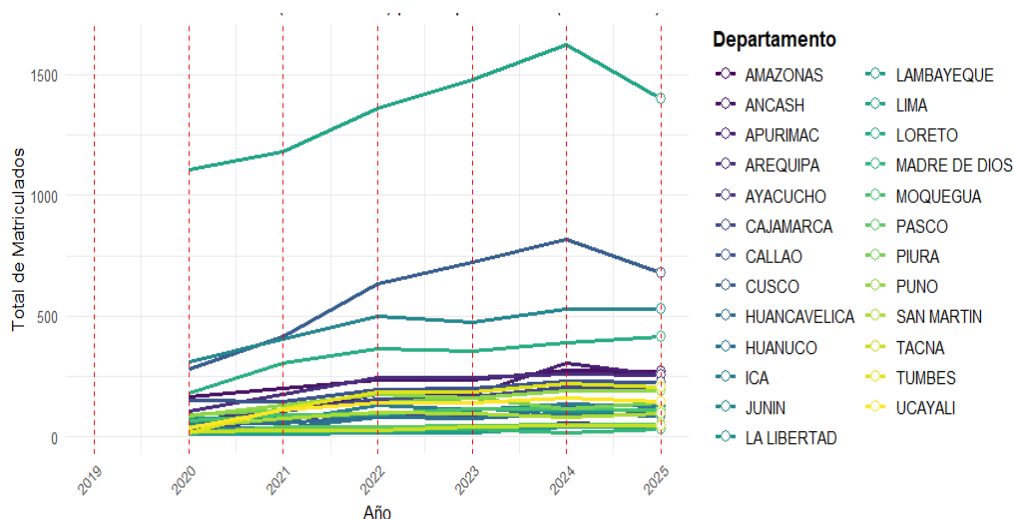


Figure 5. Projected Trend of Educational Demand in Special Basic Education (EBE) by Department (2019–2025).

Note. Chart created by the authors based on the trend table published on Figshare (2025). Available at: [10.6084/m9.figshare.29323010](https://figshare.com/10.6084/m9.figshare.29323010)

Figure 5 displays the projected enrollment demand at the Basic level (including services for ages 0 to 2, Preschool, and Primary) for students with special educational needs, disaggregated by department between 2019 and 2025. A sustained increase is observed across most regions until 2023, followed by a marked decline in 2025, particularly in Lima and Arequipa, which concentrate the highest educational demand nationwide. Lima leads with more than 1,600 projected students in 2023, followed by Arequipa, Junín, Cusco, and La Libertad, all showing upward trajectories, while departments with lower population density such as Madre de Dios, Moquegua, Tumbes, and Ucayali remain at low and relatively stable levels throughout the period.

The predictive model achieved a coefficient of determination of $R^2 = 0.917$, indicating that 91.74% of the variation in demand is explained by the included variables, while the remaining 8.26% may be associated with unaccounted factors such as education policy shifts, infrastructure limitations, or external disruptions. Additional performance metrics confirmed its robustness (MSE = 8,952.78; RMSE = 94.62; MAE = 49.36), supporting the model's accuracy and reliability. With a 95% Confidence Interval (CI) ranging from 1,415 to 1,785, the margin of error ($\pm 11.8\%$) indicates moderate but controlled variability, demonstrating the model's strong capacity to estimate regional fluctuations. This range confirms that the projections exhibit adequate statistical robustness and that the estimated values are consistent with trends observed in previous years.

The projected decline in 2025 could be associated with a stabilization of demand or the lagged effect of recent disruptive events. However, this possible stabilization should be interpreted with caution, since the database used does not include external indicators (such as demographic or economic variables, or changes in education policies) that would allow this trend to be empirically contrasted. Therefore, the explanation is based solely on the internal dynamics of the

predictive model and on historical variations in enrollment observed in the analyzed series. Overall, these results provide valuable evidence for guiding inclusive educational planning, particularly in regions with higher projected growth.

