ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.291

Technological Model Based on Machine Learning to Improve the Enrollment Process in Educational Institutions in Northern Lima

Ricardo Ñiquen[1]; Jesus Luna[2]; Juan Aguirre[3]
[1,2,3]Universidad Peruana de Ciencias Aplicadas, Perú, *u20191c983@upc.edu.pe, pcapjlun@upc.edu.pe, u20191e241@upc.edu.pe*

Abstract*– The enrollment process in public schools in Northern Lima is predominantly managed through manual forms, leading to inefficiencies, risk of information loss and administrative overload. This research proposes a machine learning-based model to optimize the management of enrollment records through risk-based classification, enabling the early identification of cases that may require additional administrative review. A dataset of 4,500 anonymized enrollment records collected between 2024 and 2026 from a public educational institution was used. Three supervised learning algorithms: Random Forest, Logistic Regression and Decision Tree were implemented and evaluated using a time-based 80/20 train-test split where results show that Random Forest achieved an accuracy of 60.6% and the lowest Brier Score (0.238), while Logistic Regression obtained the highest F1-Score (47.2%) and ROC-AUC (60.4%). Based on these results, Random Forest was selected due to its probability calibration and operational applicability. The proposed model enables the classification of enrollment records into risk levels, supporting administrative prioritization and improving decision-making in enrollment management. The study contributes to the digital transformation of the public education system in Peru, with a high potential for scalability at regional and national levels.

Keywords*-- Machine learning, predictive modeling, digital transformation, enrollment management, public education

I. Introduction

In recent years, digitalization has become an important tool to improve efficiency of administrative processes across many sectors. Nevertheless, on the Peruvian educational side, manual management still exists, this slows down the enrollment process, especially in public ones. According to data from the Educational Census [1], more than 90% of enrollment forms in Northern Lima are still managed on paper, leading to inefficient storage, risk of document losses and possible data duplication. This scenario is further exacerbated due to the high student density in public schools, which accounts for more than half of the students in the Unidad de Gestión Educativa Local 04 (UGEL 04), which is the local educational management unit of the sector of Northern Lima for the school sample on this research [2].

Many studies highlight the positive impact that digital transformation has had on educational processes, such as the NEIS (National Education Information System) in South Korea, as well as eKool in Estonia, which have demonstrated that digitalization not only improves administrative efficiency but also enhances data transparency and traceability [3], [4]. On the Peruvian context, the Ministry of Education has promoted guidelines to support digital transformation in public education [5]; however, despite these efforts, there is still no system that aims to optimize the registration, validation and prioritization of enrollment forms at an educational level.

In this context, machine learning emerges as a viable alternative as it allows the analysis of large volumes of information, identification of patterns and classification processes [6]. Also, recent academic literature shows that the performance of algorithms varies depending on the information included in the datasets, which makes even more critical the correct selection of techniques for a reliable and applicable model in education [8]. Additionally, international studies have shown that the integration of explainable models (XAI) have been successfully implemented to predict academic trajectories in high school students, providing transparency and confidence in the adoption of these technologies [7]. Applying this technology to the enrollment process would reduce human mistakes, expedite access to information and ensure the security of sensitive data of students. This paper proposes a technological model based on Random Forest that functions as an intelligent risk classification assistant in public schools in Northern Lima, taking in consideration its current limitations and contributing to the modernization of the Peruvian educational system.

II. Related Studies

Over the last few years, several studies have highlighted the impact of automation technologies on improving administrative processes within educational environments. Osco Pupe and Aguilar-Alonso propose a web architecture with integration of this technology in order to automate administrative workflows in a pre-university environment. This model allows for modularity, scalability and intelligent processing of academic information, significantly reducing processing time and dependence on human intervention [9]. Similarly, in the field of automated academic management, Viera-González et al. [10] implemented a low-code pipeline to manage scheduling-related incidents, achieving a reduction of service time by up to 80% while standardizing response processes. Although these studies do not focus specifically on enrollment processes, they demonstrate how automation can optimize administrative tasks, providing a relevant foundation for improving enrollment management in schools.

