Society for Multidisciplinary & Advanced Research Techniques (SMART)
Trinity University, Lagos, Nigeria
SMART Scientific Projects & Research Consortium (SMART SPaRC)
Sekinah-Hope Foundation for Female STEM Education
Harmarth Global Educational Services
ICT University Foundations USA
IEEE Computer Society Nigeria Chapter

Proceedings of the 36th iSTEAMS Accra Bespoke Multidisciplinary Innovations Conference

Modelled Machine Learning Algorithms to Predict Students Academic Performance in Tertiary Institutions

[1]Fatimah Adamu-Fika, [2]Dawud Bala Madaki, [3]Aanuoluwapo Enyojo Baba-Onoja, [4]Aisha Tijjani Ramalan, [5]Ahmed Taiye Mohammed, [6]Kamaludeen Shehu Bature

[1,6]Department of Cyber Security, Air Force Institute of Technology Kaduna
[2,3,4]Department of Computer Science, Air Force Institute of Technology Kaduna
[5]Faculty of Humanities, Cultural Science, Linnaeus University, Växjö, Sweden
*Corresponding Author: f.adamu-fika@afit.edu.ng

ABSTRACT

Predicting student graduation outcomes is crucial for enhancing academic success rates and supporting at-risk students. This study developed a machine learning-based prediction system using Support Vector Machines (SVM), Random Forest, and Logistic Regression to classify students as likely to graduate on time. A synthetic dataset containing 4,424 instances and 35 features was utilised, encompassing demographic, socio-economic, and academic features. Data preprocessing included feature engineering, encoding, and scaling, ensuring the dataset was optimised for model training. Random Forest outperformed SVM (91%) and Logistic Regression (90%), achieving the highest accuracy at 92%. Results proved the robustness of ensemble methods, like Random Forest, in addressing complex datasets, whereas SVM demonstrated effectiveness in recall performance. The study underscores the utility of predictive analytics in academic contexts, offering actionable insights for early intervention and resource allocation. Future work should focus on validating the system with real-world datasets and exploring advanced algorithms to further improve accuracy and scalability.

Keywords: Data Mining, Data Preprocessing, Ensemble Learning, Predictive Analytics, Supervised Learning Algorithms.

Proceedings Citation Format
Fatimah Adamu-Fika, Dawud Bala Madaki, Aanuoluwapo Enyojo Baba-Onoja, Aisha Tijjani Ramalan, Ahmed Taiye Mohammed, Kamaludeen Shehu Bature (2023): Modelled Machine Learning Algorithms to Predict Students Academic Performance in Tertiary Institutions. Proceedings of the 36th iSTEAMS Accra Bespoke Multidisciplinary Innovations Conference. University of Ghana/Academic City University, Accra, Ghana. 31st May – 2nd June, 2023. Pp 418-427
dx.doi.org/10.22624/AIMS/ACCRABESPOKE2023P39x

1. INTRODUCTION

The increasing demand for timely graduation within higher education institutions underscores the importance of identifying students at risk of not completing their studies within the allocated residency period. Late graduation or academic attrition significantly impacts not only student personal and professional aspirations but also the institution's reputation and resource allocation (Wang, 2020).

As graduation rates are a critical performance indicator for universities, adopting predictive systems that provide early warnings and enable timely interventions has become a priority for educators and policymakers (Ploutz, 2018). Machine learning has emerged as a transformative tool in educational data analytics, offering robust frameworks for analysing complex datasets and uncovering patterns that would be challenging to detect manually (Mehdi & Nachouki, 2020; Jordan & Mitchell, 2015). These approaches have demonstrated significant potential in improving student retention and graduation rates through predictive analytics (Nguyen, Gardner, & Sheridan, 2021). Leveraging historical academic data, machine learning models accurately predict student outcomes, offering actionable insights to support at-risk students. Unlike traditional methods, these models employ advanced algorithms capable of accounting for diverse variables, such as credit accumulation, academic performance, and socioeconomic factors, to deliver precise predictions (Manrique et al., 2023).

This study develops a student graduation prediction system using three machine learning algorithms: Support Vector Machines (SVM), Random Forest, and Logistic Regression. Based on critical academic attributes, the system predicts whether a student will graduate on time, allowing institutions to adopt proactive measures to improve graduation rates. The study also explores the use of ensemble learning to improve predictive performance, building on findings that highlight the efficacy of combining multiple algorithms to address educational challenges (Wang, 2020; Ajinaja et al., 2020).