The most notable increases in Lima, Arequipa, and Junín may be associated with the sustained expansion of public special education services and a greater capacity for student registration, while the projected declines in Madre de Dios and Moquegua could be related to budgetary constraints, limited availability of specialized personnel, and difficulties in maintaining specialized programs in low population density areas. These regional differences indicate that the observed trends depend not only on urban density but also on structural and management factors that condition the effective supply of special education in each territory. Figure 6 shows the projected trend of enrollment demand at the Preschool level in Special Basic Education (EBE) by department for the 2019–2025 period, revealing a steady increase up to 2023 followed by a slight decline in subsequent years.

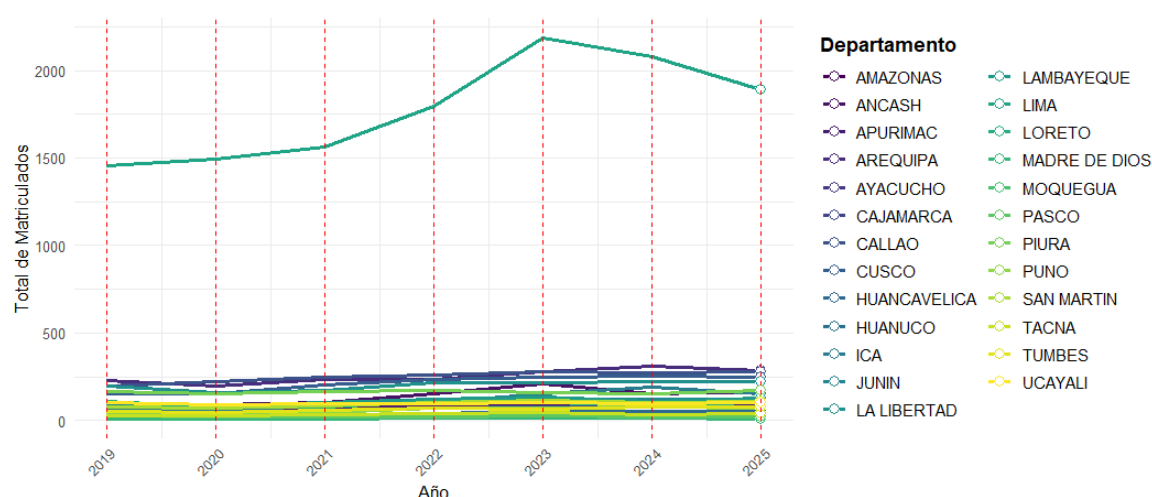


Figure 6. Trend of Demand at the Preschool Level by Department (2019–2025).

Note. Chart created by the authors based on the trend table published on Figshare (2025). Available at: [10.6084/m9.figshare.29323010](https://figshare.com/10.6084/m9.figshare.29323010)

Figure 6 shows the projected enrollment at the Preschool level (ages 3 to 5) for students with special educational needs between 2019 and 2025, disaggregated by department. At the national level, an upward trend is observed from 2019, peaking in 2023 and followed by a slight decline in the two subsequent years. This pattern suggests an initial expansion in coverage, possibly followed by stabilization in the provision of support to this population. Lima clearly leads in demand, surpassing 2,000 projected students in 2023, followed by Arequipa, Junín, Cusco, and La Libertad, which also show rising trajectories. In contrast, departments with lower population density, such as Madre de Dios, Huancavelica, Tumbes, and Ucayali, maintain low and stable levels throughout the analyzed period.

The predictive model underlying these projections achieved a coefficient of determination of $R^2 = 0.978$, indicating that 97.83% of the variability in demand is explained by the variables considered, thus reflecting an excellent fit. Additional performance metrics confirm its robustness (MSE = 3,366.07; RMSE = 58.02; MAE = 25.27), demonstrating low error levels and strong generalizability. With a 95% confidence interval (CI) ranging from 1,886 to 2,114, the variation margin of only $\pm 5.6\%$ reflects high model precision and excellent prediction stability. These results demonstrate that the generated estimates are robust and that the observed growth does not correspond to random fluctuations, but rather to consistent temporal patterns captured by the Random Forest model. These findings confirm that the projections exhibit low error levels and a high capacity for generalization.