In addition to automation, several studies have analyzed the role of digital transformation in educational institutions. Palm, Asp and Håkansta [11] examined how the adoption of digital environments modifies the relationship between teachers, administrators and technology within the educational system; approximately 40% of participants reported that digital tools improved the effectiveness of work organization. Similarly, a recent study in Hong Kong researched

the role of digital leadership in the adoption of artificial intelligence (AI) within the educational context. Similarly, Cheng and Wang [12] conducted a quantitative study in 60 educational institutions with, approximately, 204 participants to evaluate the relationship between digital leadership, perceived barriers and AI integration strategies; this model presented high reliability levels ($\alpha = 0.836 – 0.966$) and a satisfactory fit (CFI = 0.916, RMSEA = 0.067), which suggests that there is a strong and positive relationship between digital leadership and AI adoption approaches. These findings emphasize the importance of institutional readiness and strategic direction in enabling technological integration. At a broader level, Dubois [13] analyzed digitalization processes within the European Union, highlighting how digital identity and interoperability mechanisms contribute to institutional transparency. The study reported a reduction of approximately 40% in administrative processing time, reinforcing the impact of digital transformation on organizational efficiency.

Machine learning techniques have been widely applied in educational contexts to model complex patterns and support data-driven decision-making processes. Ahmed et al. [14] applied ten regression model types based on machine learning (CatBoost, XGBoost, AdaBoost, Linear Regression, KNN), these were integrated into a model called Voting Regression, which reached a R^2 of 0.9890 on a first set of 10,000 records and R^2 of 0.7716 on a second set of 6,607 records, this approach enables the identification of the most influential variables in academic performance. Similarly, Delen et al. [15] combined deep learning models with traditional algorithms such as XGBoost and Random Forest, achieving an average accuracy above 85% and an AUC of 0.88, outperforming conventional statistical methods by approximately 12%. Their findings highlighted academic participation and prior performance as key predictive factors, reinforcing the importance of early-stage data in educational modeling. In this context, several studies have focused on the development of predictive systems using early-stage or pre-enrollment data. Carballo-Mendívil et al. [16] proposed an early warning system based on the XGBoost algorithm using pre-registration data from approximately 40,000 students. The model achieved an AUC-ROC of 0.6902, an F1-score of 0.6946 and a sensitivity of 88%, outperforming Random Forest and demonstrating the feasibility of transforming initial enrollment data into predictive indicators of dropout. Niyogisubizo et al. [17] introduced a two-layer machine learning architecture integrating Random Forest, Gradient Boosting, Logistic Regression and SVM, achieving an average accuracy of 94.8% and an AUC above 0.92. This ensemble-based approach demonstrated strong generalization capabilities, particularly when handling imbalanced datasets. Similarly, Seo et al. [18] developed predictive models using more than 144,000 longitudinal records from a South Korean university, where LightGBM achieved a ROC-AUC of 0.875, followed closely by XGBoost at 0.873, maintaining performance levels above 0.85 across multiple scenarios. Finally, Marcolino et al. [19] implemented a predictive model using CatBoost, optimized through a multi-objective NSGA-II algorithm and ADASYN data balancing techniques. The model achieved an F1-score of approximately 0.8, demonstrating a significant improvement in the detection of students at risk.

These studies demonstrate the effectiveness of machine learning techniques in educational data analysis, particularly in handling large-scale datasets and identifying key predictive variables and supporting early detection of risk-related patterns.

More specifically, several studies have explored the application of machine learning techniques in enrollment-related processes within the educational field. Abideen et al. [20] developed a predictive model to analyze enrollment criteria in secondary schools in Punjab, Pakistan. Using data collected from 100 institutions and a dataset of 500 records, multiple supervised learning algorithms were evaluated, where Random Forest achieved a precision of 98%, outperforming Decision Tree (90%) and SVM (88%). These results demonstrate the effectiveness of machine learning techniques in identifying key factors influencing enrollment decisions with high reliability. Similarly, He et al. [21] applied five machine learning classifiers to predict college enrollment among low-income students. The Random Forest model achieved an accuracy of 67.73%, followed by SVM and Gradient Boosting algorithms, confirming the ability of machine learning models to detect complex patterns in student admission processes. In the context of decision-making optimization, Liu et al. [22] proposed a machine learning-based approach to improve university admission processes by optimizing the number of students admitted and enrolled. The final model achieved an AUC of 0.86 and an adjusted Brier Score of 0.14, reducing variability in enrollment prediction by approximately 20%. Fiorini et al. [23] applied machine learning techniques, including Random Forest, Logistic Regression, and XGBoost, to analyze enrollment decision patterns and institutional assignment. Their model achieved an AUC of 0.91 and an overall accuracy of 87%, reducing predictive error by 18% compared to traditional statistical methods. These findings collectively demonstrate the effectiveness and reliability of machine learning techniques in modeling complex educational processes such as enrollment management.