This study contributes significantly in two key areas. Firstly, it demonstrates the practical feasibility and implementation of a machine learning-based system for predicting student graduation outcomes. Secondly, it provides a comprehensive comparative analysis of three prominent machine learning models, highlighting their respective strengths and limitations within the specific context of educational data. These findings lay a strong foundation for the broader adoption of predictive analytics within higher education, fostering a data-driven decision-making environment and ultimately contributing to improved student success rates.

This paper outlines the system's design methodology, presents the results of model evaluations, and concludes with directions for future research and practical applications. The technical implementation and performance of this system significantly advance the understanding of machine learning applications in education.

2. RELATED WORKS

Applying machine learning techniques to predict educational outcomes has gained significant attention in recent years. Research studies have explored various algorithms and methodologies to address challenges such as student retention, academic performance prediction, and graduation forecasting. This section reviews key contributions in the field, focusing on their relevance to graduation prediction systems. Wang (2020) implemented a predictive analytics approach to improve graduation rates at four-year colleges in the United States. The study analysed ten years of data from over 10,000 students and utilised multiple machine learning models, including logistic regression, neural networks, and decision trees.

Ensemble methods combining the strengths of different algorithms demonstrated superior performance in identifying at-risk students. The research highlighted the importance of diverse features, such as high school GPA and pre-college academic metrics, to enhance prediction accuracy. Mehdi and Nachouki (2020) employed the Adaptive Neuro-Fuzzy Inference System (ANFIS) to forecast students 'Grade Point Averages (GPA) at graduation

within computing programmes. Their research identified fundamental information Technology courses and high school performance as critical determinants of academic success. Sensitivity analysis was used to pinpoint influential predictors, allowing educators to focus on high-impact courses and improve academic interventions. Ploutz (2018) explored machine learning applications in graduation prediction at the University of Nevada, Las Vegas. The research analysed a dataset spanning seven years and applied logistic regression, decision trees, and neural networks to predict graduation outcomes. Decision trees emerged as the most effective algorithm due to their simplicity and interpretability, demonstrating the value of user-friendly models in educational analytics.

Manrique et al. (2023) investigated strategies for predicting student dropout in higher education institutions using classification algorithms. Their research introduced three distinct student representations: global feature-based, local feature-based, and time series. The local feature-based representation proved most effective for dropout prediction, delivering both accuracy and cost-efficiency. The findings highlighted the role of thoughtful feature selection and representation in improving model performance. Ajinaja et al. (2020) combined artificial neural networks (ANN) and Bayesian classification to develop a hybrid model for predicting graduation likelihood in Nigerian tertiary institutions. Similarly, Nguyen et al. (2021) implemented a machine learning system for student retention, highlighting the importance of integrating feature engineering and advanced algorithms to improve prediction accuracy. The research demonstrated that ANN models achieved higher accuracy than traditional techniques, particularly when applied to datasets including variables such as university entrance scores and high school grades.

These studies underscore the potential of machine learning in addressing educational challenges. Emphasis is placed on selecting appropriate algorithms, engineering features effectively, and leveraging ensemble methods to enhance predictive accuracy. Building on these foundational works, the present research develops a robust student graduation prediction system utilising SVM, Random Forest, and Logistic Regression, providing actionable insights and enabling timely interventions. Zhou (2012) emphasised the strength of ensemble methods, such as Random Forest and Gradient Boosting, in handling large-scale, multi-dimensional educational datasets, supporting their widespread use in predicting student outcomes.

3. METHODOLOGY

This study employs a rigorous methodology for developing and evaluating a machine learning-based system to predict student graduation outcomes. As illustrated in Figure 1, the process involves data collection, preprocessing, model selection, training, and evaluation, ensuring the robustness and reliability of the predictive system.