Taken together, these results provide a robust empirical basis for guiding specialized educational planning at the Preschool level, allowing resources and strategies to be focused on regions with the highest projected growth in demand. In regions such as Lima, Arequipa, and Cusco, the projected growth may be linked to the strengthening of early intervention programs and the increase in early childhood diagnoses driven by inclusive education policies. In contrast, the lower growth in Huancavelica, Ucayali, or Tumbes may reflect coverage gaps and limited availability of specialized centers for early age groups. Figure 7 depicts the projected trend of enrollment demand at the Primary level in Special Basic Education (EBE) by department from 2019 to 2025, showing a steady rise until 2023 followed by a slight stabilization and decline by 2025.

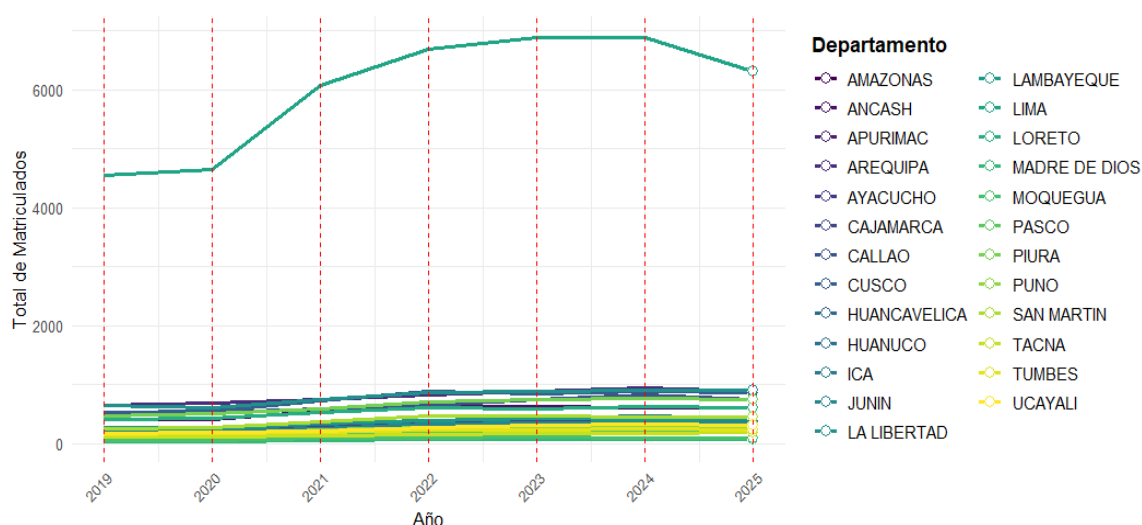


Figure 7. Trend of Demand at the Primary Level by Department (2019–2025).

Note. Chart created by the authors based on the trend table published on Figshare (2025). Available at: [10.6084/m9.figshare.29323010](https://figshare.com/figures-and-tables/29323010)

Figure 7 presents the projected enrollment at the Primary level for students with special educational needs between 2019 and 2025, disaggregated by department. A sustained increase in demand is observed up to 2023, followed by slight stabilization and a decline by 2025. This pattern suggests a progressive expansion of service coverage during the early years of the period, possibly associated with improvements in detection systems, access, and inclusive education, followed by a recent adjustment in projected demand. Lima stands out as the main center of service provision, with more than 6,000 projected students in 2023, followed by La Libertad, Junín, Cusco, and Arequipa, which display upward trajectories at intermediate levels of concentration. In contrast, regions such as Madre de Dios, Moquegua, Tumbes, and Huancavelica maintain low figures with no significant variation over time, likely reflecting structural limitations or limited coverage of specialized services.

The predictive model achieved a coefficient of determination of $R^2 = 0.978$, indicating that 97.83% of the variability in demand is explained by the variables considered, which reinforces the strength of the estimates. Additional performance metrics ($MSE = 35,882.62$; $RMSE = 189.43$; $MAE = 67.21$). With a 95% confidence interval (CI) ranging from 5,629 to 6,371, this narrow margin ($\pm 6.2\%$) confirms the model's precision even in scenarios with larger data volumes. The consistency of the estimated values supports the reliability of the projections and reinforces the validity of the upward trend described for this educational level.