III. Architecture Model

The architecture of the proposed model aims to represent the whole information workflow in a structured manner, from data collection to generating useful predictions for the administrative staff of the public school. The predictive approach considers multiple supervised machine learning algorithms, including Random Forest, Logistic Regression and Decision Tree. Among these, Random Forest is selected as the primary model due to its robustness and ability to capture non-linear relationships, while the other models are used as baseline references for comparative evaluation. This model will make decisions based on previously analyzed and observed patterns derived from historical data. It is built using multiple decision trees that, by combining their results, will generate more stable and reliable predictions. In Fig. 1, the methodology shown covers the process of constructing and developing the algorithm.

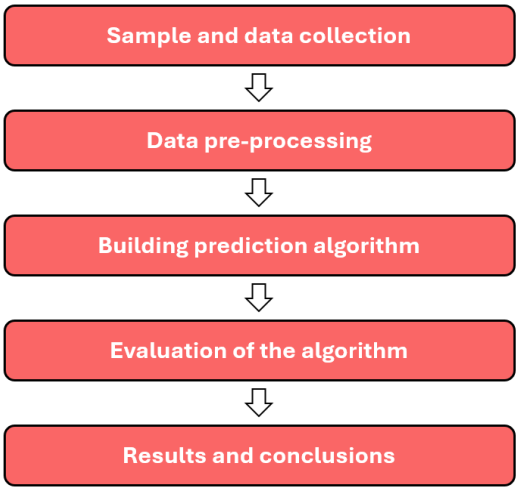

Fig. 1 Experimental process of the study
Adapted from [24]
Each step represents a stage, such as dataset normalization, algorithm construction and refinement of the predictive model.

The proposed architectural design aims to optimize the enrollment process through early detection of errors, automatic classification of requests and prioritization of cases based on their risk level. Therefore, this Random Forest model functions as an intelligent assistant that learns from the enrollment historical data to anticipate potential delays, suggest priorities and reduce administrative processing times. Furthermore, the feedback capability of the model allows it to be updated at the end of each enrollment period which takes place between January and March. New observations and results obtained during each operational cycle will be incorporated into the dataset in order to progressively improve the model's predictive capacity.

In summary, this architecture provides a robust framework for applying machine learning techniques in school administration, enabling not only the prediction of delays but also the digital transformation of the enrollment process into an intelligent digital workflow supported by data-driven decision-making and focused on operational efficiency.

IV. Method

The development of the model was carried out in the Google Colab platform, using standard Python libraries such as pandas, scikit-learn and matplotlib. In this environment, the dataset was imported, cleaned and prepared for the implementation, training and evaluation of multiple supervised machine learning models including Random Forest, Logistic Regression and Decision Tree. The process included the encoding of relevant categorical variables, handling of missing values and the transformation of selected variables derived from enrollment records.

A. *Sample and Data Collection*

For the construction of the predictive model, data were obtained from enrollment form records corresponding to the years 2024, 2025 and 2026, which were provided by the public school IE 2022 Sinchi Roca. These records were used as a reference to understand the structure and operational characteristics of the enrollment process. In this first stage, historical enrollment records were analyzed to identify relevant patterns and attributes associated with the administrative workflow. During this process, tasks such as detecting and removing duplicate entries, standardizing variable names and preparing the dataset structure were carried out.

To ensure privacy and data protection, all personal information such as student names, identification numbers and contact details were either removed or anonymized prior to analysis. As a result, a clean, privacy-compliant and consistent dataset of 4,500 records was obtained to move on to the next phase.

B. *Data pre-processing*

In this stage, missing values were handled using statistical imputation methods and numerical variables were standardized to ensure consistent scale across features, while categorical variables were encoded using One-Hot Encoding to allow their integration into the different machine learning models shortly after.

To maintain model consistency, only variables available at the initial stage of the enrollment process were included such as age, gender, school level, grade, number of submitted documents, type of identification document, district of residence and registration date.

To maintain a realistic prediction scenario, variables associated with post-process outcomes such as observations, revisions and total processing time were excluded from the feature set. Additionally, columns related to identifiers such as student and guardian information were removed from the training data as they do not contribute to predictive performance.

Therefore, the model receives data with greater coherence and predictive relevance, which enhances its accuracy and generalization capacity to operate in real-world enrollment scenarios.

