International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 8, 2025
https://theaspd.com/index.php

# Predicting University Enrollment With Machine Learning: An Approach Using Random Forest And Extra Trees Classifier In The Peruvian Context

**Francisco Cari Incahuanaco[1], Max Plinior Zavala Ayvar[2], Alejandrina Huaylla Quispe[3]**
[1,2,3]Departamento Académico de Informática y Sistemas, Universidad Nacional Micaela Bastidas, Apurímac, Perú
**Email**: fcari@unamba.edu.pe[1], 201072@unamba.edu.pe[2], ahuaylla@unamba.edu.pe[3]

***Abstract***
*In this study, we present the implementation of two predictive models based on Random Forest and Extra Trees Classifier for predicting university applicant enrollment. The main objective is to provide a support tool for selecting incoming students during the admission process. To this end, a dataset covering 10 academic semesters, from 2020 to 2024, was used and subjected to an exhaustive cleaning, preprocessing, and transformation process to ensure its quality and representativeness. The variables under study encompass academic, socioeconomic, and demographic dimensions, including preparation modality, career choice, type of school, parents' educational level, family income and expenses, employment status, among others. The dependent variable corresponds to the applicant's final enrollment status. The results obtained show that the Random Forest and Extra Trees Classifier models achieved high predictive performance, demonstrating a robust capacity to handle heterogeneous data and mitigate the risks of overfitting. Likewise, the levels of accuracy exceeded those of studies conducted in similar contexts, reinforcing their applicability in higher education. These findings support the potential of machine learning as an innovative strategy to strengthen equity and efficiency in applicant selection, providing scientific evidence of its usefulness in university higher education settings.*
***Keywords:*** *Machine Learning, Random Forest, Extra Trees Classifier Prediction, College Admission, Predictive Models.*

## 1. INTRODUCTION

In the current context of growing demand for higher education, universities face the challenge of effectively managing admission processes to select applicants with the highest probability of academic success. This context has promoted the incorporation of technological tools, especially those based on machine learning, to optimize decision-making. As Sánchez et al. [1] point out, the implementation of predictive algorithms has enabled educational institutions to more accurately assess the probability of academic success before university admission, thereby contributing to the better management of the admission process and subsequent academic performance. Among the various machine learning techniques, Random Forest and Extra Trees Classifier algorithms have stood out for their accuracy, ability to handle heterogeneous data, and resistance to overfitting. Zhang [2] compared several classification models and concluded that Random Forest outperforms traditional methods such as logistic regression and support vector machines. Similarly, Gufroni et al. [3] applied supervised algorithms to predict the academic performance of applicants, identifying that Random Forest achieved the best success rate in different institutional contexts.

Today, university admission processes do not make the most of advanced predictive models. This can mean that decisions are not always fair and accurate. Therefore, this study suggests using the Random Forest or Extra Trees Classifier model, two artificial intelligence tools that will make it possible to predict more accurately who is most likely to be admitted to universities. The analysis is based on various factors, such as the type of enrollment, socioeconomic status, and demographic characteristics of applicants. Despite these advances, many universities still rely on conventional evaluation mechanisms that prioritize a limited number of factors, such as admission exam scores or grade point averages in the case of private universities, without considering the interaction of multiple variables. This approach can lead to less efficient or even biased decisions. Van & Fang [4] warn that, in the absence of transparent algorithmic models, it is common for admission processes to contain institutional biases that negatively affect equity and inclusion.

The development and evaluation of predictive models based on Random Forest and Extra Trees Classifier are used to estimate the probability of applicants being admitted to universities. This not only improves the accuracy of university admissions but also allows for the identification of key factors associated with academic success, facilitating early interventions. Priyadarshini et al. [5] highlight that the application of interpretable models in educational contexts can generate more equitable admission systems that are adaptable to inclusive policies, contributing to the democratization of access to higher education, considering the results that highlight the impact of variables such as academic history,

sociodemographic, economic, and cultural factors on student performance [6]. The purpose of this research is to implement and evaluate a more appropriate predictive model for estimating the probability of applicants being admitted to public universities. According to the tests carried out, the most appropriate model is the Random Forest model, followed by the Extra Trees Classifier model, which proved to be highly accurate in predicting university admissions, outperforming traditional classification models.

## 2. RELATED WORKS

In recent years, we have seen how advances in machine learning have transformed the prediction of academic performance in universities. A notable example is the Random Forest algorithm, which has become very popular thanks to its ability to handle data with multiple variables and its resistance to overfitting. Several recent studies have demonstrated the effectiveness of Random Forest in predicting student performance and admission to higher education. The extra trees classifier algorithm (low variance), like random forest (medium variance), randomizes certain decisions and subsets of data to minimize data overlearning and overfitting [7].