These results are consistent with the projections observed in the Preschool and Basic levels and strengthen the usefulness of the model as a decision support tool for inclusive education policies. Its application allows for the

identification of priority areas and the adjustment of resource allocation according to the expected evolution of demand. The sustained increase in Lima and La Libertad can be explained by educational continuity within consolidated regional systems with greater public funding, while stabilization in regions such as Moquegua or Huancavelica may be associated with the limited expansion of enrollment and the reduction in the number of specialized institutions in rural or sparsely populated areas. Figure 8 illustrates the overall projected trend of enrollment demand in Special Basic Education (EBE) by department from 2019 to 2025, showing a continuous upward trajectory until 2024 followed by a slight contraction in 2025.

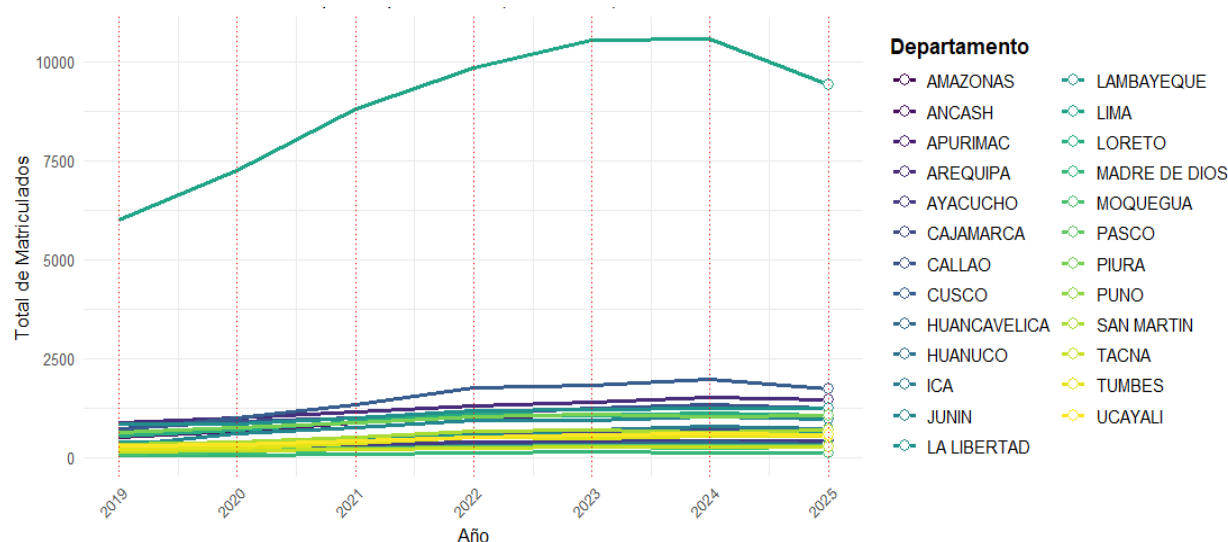


Figure 8. Overall Demand Trend by Department (2019–2025).

Note. Chart created by the authors based on the trend table published on Figshare (2025). Available at: [10.6084/m9.figshare.29323010](https://figshare.com/10.6084/m9.figshare.29323010)

The figure 8 presents the aggregated projection of enrollment for students with special educational needs in early childhood care (ages 0 to 2), Preschool, and Primary levels between 2019 and 2025, disaggregated by department. At the national level, a consistent upward trend has been observed since 2019, reaching a peak of nearly 28,000 students in 2024, followed by a slight contraction in 2025 to 26,800 enrollments. This pattern suggests a phase of sustained expansion, possibly associated with improvements in identification and coverage, followed by a stage of adjustment or stabilization.

