


Predicting Undergraduate Applicants for Enrollment Using Binary Classification Machine Learning Techniques

Christian G. Guillermo

Office of Student Affairs and Services, Isabela State University, Echague, Isabela, Philippines

✉ christian.g.guillermo@isu.edu.ph

RESEARCH ARTICLE INFORMATION	ABSTRACT
<p>Received: April 3, 2025 Reviewed: April 25, 2025 Accepted: June 16, 2025 Published: June 30, 2025</p> <p> Copyright © 2025 by the Author(s). This open-access article is distributed under the Creative Commons Attribution 4.0 International License.</p>	<p>This study aimed to develop a binary classification machine learning model to predict undergraduate applications for enrollment. It used the 2024 student admissions data, such as the applicant's general weighted average, College Admission Test results, interview score, and personal information, and employed Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Machine models. The dataset was preprocessed using imputation, one-hot label encoding, standardization, and SMOTE to handle the class imbalance. The model performance was evaluated using accuracy, precision, recall, and F1 score, with Support Vector Machine emerging as the best-performing model, with an accuracy of 82%. To enhance model transparency and stakeholder trust, explainability methods under Explainable AI (XAI) were employed to interpret how and why predictions were made. These findings support the ethical use of artificial intelligence in admissions and provide a policy framework for a data-driven selection process. The model's predictive and interpretative capabilities can help the university streamline the admission process, optimize resources, and maintain fairness. Future researchers can include real-time data and broader factors to improve the adaptation and support inclusive education goals that are associated with SDGs and the Times Higher Education Impact Rankings.</p>

Keywords: *Binary classification, machine learning model, enrollment prediction, Support Vector Machine (SVM), higher education admissions*

Introduction

Higher Education Institutions (HEIs) are increasingly implementing cutting-edge technologies to streamline the admission and enrollment process for incoming first-year students. Predicting undergraduate applications for enrollment is a critical undertaking to forecast the qualified students on a precise basis for decision-making. The study of Dela Cruz et al. (2020) emphasized that prediction is essential when it comes to strategic and tactical decision-making processes that lead to effective and efficient management. According to Stemler (2012) and Mountford-Zimdars et al. (2016), admission decisions have usually relied on standardized test results, academic records, letters of reference, and personal statements. This technique frequently neglects to encompass the comprehensive nature of an individual's abilities and potential. Furthermore, the subjective nature of human judgment may inject prejudice or inconsistencies into the decision-making process.

The manual admission process for government institutions necessitates the completion of time-consuming but practical procedures (Sahagun, 2022). The implementation of Republic Act 10931, or the Free Higher Education Act, has transformed the educational landscape in the Philippines, leading to increased applications in State Universities and Colleges (SUCs). At Isabela State University (ISU), the Office of Student Affairs and Services (OSAS) frequently experiences an increase in applications that exceeds the University's capacity. For this reason, some applicants are unable to enroll due to the ranking of applicants and limited resources like classrooms and faculty. To address these challenges, universities are increasingly using data-driven methodologies, including machine learning, to better predict and manage the acceptance of incoming first-year applications.

Consequently, first-year applicants have historically been evaluated through a ranking system using a combination of the College Admission Test (CAT), academic records, and interview scores. These situations often involve the acceptance of underprivileged but well-deserving students. While the current system provides a formal framework for review, it cannot properly predict whether the applicant will enroll or not, even after being admitted to the program. This gap poses a significant challenge in the admission planning and resource allocation, particularly when the accepted applicants decline the offer. In line with Basu et al. (2019), academic institutions devote significant time and resources each year to influencing, predicting, and understanding the decision-making choices of admitted applicants. The significance of this study lies in demonstrating that machine learning methods can be applied by institutions to improve the accuracy of entering class size estimations, thereby enabling more efficient resource allocation.

By filling this gap through more precise and reliable forecasts, the application of advanced machine learning techniques for binary classification supports ISU's dedication to upholding a well-rounded and prepared population of students. The study of Patungan and Francia (2022) found that in higher education, predicting enrolment has become an essential component of institutional planning procedures. Because the budget and expenditures are based on the number of students enrolled, the yearly projection of enrolled students plays a critical role.