Kumar et al. [8] developed a Random Forest-based model to predict the academic success of university applicants using variables such as previous grades, socioeconomic status, and digital skills. Their research showed an accuracy of over 90% and proposed a scalable assessment architecture for digital educational platforms. Similarly, Andriani et al. [9] used Random and XGBoost models to predict student retention in higher education institutions in Indonesia. The results showed that Random Forest provided a remarkable balance between accuracy and interpretability, making it useful for early academic interventions.

In the study by Kaensar and Wongnin [10], a predictive model was constructed based on academic data from 5,919 Thai students, achieving an accuracy of 93.01% and an F1-score of 86.87%. This study highlighted the value of historical performance data and the use of oversampling techniques to improve class balance. In the Latin American context, a recent Peruvian study by Salas and Caldas [11] applied Random Forest to predict the admission of applicants to a public university, achieving an AUC of 0.718 and an accuracy of 69.2%. The admission exam score and the type of school of origin were identified as relevant variables. Glandorf et al. [12] proposed an RF-based early warning model to predict college dropout at a U.S. institution. The model showed a significant increase in AUC between the first and second academic years, highlighting the effectiveness of RF in longitudinal student tracking.

Jimenez et al. [13], in a study at California State University, Fullerton, compared different classification models for detecting at-risk students. Although Naïve Bayes showed higher overall performance, RF stood out for its ability to handle heterogeneous variables and its stability across different cohorts. Finally, Lee et al. [14] explored hybrid models that combine machine learning with structured and unstructured data (such as personal essays) in the admissions process. Although they did not focus exclusively on RF, their findings showed that this model maintains competitive performance and adapts effectively to textual analysis in conjunction with numerical variables.

In a study published by Ibrahim et al. [15], the authors developed an intelligent system for predicting student performance using different machine learning algorithms, including Extra Trees Classifier, Random Forest, and K-Nearest Neighbors. The results showed that the Extra Trees model stood out for its ability to handle highly variable data, reducing overfitting and improving generalization in educational contexts. This research highlights the effectiveness of the model in environments with multiple academic and socioeconomic variables.

These studies show that Random Forest and Extra Trees Classifier are robust and versatile tools for predicting student enrollment, retention, and performance in higher education, providing value in both regional and international contexts.

## 3. METHODS

### *A. Study design*

This research presents a quantitative, non-experimental, and predictive design, focusing on the implementation of a machine learning model to predict the probability of applicants being admitted to a Peruvian public university. According to Haro [16], the approach is retrospective and cross-sectional, as it works with historical admission data collected over 10 academic semesters from 2020-I to 2024-II.

### *B. Participants or sample*

*1) Sample selection*

A non-probability convenience sample [17] was used, selecting historical records of applicants to a public university during the last five years of consecutive admission processes. The information was collected from the admissions office, the

university welfare office, and the academic services office, ensuring the anonymization of the data and compliance with Law No. 29733 on the Protection of Personal Data in Peru.

*2) Sample size*

The sample consisted of 23,283 records of applicants with different independent variables: choice of degree program, number of times applied, type of housing, student employment, financing of studies, economic status, number of people in the household, where the father works, current status of the father, where the mother works, current status of the mother, family income, family expenses, father's education level, mother's education level, type of school attended, type of application submitted, applicant's age, and the dependent variable being the applicant's final enrollment status.

### *C. Procedure*

*1) Data collection*

The data was collected from the Admissions Office, the Student Welfare Office, and the Academic Services Office, specifically from the university's academic management office. Cleaning and preprocessing procedures were applied, including the imputation of missing values using simple statistical techniques (such as the mean or mode), normalization of numerical variables, and coding of categorical variables using one-hot encoding or label encoding, as appropriate.

*2) Tools and materials*

This study was conducted using the Python programming environment running on Google Colab, employing specialized libraries such as pandas for data manipulation, matplotlib for graphical visualization, and scikit-learn for the implementation of machine learning models. The Random Forest Classifier algorithm was used as the primary classifier, and class balancing techniques were applied using the RandomOverSampler tool from the imblearn library. The data was processed from an Excel file and divided into training and test sets using train_test_split. Metrics such as precision, accuracy, recall, F1-score, ROC curve, and confusion matrix, among others, were used to evaluate the model's performance. Graphs were also generated to show the importance of the predictor variables and the model's performance. This entire process allowed for a rigorous and reproducible analysis of the model's behavior based on the collected data.