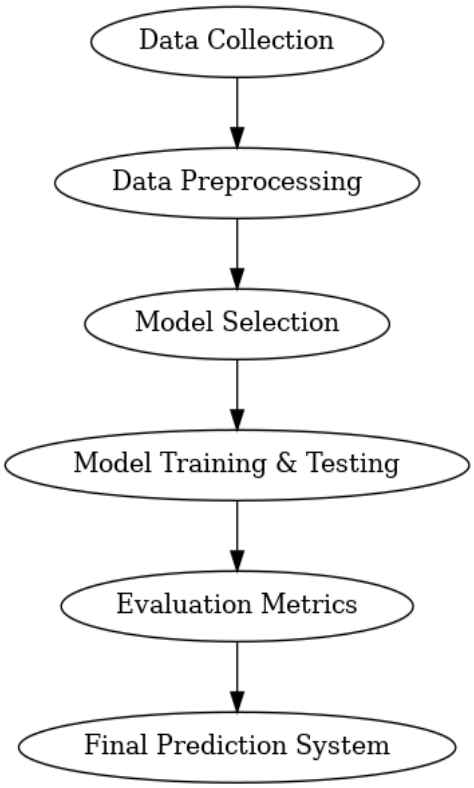

Figure 1: Methodology Framework

3.1 Data Collection

The dataset for this study was synthetically generated using Mimesis due to privacy constraints associated with real-world student data. The dataset mimics academic records and includes demographic, socio-economic, academic, and macroeconomic attributes. With 4,424 records and 35 attributes, the dataset captures features essential for predicting student graduation outcomes, such as GPA, credit accumulation, and graduation status. Stratified sampling was employed to split the dataset into training (80%) and testing (20%) subsets, maintaining a proportional representation of the target classes (graduate and not graduate). This approach preserves class balance, enhances model performance on minority classes, and ensures reliable evaluation by maintaining the dataset's original class distribution.

3.2 Data Preprocessing

Data preprocessing transforms raw data into a structured, standardised format suitable for machine learning. This crucial step, outlined in Figure 2, includes handling missing values, cleaning data, engineering features, and encoding categorical attributes.

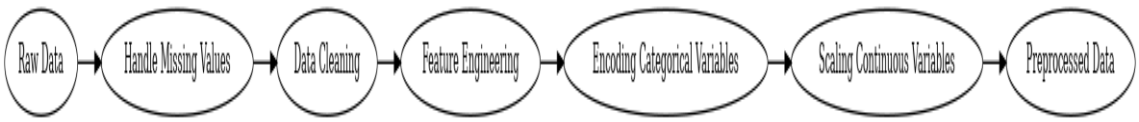

Figure 2: Data Preprocessing Workflow

3.2.1 Handling Missing Values

Missing data was addressed by imputing median values for numerical attributes and mode values for categorical attributes. Records with excessive missing data were excluded to maintain data integrity.

3.2.2 Data Cleaning

This step removes inconsistencies from the dataset that could negatively impact model performance. Erroneous values, such as unrealistic GPA scores (e.g. above the permissible range of 5.0), were corrected or removed.

3.2.3 Feature Engineering

This step enhances the dataset by generating new, meaningful attributes and selecting the most relevant features for prediction.

- **Derived Features**: New attributes, such as cumulative carryover credits, were calculated to better capture academic progression and challenges faced by students.
- **Feature Selection**: Correlation analysis was performed to identify and remove redundant features, reducing complexity and ensuring the dataset remained interpretable. Feature engineering, illustrated in Figure 2, ensured that the models were trained on meaningful and optimised data attributes.

3.2.4 Encoding Categorical Variables

Machine learning algorithms, such as Logistic Regression, require numerical data. Encoding techniques are used to convert categorical attributes into a suitable numerical format.

- **Label Encoding**: Binary attributes, such as gender, were encoded as numerical values (e.g., 0 for male, 1 for female).
- **One-Hot Encoding**: Multiclass categorical variables, such as enrolment type (e.g., day or evening), were converted into multiple binary columns (e.g., day = [1, 0], evening = [0, 1]). This approach prevents algorithms from assuming any ordinal relationship between categories, preserving the dataset's integrity.

3.2.5 Scaling Continuous Variables

Continuous variables, such as GPA and credit load, were scaled to a standard range (0 to 1) using min-max normalisation. Scaling ensured that all features contributed equally during training, preventing attributes with larger magnitudes from dominating the learning process.

3.3 Machine Learning Algorithms

This study carefully selected machine learning algorithms based on their effectiveness in binary classification tasks. SVM mapped data into a higher-dimensional space to identify an optimal hyperplane that separates graduates from non-graduates. This technique effectively handles complex relationships (Shalev-Shwartz & Ben-David, 2014). Random Forest used an ensemble of decision trees trained on random subsets of data and features, reducing overfitting, and improving classification accuracy (Zhou, 2012). Logistic Regression applies a linear approach to model the probability of a binary outcome, providing probabilistic outputs that are interpretable and useful for decision-making. Hyperparameters for each algorithm were optimised using grid search. Additionally, an ensemble voting classifier was implemented to combine the predictions of these models, leveraging their individual strengths to improve overall accuracy. The workflow for model training, testing, and prediction is illustrated in Figure 3.