Additionally, machine learning is an innovative solution to the progressively intricate problem of university admission. In particular, binary classification machine learning techniques such as Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and

Support Vector Machine are effective in identifying applicants who are likely to enroll or not. Unlike other data mining tools, these models can accommodate complex, non-linear patterns and interactions among multiple predictors. The findings of Esquivel and Esquivel (2021) suggest that, given limited knowledge about prospective students, higher education institutions can use machine learning approaches to enhance management decisions. In addition, Zub et al. (2023) emphasized that one possible solution to this problem is to create a specific intellectual information system to assist in decision-making. The benefits of machine learning include its capacity to naturally expand classic statistical approaches that focus primarily on predictions and to automatically find patterns in data.

The primary feature of this study is its use of explainability approaches to interpret machine learning outputs. Explainable AI (XAI) is crucial in educational contexts where decisions impact individual lives. Johora et al. (2025) reported that the model's transparency was improved with the application of XAI techniques, which pinpointed significant features affecting predictions, including attendance, academic habits, and demographic factors. These findings can assist educators and policymakers in customizing interventions to enhance student outcomes. Likewise, Nnadi et al. (2024) emphasized that this multifaceted interpretability method bridges the gap between machine learning performance and instructional relevance by presenting a model that predicts and explains the dynamic aspects that influence student adaptability. The synthesized findings urge for educational policies that consider socioeconomic considerations, instructional time, and infrastructure stability in order to improve student adaptation. The consequences extend to informed and individualized educational interventions, which provide an adaptable learning environment. This methodological research helps to develop responsible AI applications in education by providing predictive and interpretable models for equitable and successful educational initiatives.

Thus, this study aligns with the Sustainable Development Goals, using the predictive approaches that will support an inclusive, diverse learning environment and assist the University's Times Higher Education Impact Rankings documentation. By implementing machine learning, the Office of Student Affairs and Services (OSAS) can streamline the admissions process and guarantee the admission of students who are predicted to make valuable contributions to the university and excel academically. This research study aimed to evaluate and compare the performance of binary classification machine learning techniques using metrics derived from the confusion matrix, identify the most effective model for predicting applicant status, develop a predictive model, and provide a policy framework to Isabela State University for optimizing admission processes based on the model's insights.

Methods

The researcher employed an explainability method research design that integrates machine learning (ML) predictive techniques with explainable artificial intelligence using student admission data from the university for analysis. This approach seeks not only to predict the undergraduate student who will enroll, but also to determine which attributes were most influential in the prediction. Explainability is a technique for identifying which model components or attribute combinations influenced a specific model result. According to Khare et al. (2023), explainable Artificial Intelligence (XAI) is a technique for designing AI systems that aims to provide explicit

and understandable explanations for AI model decisions. Decision-making in AI models, such as SVM, can be complex to understand. Also, Ali et al. (2023) highlighted that data explainability refers to a set of strategies for improving understanding of the datasets used to train and create AI models. In addition, the study of Siachos and Karacapilidis (2024) emphasized that explainability strategies help users comprehend the elements that influence feedback clustering and summarization, thereby enhancing system confidence.

Machine Learning Framework

The researcher employed binary classification machine learning approaches, including Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Machine, to develop prediction models. A comparative analysis was conducted to evaluate and compare the efficiency of each model generated from each data set, utilizing the performance measures of binary classification machine learning techniques through the framework of Tamascelli et al. (2020) as shown in Figure 1.