This stage was implemented using a ColumnTransformer pipeline, where numerical features were standardized using StandardScaler and missing values were handled using SimpleImputer, while categorical variables were encoded using OneHotEncoder.

C. *Building prediction algorithm*

As shown in Fig. 2, the core of the predictive approach is based on the Random Forest algorithm, evaluated alongside Logistic Regression and Decision Tree as baseline models. In this model, each decision tree is trained on a random sample of the dataset and it evaluates a subset of the variables. In this context, the model learns to distinguish between enrollment records without administrative issues and those with a higher likelihood of risk. Each tree generates an individual classification and the final prediction is obtained by aggregating the outputs of all trees. This ensemble reduces dependence on a single decision rule.

The models were implemented using the scikit-learn library with specific hyperparameter configurations. No extensive hyperparameter tuning was performed; instead, configurations were selected based on standard practices to ensure model stability and generalization. The Random Forest classifier was configured with 200 decision trees (n_estimators = 200), a maximum depth of 10 levels (max_depth = 10), a minimum of 10 samples required to split an internal node (min_samples_split = 10) and a minimum of 5 samples required at each leaf node

(min_samples_leaf = 5). The model also incorporated class weighting (class_weight = "balanced") to address class imbalance and a fixed random state (random_state = 42) to ensure reproducibility.

The Logistic Regression model was implemented with a maximum of 1000 iterations (max_iter = 1000) and balanced class weights (class_weight = "balanced") to improve convergence and handle class imbalance.

The Decision Tree classifier was configured using a fixed random state (random_state = 42) and class weighting (class_weight = "balanced") to maintain consistency across models and address the imbalanced distribution of the target variable.

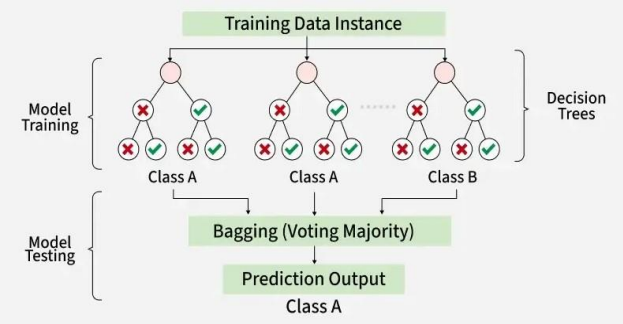

Fig. 2 Diagram of the Random Forest Model Operation
Adapted from [25]

Each decision tree learns specific patterns from the enrollment records (age, grade level, number of submitted documents, etc.).

D. Evaluation of the algorithm

The evaluation of the built model was carried out using a time-based validation strategy, which suggests using 80% of the historical records for training and the remaining 20% for testing in a controlled environment. This method reflects, in a more realistic way, the operational conditions in a real-world context within the educational institution. The final split consisted of 3,600 records for training and 900 records for testing. The target variable was defined as a binary classification indicating whether an enrollment record presents administrative risk (1) or not (0), based on observed patterns derived from historical enrollment data. The dataset is moderately imbalanced, with a higher proportion of non-risk cases. Out of the total 4,500 records, 65.73% correspond to non-risk cases (class 0), while 34.27% correspond to records with administrative risk (class 1). Precision, Recall and F1-Score were calculated using a binary classification approach, focusing on the positive class. Metrics were computed without macro or weighted averaging, in order to directly evaluate the model's ability to detect high-risk cases. Given the imbalanced nature of the dataset, accuracy alone is not sufficient to evaluate performance. Therefore, greater emphasis is placed on Precision, Recall and F1-Score.

To evaluate the classification performance and probability calibration of the models, a total of six metrics were used. The notation used is defined as follows: TP = true positives, TN = true negatives, FP = false positives, FN = false negatives, $\hat{p}$ = predicted probability, y = real value, N = total number of observations.

Additionally, baseline models such as Logistic Regression and Decision Tree were implemented to provide a comparative analysis of model performance, ensuring that the analysis does not rely exclusively on a single algorithm.

(1) Accuracy: Total percentage of correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

(2) Precision: Of the cases that the model marked as "risk", how many are actually risky.

$$Precision = \frac{TP}{TP + FP}. \quad (2)$$

(3) Recall (Sensitivity): Of all the cases that actually had observations, how many were detected by the built model.

$$Recall = \frac{TP}{TP + FN}. \quad (3)$$

(4) F1-Score: It is the harmonic mean between precision and recall, which helps measure the balance between both metrics.