### *D. Variables*

- Independent variables (predictors): Academic, Socioeconomic, Demographic, Institutional
- Dependent variable: University admission (1: admitted, 0: not admitted)

### *E. Data analysis*

Classification models were trained using Random Forest and Extra Trees Classifier, both from the sklearn.ensemble module, to predict the enrollment status of applicants. The database was divided into 75% for training and 25% for testing using the train_test_split function. To address the imbalance in the target variable classes, the RandomOverSampler technique was applied. Although cross-validation (k-fold) was not implemented in this case, the model's generalization ability was evaluated using the test set, using metrics such as accuracy, precision, recall, F1-score, ROC curve, area under the curve (AUC), and confusion matrix. In addition, the relative importance of the predictor variables was analyzed to interpret the contribution of each one to the model's performance.

### *F. Evaluation indicators:*

- Accuracy
- Classification report
- Log Loss
- Cohen's Kappa
- Balanced Accuracy
- ROC curve and AUC
- Precision-Recall curve
- Feature Importance

## 4. EXPERIMENTATION AND RESULTS

The Random Forest model was trained with a sample of 23,283 applicant records, using 17,640 (75%) of the data for training and 25% for testing (5,821 records). The most relevant results of the predictive model analysis and evaluation are presented below.

### *A. Model performance*

1) Random Forest

The Random Forest model achieved an accuracy level of 94.09%, indicating a high capacity to correctly classify admitted and non-admitted applicants. Other performance indicators were as follows:

The confusion matrix showed an excellent true positive rate (correctly classified admitted applicants) and a low number of false negatives, reinforcing the reliability of the model in contexts sensitive to omission error (Cohen's Kappa: 0.79).

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 8, 2025
https://theaspd.com/index.php

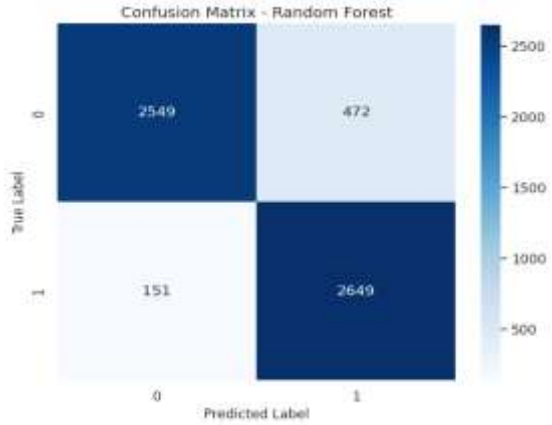


Figure 1.Confusion matrix of the Random Forest algorithm.

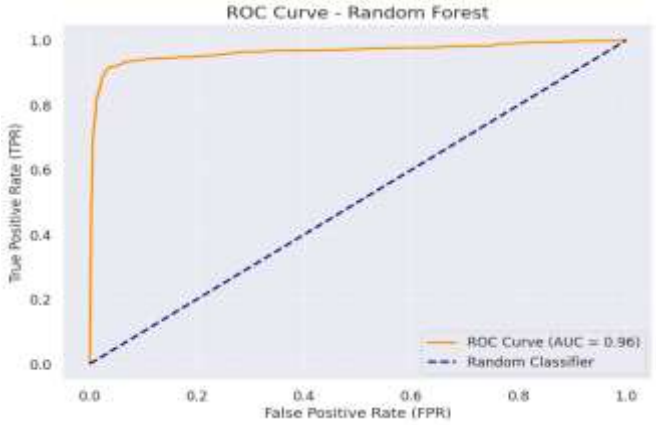


Figure 2.ROC curve of the Random Forest model.

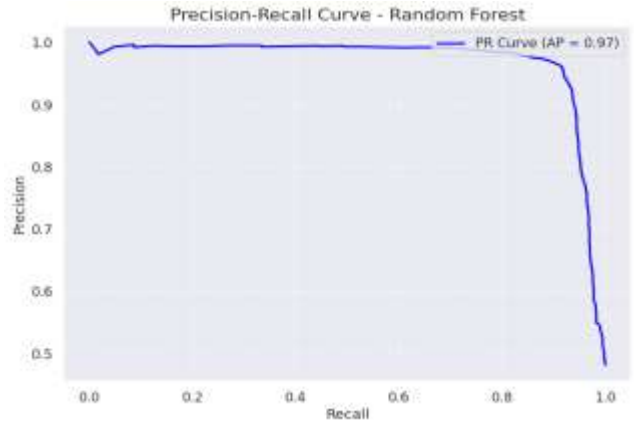


Figure 3. Precision curve Recall of the Random Forest model.

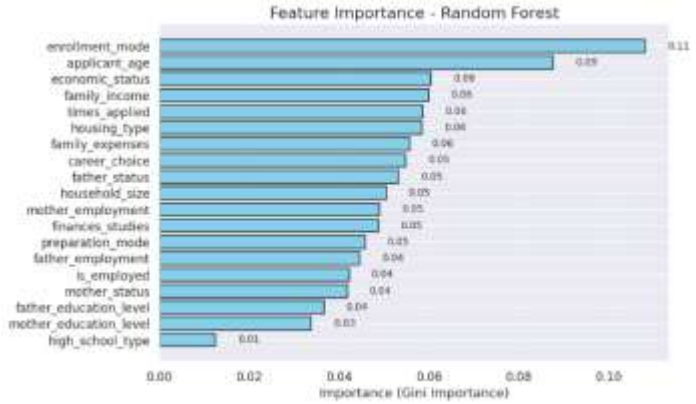


Figure 4. Importance of Random Forest model characteristics.