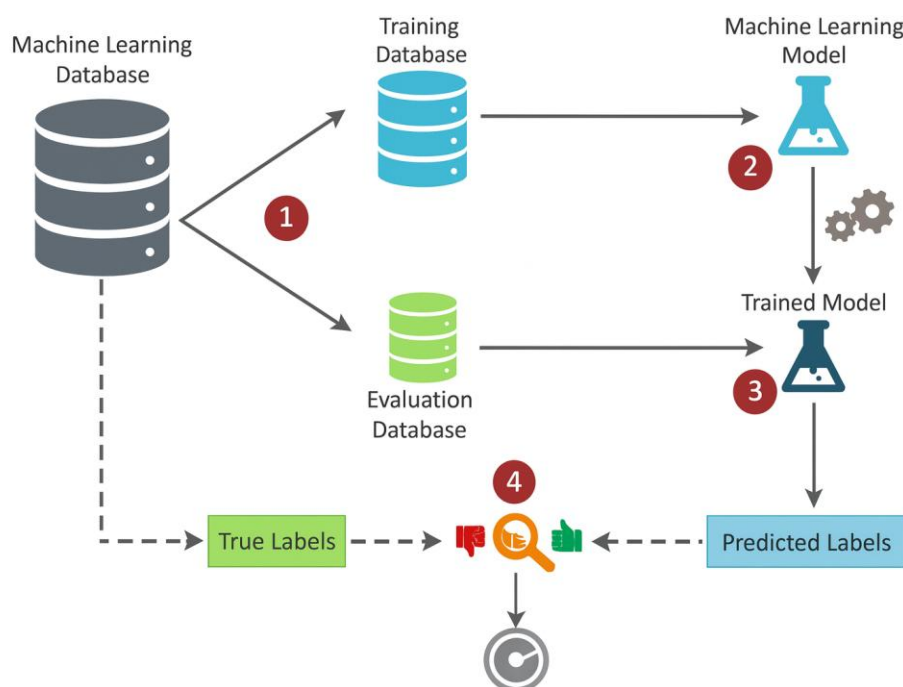


Figure 1. Machine Learning Classification Framework

Data Collection

The researcher utilized the 2024 undergraduate admissions data gathered from the Office of Student Affairs and Services (OSAS). The structured dataset comprised 8,226 college admission applications, of which 2,932 applicants were deemed eligible for enrollment in the First Semester of SY 2024–2025.

Table 1. Dataset Description

Feature Name	Description
SHS_GWA	Senior High School General Weighted Average (GWA) of the applicant
CAT_Score	Score obtained in the ISU College Admission Test
Interview_Score	Score obtained during the college admission interview
Birth_Order	Order of birth among siblings (e.g., 1st child, 2nd child, etc.)
Gender_Identity	Self-identified gender (e.g., male, female, transgender, non-binary, etc.)
Gender_Expression	How the applicant expresses their gender (e.g., masculine, feminine)
Biological_Sex	Sex assigned at birth (e.g., male, female)
Sexual_Orientation	Applicant's sexual orientation (e.g., heterosexual, homosexual, bisexual, etc.)
Monthly_Income	Total monthly family income
Disability_Status	Indicates if the applicant has a disability (Yes/No or 1/0)
SHS_Strand	Senior High School academic strand (e.g., STEM, HUMSS, ABM, TVL)
Admission_Status	Admission decision: 1 = Admitted, 0 = Rejected

Data Preprocessing and Cleaning

The researcher used the data mining tool for the retrieval and transformation of the datasets from the Office of Student Affairs and Services.

1. *Missing Values*: Numeric attributes such as SHS GWA, CAT score, and interview score were treated using mean imputation, while categorical variables were imputed using mode imputation.
2. *Encoding Categorical Variables*: The one-hot encoding was applied to nominal features such as the gender identity, gender expression, biological sex, sex orientation, and SHS strand to prevent ordinal bias, and the label encoding was used for binary categorical features such as the disability status (yes or no).
3. *Standardization*: Numeric features were standardized to ensure a uniform scale across features, especially relevant algorithms like the KNN and SVM.
4. *Bias and Limitations*: To address the class imbalance of the data set (admitted vs not admitted), the SMOTE technique was used. Overfitting was mitigated through k-fold cross-validation to ensure that the model generalized well across different subsets of the dataset.

Model Training and Evaluation

The researcher randomly split the datasets into training (70%) and testing (30%). The four binary classification machine learning techniques were trained and evaluated using the 10-fold cross-validation, which was used to minimize overfitting and validate the generalizability. According to Delgadillo and Atzil-Slonim (2022), cross-validation is a key concept in the field of machine learning. It entails utilizing certain samples to train a model and others to test its performance. This can be performed by using different sources of data for the training and evaluation stages, or by dividing a big dataset into separate subsets using split-half (70:30 train-test) sampling techniques. The confusion matrix was used to derive the performance metrics: accuracy, precision, recall, and F1-

score. In the context of admissions, recall is critical because the false negative has more severe implications than the false positive. Thus, while overall, the accuracy is vital, the models with high recall and F-1 score were prioritized, particularly the SVM.

Ethical Considerations

The researcher notified the Office of Student Affairs and Services (OSAS) of the confidentiality, privacy, and anonymity upheld throughout the study report's data collection, storage, and publication processes. The Data Privacy Act regulated and safeguarded any data produced throughout the research investigation.