Lima accounts for the largest share of demand, surpassing 10,000 students in 2024, which represents approximately 37% of the national total. It is followed by departments such as La Libertad, Arequipa, Cusco, Callao, and Junín, which also show upward trajectories. In contrast, regions such as Madre de Dios, Tumbes, and Moquegua maintain low and stable projected levels, which may reflect structural limitations or persistent gaps in specialized service provision.

The predictive model applied to this combined projection achieved a coefficient of determination of $R^2 = 0.974$, indicating an outstanding fit. A total of 97.39% of the observed variability was explained by the variables considered, including year, region, and historical enrollment data. The error metrics obtained (MSE: 101,793.9; RMSE: 319.05; MAE: 132.21). With a 95% CI ranging from 27,375 to 28,625, the margin of error was only $\pm 2.3\%$, reflecting very high stability and confirming the statistical soundness of the national-level predictions. These results consolidate the performance and reliability of the Random Forest model as a valid tool for generating realistic projections grounded in empirical evidence, although with slightly greater dispersion compared to the individual models, which is expected given the heterogeneous nature of the integrated educational levels.

Overall, the projected regional variations reflect both the historical concentration of resources in urban regions and the persistent gaps in infrastructure and specialized personnel in Amazonian and Andean areas. The increases observed in

Lima, La Libertad, and Arequipa suggest institutional consolidation of services, while the declines in Madre de Dios and Tumbes reveal operational fragility and limitations in the continuity of specialized educational services.

These findings reinforce the usefulness of the model as a reliable tool for strategic planning in special education, allowing for the anticipation of growth scenarios, the guidance of resource allocation, and the definition of territorial priorities within the framework of sustained inclusive education policies. Likewise, the model enabled the translation of projections into empirical evidence useful for educational management and territorial planning in Special Basic Education in Peru, contributing to the strengthening of equity, the optimization of resource distribution, and the promotion of inclusive care at the regional level.

In terms of educational policy, the results make it possible to move from reactive to prospective, evidence-based planning. Based on the generated projections, education authorities could prioritize three concrete lines of action: (1) redistribute specialized human resources toward regions with the highest projected demand growth; (2) plan phased investments in infrastructure and equipment for Special Basic Education Centers (CEBE) in areas with coverage deficits; and (3) strengthen registration and early diagnosis systems to reduce the underreporting that persists in rural and Amazonian regions. These strategies not only optimize the use of existing resources but also help to guide future budget allocations and public investment projects in a more equitable and targeted manner.

5. Discussion

The results obtained in this study demonstrate that the Random Forest model, applied in its regression mode, allows for highly accurate prediction of educational demand for students with special needs in Peru. In all models developed (by educational level and in the combined model), R^2 values above 0.97 were achieved, indicating an excellent fit between historical data and the projections generated. This suggests that the evolution of demand responds to temporal and territorial patterns that can be effectively captured using machine learning techniques.

In practical terms, the performance metrics obtained across the different models (R^2 , RMSE, MAE, and MAPE) provide a sufficient level of accuracy for application in educational planning. An R^2 value above 0.97 indicates that most of the variability observed in enrollment can be explained by the model, while average error values (MAE and MAPE) below 10% reflect minimal deviations between projections and actual records. For public policy decision-makers, this means that the estimates can be used with high reliability to anticipate regional demand for services, project teacher and budgetary requirements, and guide infrastructure expansion with a statistically controlled margin of uncertainty.

Moreover, the projections reveal an unequal distribution of demand, with clear concentration in urban regions such as Lima, La Libertad, Arequipa, and Junín, and lower levels in Amazonian and highland departments. This asymmetry reflects not only population density but also structural differences in access to diagnostic services and the installed capacity for specialized attention.

These findings are consistent with international studies that have demonstrated the potential of machine learning to improve the understanding and anticipation of complex educational phenomena. Broda showed that decision trees can predict employment trajectories for individuals with disabilities more accurately than traditional methods [20]. Similarly, Ganggayah et al. highlight that explainable models based on artificial intelligence, such as Random Forest, can identify relevant factors for decision making in inclusive educational settings [16].