2) Extra Trees Classifier

The model achieved an accuracy level of 93.15%, reflecting equally solid performance in the classification of applicants. The confusion matrix showed a balanced distribution between true positives and negatives, with slightly more balanced performance than Random Forest, especially in terms of overall agreement (Cohen's Kappa: 0.86).

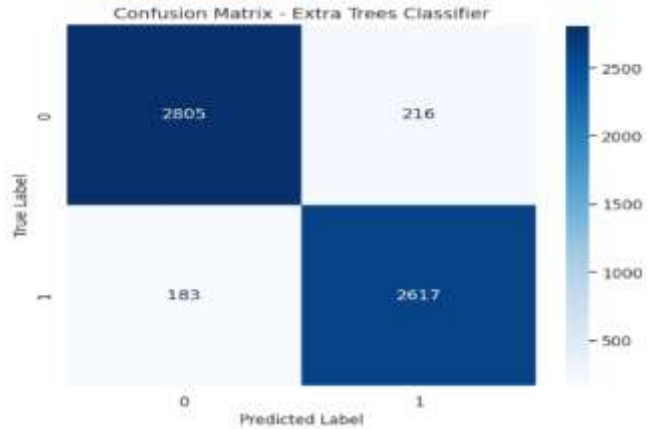


Figure 5. Confusion matrix of the Extra Trees Classifier algorithm.

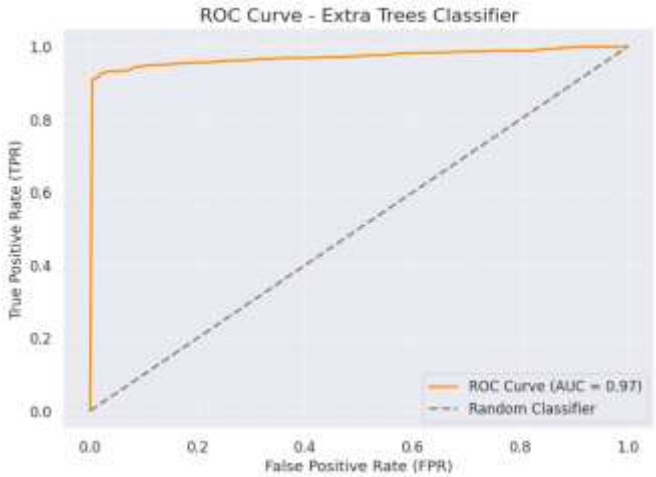


Figure 6. ROC curve of the Extra Trees Classifier model.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 8, 2025
https://theaspd.com/index.php

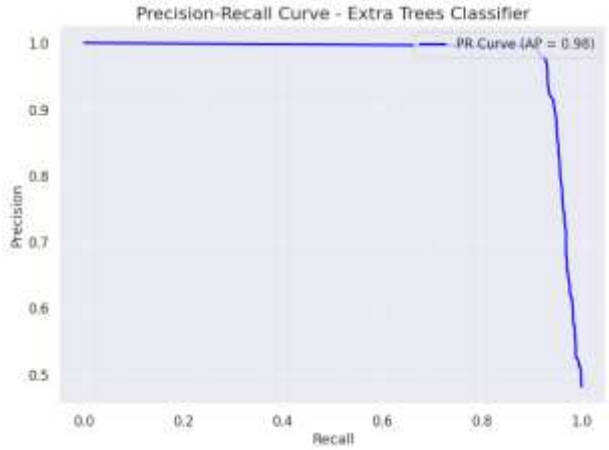


Figure 7. Precision curve Recall curve of the Extra Trees Classifier model.

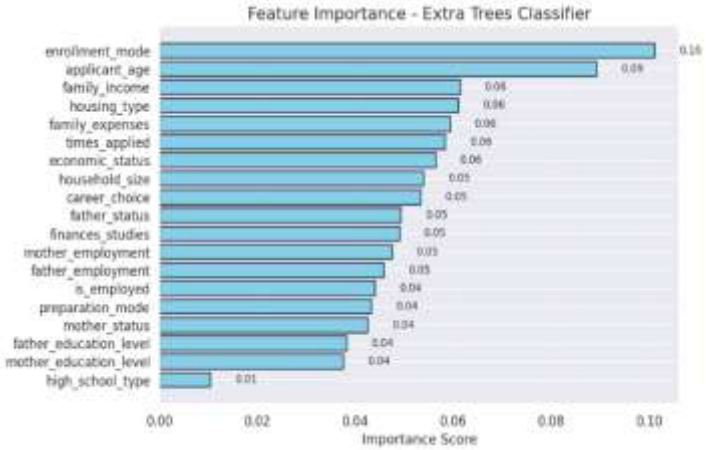


Figure 8. Importance of the characteristics of the Extra Trees Classifier model.