Results and Discussion

The primary objective of this research was to develop a model in predicting undergraduate applications for enrollment using binary classification machine learning that will assist the University, specifically the Office of Student Affairs and Services (OSAS), in making an informed decision on the admission process.

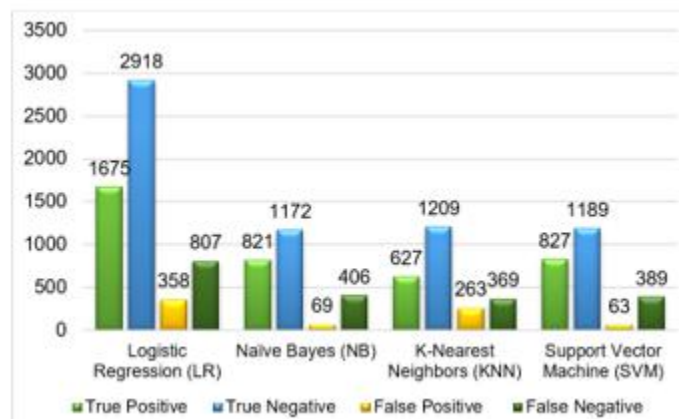


Figure 2. Performance Comparison of Binary Classification Models in Predicting the Status of Undergraduate Applicants Using the Confusion Matrix

Logistic Regression (LR)

Figure 2 presents the confusion matrix of the Linear Regression Model, which shows that the model produced many true negatives and true positives. The results indicate that it is generally effective in predicting students who will and will not enroll. However, the presence of false negatives and false positives shows that there are still cases where the model misclassifies the applicants' intentions, which could impact the overall precision and recall.

Naïve Bayes (NB)

In general, the Naïve Bayes model can reliably distinguish between individuals who want to enroll and those who do not, as evidenced by its reasonably high performance with a significant number of true positives and true negatives. However, it produced 406 false negatives, indicating that the model failed to correctly detect a large proportion of applicants who are eligible to enroll. As a result, the model's recall may be affected, suggesting that although it may make sensitive forecasts, it may not be able to identify every possible enrollee. Likewise, the presence of 69 false positives may impact

the model's precision and cause certain overestimations in the prediction of student enrollment.

K-Nearest Neighbors (KNN)

Based on the result, the performance of the K-Nearest Neighbors model could be more consistent. Including 263 false positives indicates that the model is overestimating the number of students who want to enroll, which could result in an overprediction of enrollment numbers, even though it has an acceptable number of true positives and true negatives. The 369 false negatives further suggest that the model needs a significant proportion of actual enrollees, which limits its capacity to identify all prospective students who plan to enroll precisely.

Support Vector Machine (SVM)

The SVM model's performance suggests a typically strong capacity to differentiate between applicants who will and will not enroll. The 63 false positives indicate that the model is exact, reducing the likelihood of inaccurate enrollment forecasts. The recall of the model is impacted by the 389 false negatives, which draw attention to a significant difficulty in enrolling every possible candidate. This disparity implies that even while the SVM model is conservative in its forecasts, guaranteeing that most anticipated enrollees are correct, it can overlook a sizable percentage of applicants who genuinely plan to enroll.