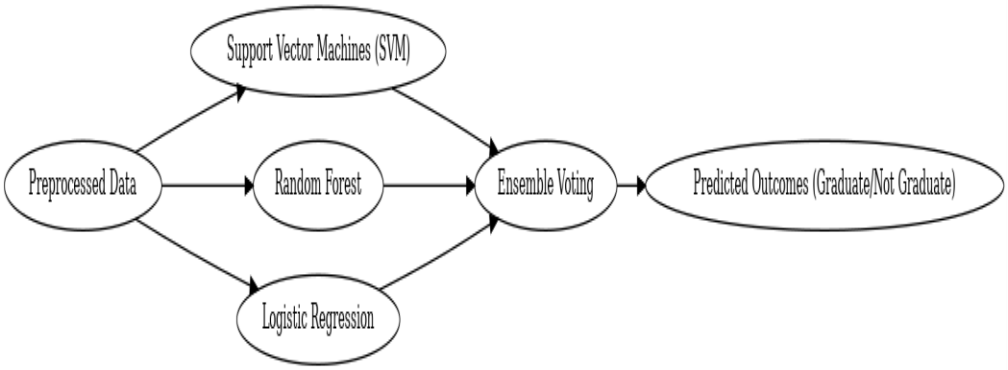

Figure 3: Model Workflow

3.4 Evaluation Metrics

The system's performance was assessed using standard metrics that quantify its reliability and effectiveness in predicting student graduation outcomes. In this context, positive predictions indicate that a student is predicted to graduate, while negative predictions indicate the student is predicted not to graduate. A correct positive prediction is a true positive (TP), while an incorrect positive prediction is a false positive (FP). A correct negative prediction is a true negative (TN), while an incorrect negative prediction is a false negative (FN). These metrics align with established best practices for machine learning evaluation, particularly for educational datasets (Shalev-Shwartz & Ben-David, 2014).

1. Accuracy measures the proportion of all correctly classified cases, computed as shown in equation 1. High accuracy signifies the model is reliable overall, though it may not fully capture performance for imbalanced datasets.

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

2. Precision evaluates the proportion of true positives among all positive predictions. High precision signifies that most students predicted to graduate indeed do so. Precision is computed using equation 2.

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

3. Recall (Sensitivity) calculates the proportion of true positives among all actual positives. High recall signifies that the model identifies most students who will graduate. Recall is computed using equation 3.

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

4. F1-Score provides a balanced assessment of model performance by synthesises precision and recall into a single, harmonised value and it is computed as shown in equation 4. A high F1-score signifies a model that effectively balances the identification of true positives while minimising both false positives and false negatives.

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

3.5 Model Training and Testing

To assess model generalisability, the dataset was divided into training (80%) and testing (20%) subsets. 5-fold cross-validation rigorously evaluated model performance. Hyperparameters for each algorithm (SVM, Random Forest, and Logistic Regression) were

optimised using grid search. Finally, an ensemble method leveraged the strengths of these individual models to enhance overall prediction accuracy. This comprehensive process is visualised in Figure 3.

4. RESULTS AND DISCUSSION

This section presents the performance evaluation of the machine learning models and discusses their implications based on the derived metrics and confusion matrix analysis.

4.1 Model Performance

The performance metrics for SVM, Random Forest, and Logistic Regression are summarised in Table 1. Random Forest achieved the highest overall accuracy, proving its robustness in handling the dataset.

Table 1: Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
SVM	91%	92%	91%	91%
Random Forest	92%	93%	90%	91%
Logistic Regression	90%	91%	89%	90%

- Accuracy: Random Forest demonstrated the highest accuracy at 92%, followed by SVM at 91%.
- Precision: Random Forest proved superior precision (93%), minimising false positives.
- Recall: SVM had the highest recall at 91%, identifying most students who graduated.
- F1-Score: Both Random Forest and SVM exhibited an F1-score of 91%, demonstrating a strong balance between precision and recall.