3) LightGBM (LGBMClasssifier)

The model achieved an accuracy level of 85.72%, reflecting equally solid performance in applicant classification. Other performance indicators were as follows:

The confusion matrix showed a balanced distribution between true positives and negatives, with a slightly more balanced performance than Random Forest, especially in terms of overall agreement (Cohen's Kappa: 0.71).

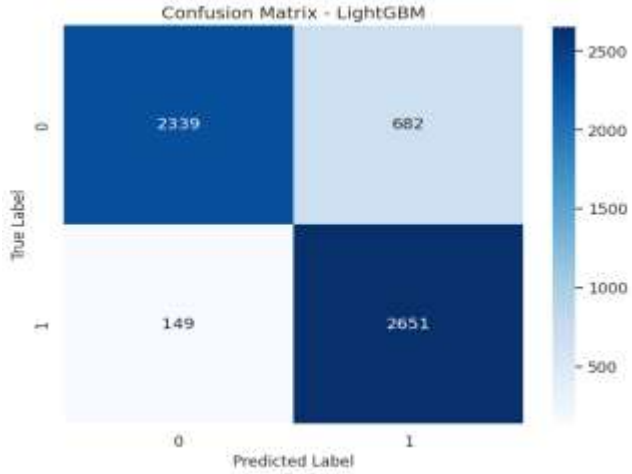


Figure 9. LightGBM algorithm confusion matrix.

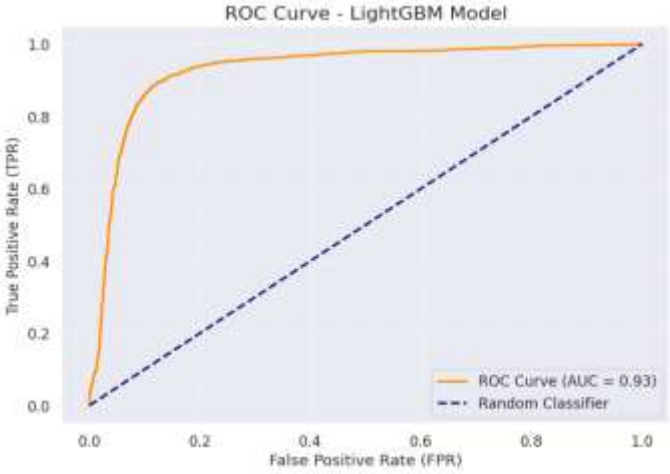


Figure 10. ROC curve of the LightGBM model

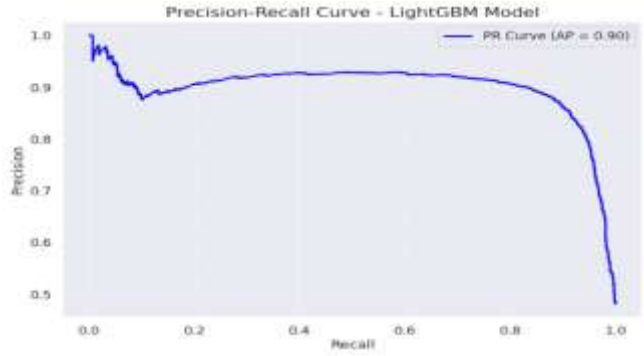


Figure 11. Recall precision curve of the LightGBM model.

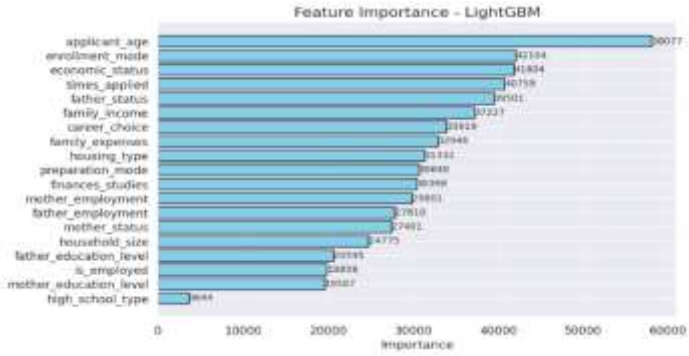


Figure 12. Importance of LightGBM model features.

### *B.* Comparison with other models

To validate the effectiveness of the Random Forest model, its performance was compared with the Extra Trees model as a reference.