Additionally, the unequal geographic distribution of projected demand aligns with the findings of Senese and Winters, who warned that school choice policies, when not accompanied by compensatory mechanisms, can reproduce existing inequalities in special education [7]. In this regard, the concentration observed in urban areas reflects not only broader institutional coverage but also a possible bias in the timely detection and official registration of students with disabilities, as also argued by Stock and Schultz [6].

While these analyses are supported by the international studies cited, most of them originate from high-income contexts characterized by higher levels of digitalization, institutional stability, and the availability of high-quality longitudinal data. It should be noted that the Peruvian education system faces significant gaps in infrastructure, diagnostic coverage, and specialized human resources, which limit the direct comparability of the findings. These contextual differences

may influence both the magnitude and dynamics of the projected demand, as the country's structural conditions still constrain the timely identification and comprehensive attention of the population with disabilities.

Furthermore, it is acknowledged that some of the cited studies correspond to very recent publications (e.g., [22]). Their inclusion responds to the purpose of updating the theoretical and methodological foundations of the field, as they provide emerging evidence on the use of machine learning in special education. However, they are interpreted in a referential and comparative manner, without assuming their full transferability to the Peruvian context, which presents structural and managerial particularities still distinct from the settings in which those studies were conducted.

From an education economics perspective, the integration of explainable predictive models makes it possible to optimize the allocation of human, financial, and logistical resources. This study provides a replicable methodology that can strengthen the institutional capacity of the state to plan based on evidence, anticipate pressure on the education system, and target policies toward more underserved territories.

The explainable nature of the model also provides a strategic advantage. It not only offers precision in prediction but also transparency in identifying key variables such as region, educational level, or management type that influence demand. This feature addresses the need highlighted by Molnar [13] and Khosravi et al. [16], who argue that the use of artificial intelligence in education must be grounded in principles of equity, inclusion, and auditability.

Despite the positive results, this study has some limitations. The database used includes only official enrollment records, which excludes students who do not have access to the education system or who have not been formally diagnosed. Furthermore, contextual variables such as socioeconomic indicators, health coverage, teacher training, or available infrastructure by region were not incorporated. Including these factors could enrich the explanation of the observed variations.

The exclusion of socioeconomic and infrastructural variables has direct implications for the formulation of policies derived from this study. By relying solely on enrollment records, the projections reflect the internal dynamics of the educational system but do not capture the structural determinants that influence access and retention among students with disabilities. Therefore, recommendations related to resource allocation or service expansion should be considered of limited scope, as they may overestimate institutional capacity in regions with high poverty levels or deficient infrastructure.

These limitations are consistent with the warnings of Wankhede et al. (2024) and Alkan (2024a), who emphasize that the application of AI in education requires robust ethical frameworks and critical contextual analysis to avoid reproducing historical biases or technifying decisions without educational justice. Integrating contextual indicators in future models (such as regional investment, school density, or the availability of specialized professionals) would allow for the development of more precise, equitable, and sustainable policies aligned with the real conditions of each territory.

It is important to highlight that the results confirmed that the Random Forest model provided a better balance between accuracy and explainability, outperforming both traditional approaches (ARIMA) and black box models (ANN) in the multivariate context of this study. To advance the strategic use of predictive models in special education, it is recommended to incorporate structural and social variables that better reflect territorial conditions. Including indicators such as poverty level, degree of urbanization, regional budget for special education, and the number of specialized professionals would allow the development of more robust models that are sensitive to reality. It would also be useful to compare the performance of Random Forest with other algorithms such as XGBoost, neural networks, or hybrid models, as well as to develop interactive visualization platforms that serve as tools for public decision-making. Finally, it is suggested to institutionalize the use of explainable predictions within educational planning systems, integrating them with monitoring mechanisms and citizen participation.

6. Conclusion

This study provides robust empirical evidence on the effectiveness of the Random Forest model for projecting educational demand for students with special needs in Peru. The predictions generated using multivariate time series showed a sustained growth trend between 2019 and 2025, especially at the Preschool and Primary education levels.

The model achieved high levels of accuracy and consistency, validating its application in educational contexts characterized by high territorial variability.