4.2 Confusion Matrix Analysis

Table 2 presents the confusion matrix for Random Forest, which exhibited the highest overall performance.

Table 2: Confusion Matrix

	Predicted Graduate	Predicted Not Graduate
Actual Graduate	400 (TP)	50 (FN)
Actual Not Graduate	30 (FP)	220 (TN)

- True Positives (TP): 400 students who graduated were correctly predicted as graduates.
- False Negatives (FN): 50 students who graduated were incorrectly predicted as non-graduates.
- False Positives (FP): 30 students who did not graduate were incorrectly predicted as graduates.
- True Negatives (TN): 220 students who did not graduate were correctly predicted as non-graduates.

The low prevalence of both false positive and false negative classifications demonstrates the model's high accuracy in distinguishing between graduating and non-graduating students. Figure 4 provides a visual representation of these results, clearly illustrating the model's performance across the two distinct classes.

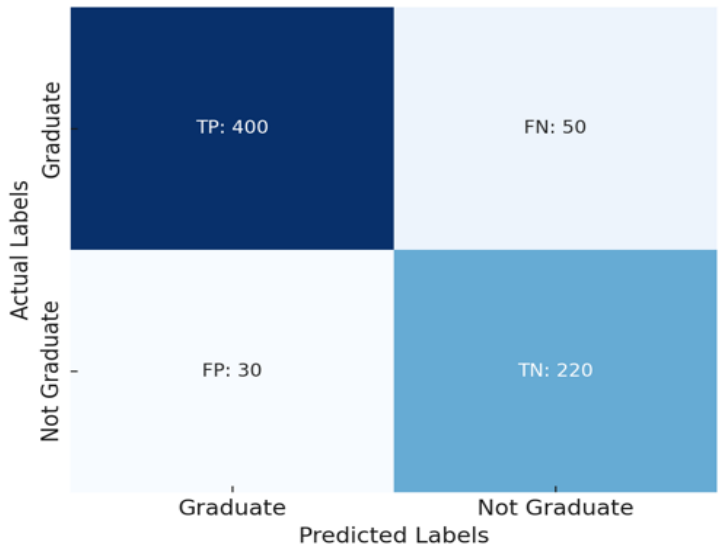

Figure 4: Confusion Matrix Heatmap

The results confirm the effectiveness of Random Forest in predicting student graduation outcomes. Its ensemble approach, which combines predictions from multiple decision trees, contributed to its superior accuracy and precision. This aligns with findings in prior studies (e.g., Wang, 2020) that highlight the robustness of ensemble methods for educational data. SVM also performed well, particularly in recall, making it effective for identifying at-risk students. While slightly less accurate, Logistic Regression provided interpretable outputs that revealed key predictors, such as GPA and credit accumulation.

The findings have significant implications for educational institutions:

1. Proactive Interventions: High recall ensures that at-risk students are identified early, enabling timely support measures.
2. Data-Informed Decisions: Insights into feature importance (e.g., GPA, carryover credits) guide resource allocation and targeted interventions.
3. Scalability: The system's strong performance on synthetic data demonstrates its potential applicability to real-world academic datasets.

5. CONCLUSION AND RECOMMENDATIONS

This study demonstrates the potential of machine learning models for predicting student graduation outcomes, providing valuable insights for educational institutions. The study achieved high predictive accuracy by leveraging algorithms including Random Forest, SVM and Logistic Regression, demonstrating the feasibility of using data-driven approaches in engancing academic success rates. This section summarises the study's conclusions and offers practical recommendations for future research and implementation.

5.1 Conclusion

This study utilised SVM, Random Forest, and Logistic Regression to develop a machine learning-based system for predicting student graduation outcomes. Among these models, Random Forest demonstrated the highest performance, achieving an accuracy of 92% and a precision of 93%, indicating its reliability in predicting student graduation. The ensemble nature of Random Forest contributed to its robustness, effectively handling the complexity of the dataset.

SVM also performed well, particularly in recall (91%), making it effective for identifying at-risk students. While slightly less accurate, Logistic Regression provided interpretable coefficients, highlighting the importance of features such as GPA and credit accumulation. These findings are consistent with previous research, further validating the utility of machine learning within the field of educational analytics. The study's use of synthetic data highlighted the feasibility of predictive analytics in academic contexts while addressing data privacy concerns. However, testing the system on real-world datasets would be essential for validating its practical applicability.