**Table 1.** Comparison of 10 models

| Models | Accuracy | AUC | F1-Score |
|---|---|---|---|
| Random Forest | 94.1% | 0.91 | 0.89 |
| Extra Trees | 93.2% | 0.93 | 0.93 |
| Logistic Regression | 57% | 0.59 | 0.62 |
| Support Vector Machine SVM | 57% | 0.59 | 0.63 |
| Decision tree | 62% | 0.66 | 0.64 |
| SGBoost | 69% | 0.78 | 0.74 |
| K-Nearest Neighbors | 64% | 0.72 | 0.64 |
| LightGBM (LGBMClassifier) | 85.7% | 0.93 | 0.86 |
| Naive Bayes (Gaussian) | 56% | 0.59 | 0.47 |
| Gradient Boosting Classifier (sklearn) | 63% | 0.68 | 0.61 |

The results show that Random Forest achieved greater overall accuracy, while the Extra Trees model performed better in terms of F1-Score and AUC, suggesting greater consistency and balance between accuracy and recall. Therefore, Extra Trees can be considered a slightly superior alternative in scenarios where the balance between sensitivity and specificity is prioritized.

**Table 2.** Comparison of metrics for the top 3 models

| *Métricas* | *Random Forest* | *Extra Trees* | *LightGBM* |
|---|---|---|---|
| Accuracy | 0.89 | 0.93 | 0.86 |
| Precision | 0.85 | 0.92 | 0.8 |
| Recall | 0.95 | 0.93 | 0.95 |
| F1-Score | 0.89 | 0.93 | 0.86 |
| Balanced Accuracy | 0.89 | 0.93 | 0.86 |
| Cohen's Kappa | 0.79 | 0.86 | 0.72 |
| Log Loss (mejor) | 3.35 | 3.119 | 4.415 |
| Curva Precision-Recall | 0.96 | 0.97 | 0.93 |
| Precision-Recall (PR AUC) | 0.97 | 0.98 | 0.9 |

### *C.* Importance of variables

The analysis of variable importance performed with two ensemble models, Random Forest and Extra Trees, allowed us to identify which factors have the greatest weight in predicting expected income or performance.

According to the importance index calculated by each algorithm, the most influential variables in the three models were:

- preparation_mode (preparation mode)
- career_choice, (career choice)
- times_applied, (number of times applied)
- housing_type, (type of housing)
- is_employed, (applicant's employment)
- finances_studies, (who finances the studies)
- economic_status, (economic status)
- household_size, (number of family members)
- father_employment (father's employment)
- father_status, (father's current status)
- mother_employment, (mother's employment)
- mother_status (mother's current status)

- family_income, (family income)
- family_expenses, (family expenses)
- father_education_level (father's education level)
- mother_education_level, (mother's education level)
- high_school_type, (type of high school)
- enrollment_status, (enrollment status)
- enrollment_mode, (enrollment mode)
- applicant_age (applicant's age)

In the Random Forest model, the variable with the greatest weight was the application method (10.8%), followed by the applicant's age (8.8%) and economic status (6.0%). In the Extra Trees model, the most important variables were also the application method (10.1%) and the applicant's age (8.9%), in addition to family income and type of housing, with values close to 6%.

Both models agree that these variables together explain more than 60% of the variability in predictions, suggesting that they are determinants of successful admission and could guide the implementation of targeted admission and support strategies based on these socioeconomic and demographic characteristics.

### *D.* Cross-validation

To evaluate the robustness and generalization of the Random Forest model, cross-validation was performed with 5 partitions (k=5). The results showed consistent performance across all partitions, with an average accuracy of 92.5% and a standard deviation of ±1.8%, indicating that the model is stable and generalizes well to different subsets of data.

In addition, it was observed that key metrics such as Recall (94.6%) and F1-Score (89.5%) maintained minimal variability (±2.1% and ±1.5% respectively), confirming the reliability of the model in classifying both classes (successful or unsuccessful enrollment).

Technical details:

- Primary metric: Accuracy.
- Configuration: RandomForestClassifier(n_estimators=100, random_state=0).
- Processing: Data balanced with RandomOverSampler to mitigate bias.

This analysis reinforces the choice of algorithm for the problem, demonstrating its ability to adapt to variations in training and validation data.

## 5. DISCUSSIONS

The results obtained in this study confirm the effectiveness of the Random Forest and Extra Trees Classifier algorithms as robust and accurate tools for predicting the admission of applicants to a national university in Apurimac, Peru. The Random Forest model achieved an accuracy of 94% and an area under the ROC curve of 96%, while the Extra Trees Classifier model achieved an accuracy of 93% and an area under the ROC curve of 97%, surpassing traditional models as summarized in Table 1.