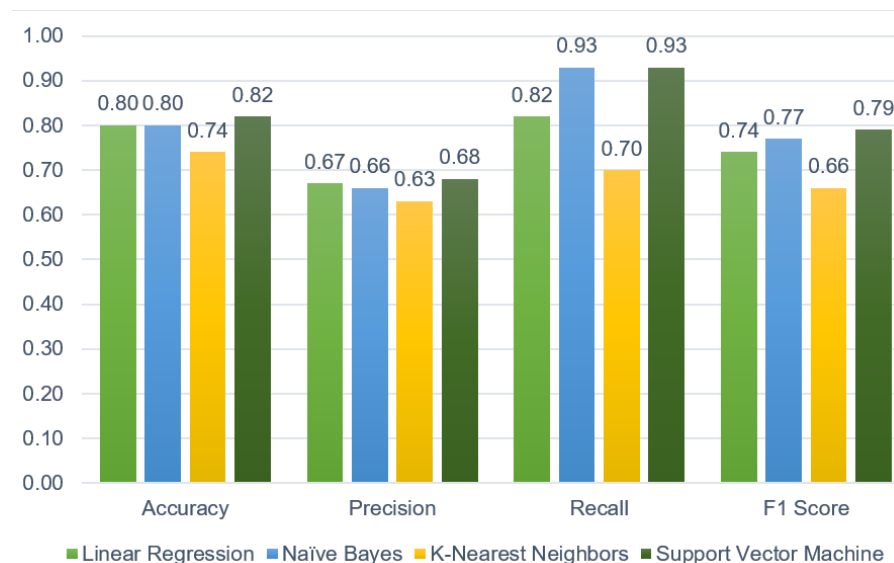


Figure 3. *Performance Evaluation of Binary Classification Machine Learning Techniques for Predicting Applicant Status Derived from the Confusion Matrix*

As shown in Figure 3, the SVM model has the highest accuracy with 0.82 or 82%, followed closely by Linear Regression and Naïve Bayes with 0.80 or 80%. On the other hand, KNN has the lowest accuracy, with 0.74 or 74%, indicating a higher rate of incorrect predictions.

In terms of precision, the Support Vector Machine (SVM) obtained the highest precision, 0.68 or 68%, which means it makes the fewest false positive predictions. The Linear Regression follows closely, with a precision of 0.67 or 67%, and Naïve Bayes is slightly lower, at 0.66 or 66%. The KNN model obtained a precision of 0.63 or 63%, the least precise among the models.

With regard to recall, the SVM and Naïve Bayes models achieve the highest recall at 0.93 or 93%. This indicates strong performance in identifying true positives. Linear regression follows with a recall of 0.82 or 82%, while the KNN model lags with a recall of 0.70 or 70%.

In terms of F1-score, the SVM model has the highest score at 0.79 or 79%, which means it balances precision and recall. This is followed by Naïve Bayes, with an F1-score of 0.77 or 77%, while linear regression has an F1-score of 0.74 or 74%. The KNN model obtained the lowest F1-score at 0.66 or 66%, which reflects its weaker overall performance.

The precision-recall trade-off is particularly significant in applications where the false positives and false negatives have different consequences. For instance, the false positive could result in the overestimation of the number of enrollments, which can result in the wasting of resources of the University, while the false negative could result in the missed opportunities in admitting competent students that were not detected by the model. In this case, the SVM model's high precision reduces the risk of overestimating enrollment, but with a recall of 93%, it may overlook up to 7% of the actual applications. This trade-off must be considered especially in allocating resources in Higher Education Institutions. The Naïve Bayes model has a high recall of 93%, making it a useful model for identifying potential enrollees even if it means overestimating enrollments (false positives). The linear regression provides a midway ground but falls short of the SVM's overall precision-recall ratio. On the other hand, the KNN model underperforms across all metrics, with a lower F1-score of 66% suggesting its weaker predictive capabilities.

Based on the performance metrics, the Support Vector Machine (SVM) model is the best-performing model for predicting undergraduate applicants' likelihood of enrollment using binary classification techniques. The SVM achieves the highest accuracy, precision, recall, and F1-scores, indicating a strong balance between correctly identifying true positives and minimizing false positives. This balance is crucial in accurately predicting enrollment numbers, ensuring that potential enrollees are identified while minimizing overestimation. The results reported by Zub et al. (2023) using a two-stage PNN-SVM ensemble model yield an accuracy of 0.940, which is the greatest value when compared to other researched approaches. The obtained results indicate that the proposed model could be employed in the following stages of developing an information system to assist HEI entrants in their decision-making process. Furthermore, the significance of this study lies in demonstrating the feasibility of employing such a model for the prediction task. This could benefit other specialists who want to build IT solutions in higher education or other fields.

These findings underscore the importance of selecting a suitable model, as it directly impacts the accuracy and efficiency of predicting undergraduate applicants for enrollment, ultimately assisting the University, specifically the Office of Student Affairs and Services, for better decisions and resource allocation.