The analysis also identified significant disparities by region and type of management, with a greater concentration of projected demand in public institutions and in urban regions such as Metropolitan Lima, Arequipa, and La Libertad. It should be noted that this territorial pattern was empirically estimated by the model based on the available historical data, while the interpretation of its concentration in urban areas corresponds to structural conditions of the educational system, such as the greater availability of services and specialized resources in those urban zones. These findings offer key insights for designing targeted policies that respond to the geographic and structural particularities of the education system.

Furthermore, the application of an explainable machine learning algorithm such as Random Forest facilitated the detection of complex nonlinear patterns and critical variables associated with demand trends, all without compromising the model's transparency. This analytical capacity makes it a valuable tool for evidence-based educational planning, particularly in dynamic and uncertain scenarios.

Finally, the methodology employed contributes not only to the academic field but also strengthens strategic decision-making aimed at inclusive, equitable, and territorially responsive education. This opens a promising path for the use of predictive models in managing education systems that seek to address diversity and the contemporary challenges of social development.

7. Declarations

7.1. Author Contributions

Conceptualization: RAGC; Methodology: EMCG; Software: WRP; Validation: RHPM and WRP; Formal analysis: RAGC, EMCG, and WRP; Investigation: RHPM; Resources: RAGC; Data curation: EMCG; Writing—original draft preparation: RAGC, EMCG, WRP, and RHPM; Writing—review and editing: RAGC, EMCG, WRP, and RHPM; Visualization: RHPM. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The tables supporting the findings of this study are publicly available in Figshare at <https://doi.org/10.6084/m9.figshare.29323010>. Additional data supporting the conclusions of this article are available from the corresponding author upon reasonable request.

7.3. Funding

This research was funded by Universidad Nacional Jorge Basadre Grohmann through the research project “Creation and design of specialized multimethod didactics to foster interest and learning of basic sciences in basic education students” (R.R. No. 10979-2023-UNJBG).