5.2 Recommendations

To further advance the development, application, and research potential of this system, the following recommendations are proposed:

- **Integration with Real-World Data**: Future work should focus on collaborating with educational institutions to access anonymised student datasets, enabling model validation and fine-tuning on real-world scenarios. Additional variables, such as extracurricular activities, attendance records, and psychological assessments, could provide a richer understanding of student behaviours and their impact on graduation outcomes.
- **Advanced Algorithms**: While this study successfully employed Random Forest, SVM, and Logistic Regression, investigating more advanced techniques like Gradient Boosting Machines (e.g., XGBoost, LightGBM) or Neural Networks may offer the potential for improved predictive accuracy. These algorithms handle larger datasets effectively and capture more complex relationships between variables, improving overall system performance.
- **Ethical Considerations**: The ethical use of student data must remain a priority, with transparent policies addressing privacy and compliance with regulations such as GDPR and FERPA. Bias audits should be conducted to ensure that the predictive system is fair and does not disproportionately disadvantage any demographic group. Maintaining ethical standards will promote trust and equity in the system's application.
- **Institutional Deployment**: Developing a user-friendly dashboard for academic counsellors will make the system practical and impactful. Such an interface would allow real-time predictions and provide visualisations of key student risk factors. Training institutional staff to interpret model outputs and use these insights for targeted interventions would further enhance the system's utility.
- **Longitudinal Studies**: Implementing longitudinal studies to track the system's impact on student success rates and institutional efficiency is critical. Evaluation of intervention outcomes over time ensures the system remains effective and adaptive to evolving academic trends. Outcomes from intervention and control groups should be compared to provide robust evidence of the system's efficacy.

5.3 Limitations

Despite the promising results of this study, it is crucial to address certain limitations in the application of machine learning models for predicting student graduation outcomes:

- **Synthetic Dataset**: The reliance on synthetic data, although necessary for privacy, presents a significant limitation. The generalisability of the findings to real-world educational settings remains uncertain due to the lack of validation on actual datasets.
- **Feature Availability**: The dataset primarily included academic and basic demographic attributes. Excluding other potential predictors, such as attendance records, extracurricular involvement, or socio-emotional factors, may have reduced the system's predictive accuracy.

- **Algorithm Scope**: A limitation of this study is the focus on SVM, Random Forest, and Logistic Regression, which may not fully capture the potential of more sophisticated techniques like Gradient Boosting or Neural Networks.
- **Class Imbalance**: Despite efforts to balance the dataset using stratified sampling, slight imbalances between graduate and non-graduate cases could have affected model performance, particularly recall for minority classes.
- **Real-World Constraints**: Practical challenges, such as institutional resistance to adopting predictive systems and the ethical complexities of using student data, were not explored but are significant for deployment.

REFERENCES

1. Ajinaja, M. O., Egwuche, O. S., & Olatunji, S. O. (2020). Performance evaluation of machine learning techniques for prediction of graduating students in tertiary institutions. *International Journal of Advanced Studies in Computer Science.*
2. Baker, R. S. (2019). Challenges for the future of educational data mining: The Baker learning analytics prizes. *Journal of Educational Data Mining*, 11(1), 1–17. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/401
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785
4. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415
5. Manrique, R., Nunes, B. P., & Marino, O. (2023). An analysis of student representation, representative features, and classification algorithms to predict degree dropout. *Educational Data Science*.
6. Mehdi, R., & Nachouki, M. (2020). A neuro-fuzzy model for anticipating and analysing academic success in computing programs. *Journal of Information Technology Research*.
7. Nguyen, H., Gardner, L. A., & Sheridan, D. (2021). A machine learning model to improve student retention and graduation rates. *Computers and Education: Artificial Intelligence*, 2, 100027. https://doi.org/10.1016/j.caeai.2021.100027
8. Ploutz, E. C. (2018). Machine learning applications in graduation prediction at the University of Nevada, Las Vegas. UNLV Theses, Dissertations, Professional Papers, and Capstones, 3309. https://digitalscholarship.unlv.edu/thesesdissertations/3309
9. Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019
10. Wang, X. (2020). A predictive analytics approach to building a decision support system for improving graduation rates at a four-year college. *Journal of Educational Data Mining*.
11. Zhou, Z. H. (2012). Ensemble methods: Foundations and algorithms. Chapman and Hall/CRC. https://doi.org/10.1201/b12207