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}. \quad (4)$$

(5) ROC-AUC: It measures the ability of the predictive model to distinguish between correct forms and those with observations; it is better as long as the metric gets closer to 1.

$$ROC - AUC = \int_0^1 TPR(FPR)\, d(FPR). \quad (5)$$

(6) Brier Score: It evaluates the calibration of probabilities where values closer to 0 indicates a good coherence.

$$Brier\ Score = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - y_i)^2. \quad (6)$$

Finally, the Random Forest model generates a risk probability on each enrollment record, which is then automatically transformed into a three-level classification. To ease the interpretation by the administrative staff, the following thresholds were defined for each enrollment form: "Fast Track" for low risk ($\hat{p} < 0.33$), "To Review" for medium risk ($0.33 \leq \hat{p} < 0.66$) and "Warning" for high risk ($\hat{p} \geq 0.66$).

V. RESULTS

The experimentation was conducted on the Google Colab platform, where the proposed methodology was carried out. The performance of the evaluated models was measured using the test dataset, which represents 20% of the most recent enrollment records (see Table I).

The comparative results obtained from the three evaluated models show that, Random Forest achieved the highest

accuracy of 60.6% and the lowest Brier Score (0.238), which indicates better calibration of predicted probabilities while Logistic Regression obtained the highest F1-Score of 47.2% and ROC-AUC of 60.4% suggesting a better balance between precision and recall, as well as a slightly stronger ability to discriminate between classes. Lastly, the Decision Tree model showed the lowest performance across most metrics, indicating limited generalization capability in the administrative context.

Although Logistic Regression achieved a higher F1-Score and ROC-AUC, Random Forest was selected as the primary model due to both technical and operational considerations. From a technical perspective, Random Forest is well-suited for handling datasets with mixed variable types and non-linear relationships, as is the case in this study, where both numerical and categorical features derived from enrollment forms are used. Additionally, its ensemble structure reduces variance and improves generalization, particularly in moderately sized datasets with class imbalance, such as the one analyzed.

From an applied perspective, Random Forest provides more stable and calibrated probability estimates, which are essential for the implementation of risk-based classification in administrative environments. Since the objective of the model is not only to classify records but to support prioritization decisions, the reliability of predicted probabilities becomes more critical than maximizing a single performance metric.

TABLE I
COMPARATIVE PERFORMANCE OF MACHINE LEARNING MODELS

Ref.	Metric	Random Forest	Logistic Regression	Decision Tree
(1)	Accuracy	**0.606**	0.576	0.580
(2)	Precision	**0.419**	0.404	0.365
(3)	Recall	0.465	**0.568**	0.346
(4)	F1-Score	0.441	**0.472**	0.355
(5)	ROC-AUC	0.593	**0.604**	0.522
(6)	Brier Score	**0.238**	0.243	0.420

In addition to classification performance, the Random Forest model was used to assign probability-based risk categories for each enrollment record in the test set.

As shown in Table II, once applied the classification to the 20% test set of the total enrollment records in the dataset, a predominance of the "To Review" level was observed, reflecting that the model identifies a high concentration of borderline cases that require administrative validation rather than clearly low-risk or high-risk scenarios.

This behaviour reflects a realistic operational context, where most enrollment forms are neither entirely error-free nor critically problematic, but require some level of review.

TABLE II
DISTRIBUTION OF RISK CATEGORIES ASSIGNED TO ENROLLMENT FORMS

Category	Count	Percentage
Fast Track	25	2.78%
To Review	857	95.22%
Warning	18	2.00%

VI. DISCUSSION

The predominance of the "To Review" category (95.22%) indicates that the model identifies a large number of borderline cases within the enrollment process. This behavior may be associated with the moderate class imbalance present in the dataset, where 65.73% of enrollment forms correspond to non-risk cases, as well as by the inherent variability of administrative patterns in early-stage enrollment data. Since the model relies exclusively on pre-registration variables, such as demographic and submission-related information, its predictive capacity is naturally constrained by the limited availability of contextual and process-related features. In this context, the obtained performance metrics reflect a realistic operational scenario, where most records are not clearly distinguishable as either low-risk or high-risk cases. Instead, they require some degree of administrative validation, which explains the concentration of predictions in the intermediate category. Therefore, rather than indicating a limitation in the model itself, these results highlight the complexity of decision-making processes in real-world enrollment environments and reinforce the importance of supporting administrative prioritization through probabilistic classification.