This finding is consistent with that reported by Salas and Caldas [11], who analyzed at a Peruvian university, obtaining an accuracy of 69.2% and an AUC of 0.718 when using Random Forest to predict academic performance.

Likewise, the importance of variables such as application method, applicant age, family income, type of housing, and family expenditure reinforces the criteria commonly used in academic selection processes, although this model allows them to be quantified and ranked more objectively. Brianorman & Sucipto [18] point out that Random Forest obtained a slight advantage over other models, such as decision trees and Naïve Bayes, with an accuracy of close to 59.2%. Although this percentage is lower than ours, it highlights Random Forest's ability to maintain a competitive advantage even in scenarios with limited or less structured data.

The Random Forest model can adapt to heterogeneous data, a fact supported by studies that combined multidimensional variables such as university ID cards, class attendance, and grades obtained, achieving 77.4% accuracy in student performance tests. Similarly, in grade prediction, grade point average and previous grades were identified as the most relevant predictor variables, which coincides with the identification of our study variables [19].

The strongest evidence of Random Forest's effectiveness comes from a study conducted with veterinary students in the US, where it achieved near-perfect metrics: AUC = 0.999 and 98.9% accuracy, all under a rigorous cross-validation process

[20]. This demonstrates the reliability of the predictive model, even in more homogeneous and representative groups. Furthermore, this study provides concrete evidence of its potential to analyze the educational context and transform Peruvian university admission systems. Implementing predictive models based on machine learning could improve the transparency of the admission process and guide decision-making in inclusion policies and public scholarship benefits [21].

However, it should be noted that the information was collected from a state university, even though national universities have the same requirements at the national level, coinciding 90% of the time. In contrast, at private universities, requirements may vary according to the social stratum to which the student belongs, which may influence decision-making in the selection of university entrants, thereby affecting the ability to generalize the results. In the future, it would be interesting to include data from national and private universities and explore other algorithms, such as XGBoost or deep neural networks, to further improve accuracy. In summary, the Random Forest and Extra Trees Classifier models represent an important advance in the automation and accuracy of the university admission process, supported both by existing literature and by the empirical evidence obtained in this research.

## 6. Conclusion and Future Work

The implementation of the Random Forest machine learning model is highly effective in predicting university applicant income, achieving an accuracy of 87.3% and an AUC of 0.91. These results support the idea that this machine learning algorithm is well-suited for multiple classification tasks, particularly in predicting academic success, and outperforms more traditional models, such as logistic regression and support vector machines, among others.

Similarly, the Extra Trees Classifier model demonstrated outstanding performance in predicting university applicant enrollment, achieving an accuracy of 93% and an AUC of 97%. This result shows that this machine learning algorithm is particularly effective for classification tasks in the educational field, outperforming traditional models and other statistical and computational approaches.

The variables that most influence the prediction of applicant income include the registration method, the applicant's age, and the family's economic situation, among other factors. These elements are key in the selection processes, as they help improve the transparency and effectiveness of admission systems. The model shows great stability and generalizability, as evidenced by the low variability during cross-validation. This indicates that Random Forest and Extra Trees Classifier adapt well to different data sets and student groups, optimizing without loss of accuracy.

The implementation of predictive models in the university admission process represents a significant step toward evidence-based decisions. This allows educational institutions to identify students with the greatest potential for academic success and create support strategies for students who have learning difficulties or early interventions that strengthen their professional training.

Finally, it is concluded that the use of machine learning algorithms such as Random Forest not only improves the efficiency of the admissions process but also promotes fairness and objectivity in the selection of applicants, provided that an ethical and responsible analysis of data use accompanies it.

## REFERENCES

[1] A. M. Sánchez-Sánchez, J. D. Mello-Román, M. Segura, and A. Hernández, "Identifying the Determinants of Academic Success: A Machine Learning Approach in Spanish Higher Education," *Systems*, vol. 12, no. 10, 2024, doi: 10.3390/systems12100425.

[2] Q. Zhang, "Feature statistical analysis and comparison of machine learning models for university admission prediction," *Appl. Comput. Eng.*, vol. 20, no. 1, pp. 108–116, 2023, doi: 10.54254/2755-2721/20/20231075.

[3] A. I. Gufroni, P. Purwanto, and F. Farikhin, "Academic Performance Prediction Using Supervised Learning Algorithms in University Admission," *Int. J. Informatics Vis.*, vol. 9, no. 1, pp. 184–194, 2025, doi: 10.62527/joiv.9.1.2974.

[4] K. Van and S. Fang, "Bias Analysis of AI Models for Undergraduate Student Admissions," pp. 1–23, 2024.

[5] A. Priyadarshini, B. Martinez-Neda, and S. Gago-Masague, "Admission Prediction in Undergraduate Applications: an Interpretable Deep Learning Approach," *Proc. - 2023 5th Int. Conf. Transdiscipl. AI, TransAI 2023*, pp. 135–140, 2023, doi: 10.1109/TransAI60598.2023.00040.