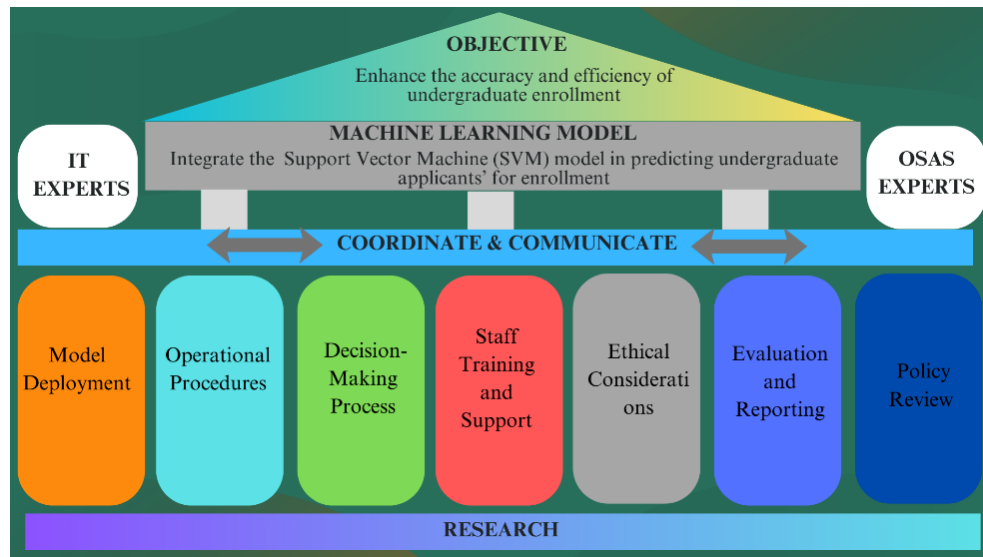


Figure 4. Policy Framework for Implementing Machine Learning Predictions for Undergraduate Enrollment

Based on the results of the study, the Support Vector Machine (SVM) model was identified as the best-performing model in predicting undergraduate applicants for enrollment. The integration of the SVM model into the university's admission process requires careful planning. The Naïve Bayes is an alternative when a high recall is necessary. There are seven (7) phases for the implementation of the model: model deployment, operational procedures, decision making process, staff training and support, ethical considerations, evaluation and reporting, and policy review. Through this framework, the University can make data-informed decisions to select qualified applicants and enhance the admission and enrollment strategies effectively.

Conclusion and Future Works

The study identified key factors in predicting undergraduate applications for enrollment using the academic records, College Admission Test, college interview score, personal information, and admission outcomes. Findings indicate that machine learning techniques can effectively create a model that supplements admission and enrollment decisions of the University. The Support Vector Machine (SVM) model is the most effective tool for predicting undergraduate applicants for enrollment, with an accuracy rate of 82%. It provides a strong balance between correctly identifying true positives and minimizing false positives. The SVM model is superior based on its performance across all key metrics, which suggests that it is well-suited for integration into the University's decision-making processes. The paper offers a policy framework for implementing machine learning predictions for undergraduate enrollment to guarantee a methodical and ethical decision-making approach that is driven by facts.

Future works should integrate supplementary variables, including parental education, geographic area (urban or rural), school type (public or private), and age group. These changes can augment precision while maintaining recall to ensure effective resource allocation and optimizing potential enrollees. In addition, future researchers should develop a system for predicting undergraduate applicant enrollment using the

SVM model. In addition, the model's ability to respond to the real-time changes of the admission process should be explored. Implementing modifications as new data, such as the volume of applications or adding demographic trends, becomes available, which can improve the accuracy of the model and responsiveness to changing enrollment patterns. This would allow the University to respond quickly to fluctuating demand and make a timely decision based on current information. Further developments in the model could include the feature of engineering, which provides more relevant and predictive variables. Hence, enhancing the models' generalization. In terms of its limitations, future researchers may take into account the potential biases of the datasets and should involve a thorough examination of the ethical implications of applying the machine learning model to university admissions. This is to ensure that the policy framework addresses issues such as fairness, transparency, and accountability.