VII. CONCLUSIONS

The results obtained from the case study at IE 2022 Sinchi Roca show that the proposed machine learning-based model can support the prioritization of enrollment records using early-stage data. Among the evaluated algorithms, Random Forest achieved the highest accuracy (60.6%) and the lowest Brier Score (0.238), indicating better probability calibration, while Logistic Regression obtained the highest F1-Score (47.2%) and ROC-AUC (60.4%). Despite these differences, Random Forest was selected due to its robustness in capturing non-linear relationships, its ensemble-based stability and its superior calibration, which is essential for risk-based classification in administrative contexts.

The model enables the classification of enrollment records into three risk levels ("Fast Track", "To Review" and "Warning"), facilitating the prioritization of administrative tasks and reducing reliance on manual validation. Within the scope of the analyzed dataset, this represents a practical contribution by introducing a data-driven approach to support decision-making in the enrollment process. Although the results are limited to a single institutional context and pre-registration variables, the proposed approach provides a foundation for future extensions involving broader datasets and more complex feature sets.

VIII. LIMITATIONS

Despite the promising results obtained, this study presents some limitations. The dataset was obtained from a single public educational institution, which may limit the external validity of the model, as its performance could vary when applied to other educational contexts with different administrative processes or data characteristics.

Furthermore, the model relies exclusively on pre-registration variables, which restricts its predictive capability in scenarios where additional information becomes available during later stages of the enrollment process. Last-

ly, the implementation of this model in a real-world environment would require integration with existing institutional systems, as well as training for administrative staff to ensure effective adoption and support a successful digital transformation process.

IX. Recommendations

Even though the results obtained demonstrate moderate predictive performance of the Random Forest based model, there are opportunities for improvement aimed at increasing the algorithm's generalization capability and practical applicability.

First, future work could explore the optimization of classification thresholds to achieve a better balance between precision and recall, particularly for the identification of higher-risk cases.

Second, the incorporation of additional contextual variables, such as institutional or operational factors, could improve the model's predictive capacity and allow a more comprehensive representation of the enrollment process.

Finally, the integration of the algorithm with explainable artificial intelligence (XAI) techniques would enhance the interpretability of predictions, allowing administrative staff to better understand the factors influencing each classification. This would support institutional trust and the adoption of the algorithm in similar educational contexts, thereby supporting broader adoption in the Peruvian educational system.

References

[1] ESCALE, *Estadística de la calidad educativa: Censo Educativo 2024*. Ministerio de Educación del Perú, 2024. [Online]. Available: https://escale.minedu.gob.pe/inicio

[2] J. Contreras, "Análisis del crecimiento de instituciones educativas públicas en Lima Metropolitana," *Revista de Educación Peruana*, vol. 18, no. 2, pp. 45–59, 2024.

[3] S. Jongwon, "The role of NEIS in digital transformation of South Korean education," *Journal of Educational Technology Studies*, vol. 12, no. 4, pp. 77–89, 2023.

[4] *Nation Africa*, "Estonia's eKool digital school platform expands to Kenya," Nation Media Group, 2023. [Online]. Available: https://nation.africa

[5] MINEDU, *Resolución Ministerial N.° 115-2023-MINEDU: Plan de Gobierno y Transformación Digital 2023-2025*. Ministerio de Educación del Perú, 2023. [Online]. Available: https://www.gob.pe/institucion/minedu/normas-legales/4203311

[6] M. A. Filz, J. P. Bosse and C. Herrmann, "Digitalization platform for data-driven quality management in multi-stage manufacturing systems," *Journal of Intelligent Manufacturing*, vol. 35, no. 6, pp. 2699–2718, 2023. doi: 10.1007/s10845-023-02056-5

[7] M. Abdalkareem and N. Min-Allah, "Explainable models for predicting academic pathways for high school students in Saudi Arabia," *IEEE Access*, vol. 12, pp. 30604–30621, 2024, doi: 10.1109/ACCESS.2024.3369586

[8] D. Oreški, I. Pihir and D. Višnjić, "Comparative analysis of machine learning algorithms on data sets of different characteristics for digital transformation," in *Proc. MIPRO 2023*, Opatija, Croatia, May 22–26, 2023, pp. 1428–1432, doi: 10.1109/MIPRO58648.2023.10444523

[9] J. W. Osco Pupe y I. Aguilar-Alonso, "Web System as Support to Automate Processes of the Administrative Area of the Pre-University Center," IEEE International Conference on Modern Trends in Information and Communication Technology Industry (MTICTI), 2021.