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] UNESCO, *Inclusion and education: All means all. Global Education Monitoring Report 2020*. Paris: UNESCO, 2020. [Online]. Available: <https://unesdoc.unesco.org/ark:/48223/pf0000373718>
- [2] W. J. Blanchett, J. K. Klingner, and B. Harry, "The intersection of race, culture, and disability: Implications for urban education," *Urban Education*, vol. 44, no. 4, pp. 389–409, 2009.
- [3] B. Harry and J. K. Klingner, *Why are so many minority students in special education? Understanding race and disability in schools*. New York: Teachers College Press, 2014.
- [4] T. A. Fiore, L. M. Harwell, J. Blackorby, and K. S. Finnigan, *Charter schools and students with disabilities: A national study. Final report (ED452657)*. Washington, DC: U.S. Dept. of Education, 2000. Available: <https://eric.ed.gov/?id=ED452657>
- [5] P. Lipman, *The new political economy of urban education: Neoliberalism, race, and the right to the city*. New York: Routledge, 2011. [Online]. Available: <https://core.ac.uk/download/pdf/236135746.pdf>
- [6] G. Stock and M. Schultz, "Medicaid expansions and special education enrollments: Evidence from administrative panel data," *Journal of Policy Analysis and Management*, vol. 44, no. 1, pp. 22–41, 2025.
- [7] Department for Education (DfE), *Special Educational Needs in England: January 2019*. London, UK: UK Government, 2019. [Online]. Available: <https://www.gov.uk/government/statistics/special-educational-needs-in-england-january-2019>
- [8] UNESCO, *Global Education Monitoring Report 2020: Inclusion and Education – All Means All*. Paris, France: UNESCO, 2020. [Online]. Available: <https://unesdoc.unesco.org/ark:/48223/pf0000373718>.
- [9] C. Nelson, P. Berman, R. Perry, and D. Silverman, *The state of charter schools 2000: Fourth-year report*. Washington, DC: U.S. Dept. of Education, 2000. [Online]. Available: <https://books.google.com.pe/books?id=fgrltWLdPXYC>
- [10] UNESCO Institute for Statistics (UIS), *Education and Disability: Analysis of Data on Children with Disabilities and Their Access to Education*. Paris, France: UNESCO, 2023. [Online]. Available: <https://uis.unesco.org/en/topic/disability-and-education>.
- [11] A. Camarillo, J. Rodríguez, and A. Salgado, "Predictive analytics for inclusive education: Data-based decision making in Latin America," *Comput. Educ. Artif. Intell.*, vol. 7, Art. no. 100275, pp. 1-12, 2024.
- [12] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [13] C. Molnar, *Interpretable machine learning: A guide for making black box models explainable*, 2nd ed. Leanpub, 2022. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/>
- [14] A. Alkan, "The role of artificial intelligence in the education of students with special needs," *International Journal of Technology in Education and Science*, vol. 8, no. 4, pp. 542–557, 2024.
- [15] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [16] H. Khosravi, S. Buckingham Shum, G. Chen, "Explainable artificial intelligence in education," *Comput. Educ. Artif. Intell.*, vol. 3, Art. no. 100074, pp. 1-12, 2022.
- [17] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, Art. e1355, pp. 1-12, 2020.
- [18] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *Int. J. Educ. Technol. High. Educ.*, vol. 16, Art. 39, no. 1, pp. 1-12, 2019.
- [19] U.S. Government Accountability Office (GAO), *Charter schools: Additional federal attention needed to help protect access for students with disabilities*. Washington, DC: GAO, 2012. [Online]. Available: <https://www.gao.gov/products/gao-12-543>
- [20] M. D. Broda, M. Bogenschutz, P. Dinora, S. M. Prohn, S. Lineberry, and E. Ross, "Using machine learning to predict patterns of employment and day program participation," *American Journal on Intellectual and Developmental Disabilities*, vol. 126, no. 6, pp. 477–491, 2021.
- [21] A. S. C. Tan, F. Ali, and K. K. Poon, "Subjective well-being of children with special educational needs: Longitudinal predictors using machine learning," *Appl. Psychol. Health Well-Being*, vol. 17, no. 1, Art. no. e12625, pp. 1-12, 2024.
- [22] L. T. Duda, J. Daniels, K. Zielinski, "Use of machine learning for behavioral distinction of autism and ADHD," *Transl. Psychiatry*, vol. 6, Art. no. e732, pp. 1-12, 2016.

-
- [23] L. Ghunaim, Y. Al-Shamma, H. Al-Ramahi, and O. Al-Qawasmi, "The Future of Pediatric Care: AI and ML as Catalysts for Change in Genetic Syndrome Management," *Jordan Medical Journal*, vol. 58, no. 4, pp. 510–528, 2024.
- [24] F. R. Waitoller, D. M. Maggin, and A. Trzaska, "A longitudinal comparison of enrollment patterns of students receiving special education in urban neighborhood and charter schools," *Journal of Disability Policy Studies*, vol. 28, no. 1, pp. 3–12, 2017.
- [25] H. Shao, J. Peng, and X. Liu, "Course enrollment prediction using machine learning: A case study," *Education and Information Technologies*, vol. 27, no. 4, pp. 5257–5277, 2022.
- [26] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, Art. e1355, pp. 1-12, 2020.
- [27] R. Tapio and D. Tarepe, "Comparative analysis of Random Forest and hybrid ARIMA Random Forest models for student enrollment forecasting in higher education," *Journal of Advances in Mathematics and Computer Science*, vol. 40, no. 3, pp. 124–136, 2025.
- [28] K. K. Adusei, "Modeling of municipal waste disposal behaviors related to meteorological and astronomical seasons using recurrent neural network models [Master's thesis]." The University of Regina (Canada), 2022. [Online]. Available: <https://www.proquest.com/openview/43375cf41062a40a8ab63a5fb34ad4eb/1?pq-origsite=gscholar&cbl=18750&diss=y>
- [29] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021. [Online]. Available: <https://otexts.com/fpp3/>
- [30] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, Art. e623, no. 1, pp 1-12, 2021.