References

- [1] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99, 101805.
<https://doi.org/10.1016/j.inffus.2023.101805>
- [2] Basu, K., Basu, T., Buckmire, R., & Lal, N. (2019). Predictive models of student college commitment decisions using machine learning. *Data*, 4(2), 65.
<https://doi.org/10.3390/data4020065>
- [3] Dela Cruz, A. P. (2020). Higher education institution (HEI) enrollment forecasting using data mining technique. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(2), 2060–2064.
<https://doi.org/10.30534/ijatcse/2020/179922020>
- [4] Delgadillo, J., & Atzil-Slonim, D. (2022). Artificial intelligence, machine learning and mental health. In *Psychotherapy* (pp. 132–142). Elsevier.
<https://doi.org/10.1016/B978-0-323-91497-0.00177-6>
- [5] Esquivel, J. A., & Esquivel, J. A. (2021). A machine learning based DSS in predicting undergraduate freshmen enrolment in a Philippine university. *International Journal of Computer Trends and Technology*, 69(5), 50–54.
<https://doi.org/10.14445/22312803/ijctt-v69i5p107>
- [6] Johora, F. T., Hasan, M. N., Rajbongshi, A., Ashrafuzzaman, M., & Akter, F. (2025). An explainable AI-based approach for predicting undergraduate students' academic performance. *Array*, 26, 100384. <https://doi.org/10.1016/j.array.2025.100384>
- [7] Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and

research recommendations. *Information Fusion*, 102, Article 102019.

<https://doi.org/10.1016/j.inffus.2023.102019>

- [8] Mountford-Zimdars, A., Moore, J., & Graham, J. (2016). Is contextualised admission the answer to the access challenge? *Perspectives: Policy and Practice in Higher Education*, 20(4), 143–150.
<https://doi.org/10.1080/13603108.2016.1203369>
- [9] Nnadi, L. C., Watanobe, Y., Rahman, M. M., & John-Otumu, A. M. (2024). Prediction of students' adaptability using explainable AI in educational machine learning models. *Applied Sciences*, 14(12), 5141.
<https://doi.org/10.3390/app14125141>
- [10] Patungan, A. J., & Francia, M. L. M. (2022). A machine learning modeling prediction of enrollment among admitted college applicants at University of Santo Tomas. *AIP Conference Proceedings*, 2560(1). <https://doi.org/10.1063/5.0100174>
- [11] Sahagun, M. A. M. (2022). Machine learning-based selection of incoming engineering freshmen in higher education institution. *International Journal of Computing and Digital Systems*, 11(1), 325–334.
<https://doi.org/10.12785/ijcds/110127>
- [12] Siachos, I., & Karacapilidis, N. (2024). Explainable artificial intelligence methods to enhance transparency and trust in digital deliberation settings. *Future Internet*, 16(7), 241. <https://doi.org/10.3390/fi16070241>
- [13] Stemler, S. E. (2012). What should university admissions tests predict? *Educational Psychologist*, 47(1), 5–17.
<https://doi.org/10.1080/00461520.2011.611444>
- [14] Tamascelli, N., Paltrinieri, N., & Cozzani, V. (2020). Predicting chattering alarms: A machine learning approach. *Computers & Chemical Engineering*, 143, 107122.
<https://doi.org/10.1016/j.compchemeng.2020.107122>
- [15] Zub, K., Zhezhnych, P., & Strauss, C. (2023). Two-stage PNN–SVM ensemble for higher education admission prediction. *Big Data and Cognitive Computing*, 7(2), 83. <https://doi.org/10.3390/bdcc7020083>

Conflict of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgements

The author expresses sincere thanks to the Almighty Father, who serves as the researcher's greatest source of strength and inspiration. Furthermore, the researcher humbly expresses his profound gratitude and appreciation to the Office of Student Affairs and Services (OSAS), which provided data for this study.

Special appreciation is extended to Dr. Albert A. Vinluan for his excellent supervision, consistent encouragement, and invaluable guidance throughout the study. His patience, motivation, enthusiasm, and vast knowledge, along with his insightful comments, greatly contributed to the improvement of this research.

The researcher also expresses deep gratitude to Dr. Helena B. Florendo, Dean of the Central Graduate School, for her unwavering support, encouraging words, and valuable suggestions. Appreciation is likewise extended to Dr. Cristine Charmaine G. San Jose, Program Chair, for her kindness, concern, and insightful contributions that helped enhance the study's clarity and presentation.

Sincere thanks are also given to the panel members—Dr. Ricardo Q. Camungao, Dr. Edward B. Panganiban, and Ms. Gloria R. Dela Cruz—for their patience, invaluable insights, and constructive suggestions, all of which were instrumental in refining this study.

Finally, the author extends heartfelt gratitude to his family, colleagues, and friends for their unwavering support and encouragement. Their contributions were essential to the success of this research.