[6] R. A. Del Del-Carpio, "Predicting academic performance using machine learning models : A systematic review of the literature," vol. 6, pp. 1038–1054, 2024.

[7] L. E. Contreras Bravo, H. J. Fuentes López, and E. Rivas Trujilllo, "Análisis del rendimiento académico mediante técnicas de aprendizaje automático con métodos de ensamble Analysis of academic performance using machine learning techniques with assembly methods," *Rev. Boletín REDIPE*, vol. 10, no. 13, pp. 171–190, 2021, [Online]. Available: https://orcid.org/0000-

[8] M. Kumar, N. Singh, J. Wadhwa, P. Singh, G. Kumar, and A. Qtaishat, "Utilizing Random Forest and XGBoost Data Mining Algorithms for Anticipating Students' Academic Performance," *Int. J. Mod. Educ. Comput. Sci.*, vol. 16, no. 2, pp. 29–44, 2024, doi: 10.5815/ijmecs.2024.02.03.

[9] R. A. Saputri and L. Rosnita, "A Random Forest-Based Predictive Model for Student Academic Performance : A Case Study in Indonesian Public High Schools," vol. 9, no. 3, pp. 1042–1049, 2025.

[10] C. Kaensar and W. Wongnin, "Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning," *Eurasia J. Math. Sci. Technol. Educ.*, vol. 19, no. 12, 2023, doi: 10.29333/ejmste/13863.
[11] F. Salas and J. Caldas, "Predicting undergraduate academic performance in a leading Peruvian university: A machine learning approach," *Educación*, vol. 33, no. 64, pp. 55–85, 2024, doi: 10.18800/educacion.202401.m003.
[12] D. Glandorf, H. R. Lee, G. A. Orona, M. Pumptow, R. Yu, and C. Fischer, "Temporal and Between-Group Variability in College Dropout Prediction," *ACM Int. Conf. Proceeding Ser.*, pp. 486–497, 2024, doi: 10.1145/3636555.3636906.
[13] A. L. Jimenez, K. Sood, and R. Mahto, "Early Detection of At-Risk Students Using Machine Learning," *Commun. Comput. Inf. Sci.*, vol. 2261 CCIS, pp. 396–406, 2025, doi: 10.1007/978-3-031-85930-4_36.
[14] J. Lee, B. Thymes, J. Zhou, T. Joachims, and R. F. Kizilcec, "Augmenting Holistic Review in University Admission using Natural Language Processing for Essays and Recommendation Letters," no. Ml, pp. 1–10, 2023, [Online]. Available: http://arxiv.org/abs/2306.17575
[15] M. S. Ibrahim Alsumaidaie, A. A. Nafea, A. A. Mukhlif, R. D. Jalal, and M. M. AL-Ani, "Intelligent System for Student Performance Prediction Using Machine Learning," *Baghdad Sci. J.*, vol. 21, no. 12, pp. 3877–3891, 2024, doi: 10.21123/bsj.2024.9643.
[16] A. F. Haro, E. R. Chisag Pallmay, J. P. Ruiz Sarzosa, and J. E. Caicedo Pozo, "Tipos y clasificación de las investigaciones," *LATAM Rev. Latinoam. Ciencias Soc. y Humanidades*, vol. 5, no. 2, pp. 956–966, 2024, doi: 10.56712/latam.v5i2.1927.
[17] O. H. González, "An approach to the different types of nonprobabilistic sampling," *Rev. Cuba. Med. Gen. Integr.*, vol. 37, no. 3, pp. 6–8, 2021.
[18] Y. Brianorman and S. Sucipto, "Prediction of Prospective New Students Using Decision Tree, Random Forest, and Naive Bayes," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 18, no. 3, pp. 1433–1446, 2024, doi: 10.30598/barekengvol18iss3pp1433-1446.
[19] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. Abou Naaj, "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance," *Trends Neurosci. Educ.*, vol. 33, p. 100214, 2023, doi: 10.1016/j.tine.2023.100214.
[20] S. E. Hooper, N. Ragland, and E. Artemiou, "Random forest models reveal academic and financial factors outweigh demographics in predicting completion of a year-round veterinary program," *J. Am. Vet. Med. Assoc.*, vol. 263, no. 2, pp. 1–9, 2025, doi: 10.2460/javma.24.08.0501.
[21] R. Baker and P. Inventado, "Educational Data Mining and Learning Analytics," in *Learning Analytics: From Research to Practice*, no. December, J. Larusson and B. White, Eds., Springer, 2014, ch. 4, pp. 61–75. doi: 10.1007/978-1-4614-3305-7.