[10] P. M. Viera-González, O. H. González-González, y G. E. Sánchez-Guerrero, "Low-Code Automation for Resolving Incidents During Class Schedule Registration," Educational Systems Automation Review, Univ. Autónoma de Nuevo León, México, 2023.

[11] K. Palm, A. Asp, y C. Håkansta, "Implementing digital technologies in the school setting – how does it relate to work environment?," *Education and Information Technologies*, vol. 29, no. 5, pp. 1–22, 2023.

[12] Y. Cheng and J. Wang, "Leading Digital Transformation and Eliminating Barriers for Teachers to Incorporate Artificial Intelligence in Basic Education in Hong Kong," *Computers and Education: Artificial Intelligence*, vol. 5, art. no. 100171, 2023, doi: 10.1016/j.caeai.2023.100171.

[13] F. Dubois, "The Digitalisation of European Union Procedures: A New Impetus Following a Time of Prolonged Crisis," *European Law and Technology Review*, vol. 11, no. 1, pp. 14–29, 2022.

[14] W. Ahmed, S. H. M. Ali, F. Hassan and T. W. Khan, "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," *Scientific Reports*, vol. 15, no. 26879, 2025, doi: 10.1038/s41598-025-12353-4.

[15] D. Delen, B. Davazdahemami and E. R. Dezfouli, "Predicting and Mitigating Freshmen Student Attrition: A Local-Explainable Machine Learning Framework," *Information Systems Frontiers*, vol. 25, 2023, doi: 10.1007/s10796-023-10344-8.

[16] B. Carballo-Mendívil, A. Arellano-González, N. J. Ríos-Vázquez and M. del P. Lizardi-Duarte, "Predicting Student Dropout from Day One: XGBoost-Based Early Warning System Using Pre-Enrollment Data," *Applied Sciences*, vol. 15, no. 16, p. 9202, 2025, doi: 10.3390/app15169202.

[17] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers & Education: Artificial Intelligence*, vol. 3, no. 4, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.

[18] E.-Y. Seo, J. Yang, J.-E. Lee and G. So, "Predictive modelling of student dropout risk: Practical insights from a South Korean distance university," *Heliyon*, vol. 10, no. 11, e30960, 2024, doi: 10.1016/j.heliyon.2024.e30960.

[19] M. R. Marcolino, T. R. Porto, T. T. Primo, R. Targino, V. Ramos, E. M. Queiroga, R. Muñoz and C. Cechinel, "Student dropout prediction through machine learning optimization: Insights from Moodle log data," *Scientific Reports*, vol. 15, no. 9840, 2025, doi: 10.1038/s41598-025-93918-1.

[20] Z. U. Abideen, T. Mazhar, A. Razzaq, I. Haq, I. Ullah, H. Alasmary and H. G. Mohamed, "Analysis of Enrollment Criteria in Secondary Schools Using Machine Learning and Data Mining Approach," *Electronics*, vol. 12, no. 3, pp. 1–25, 2023, doi: 10.3390/electronics12030694.

[21] S. He, M. Y. Naeim, Y. Cui and M. Cutumisu, "Predicting College Enrollment for Low-Socioeconomic-Status Students Using Machine Learning Approaches," *Big Data and Cognitive Computing*, vol. 9, no. 4, p. 99, 2025, doi: 10.3390/bdcc9040099.

[22] T. Liu, C. Schenk, S. Braun and A. Frey, "A Machine-Learning-Based Approach to Informing Student Admission Decisions," *Behavioral Sciences*, vol. 15, no. 3, p. 330, 2025, doi: 10.3390/bs15030330.

[23] S. Fiorini, G. D'Amico, L. Gentile and R. Martini, "A Machine Learning Approach to Predict University Enrolment Choices Through Students' High School Background in Italy," *Education Sciences*, vol. 13, no. 5, p. 512, 2023, doi: 10.3390/educsci13050512.

[24] Thao-Trang Huynh-Cam, Long-Sheng Chen and Huynh Le, "Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance," *Algorithms*, vol. 14, no. 11, p. 318, 2021, doi: 10.3390/a14110318

[25] "Random Forest Algorithm in Machine Learning," GeeksforGeeks, [Online]. Available: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/