The Current Issue and Full Text Archive of this Journal is available on
The Science Publishing House at www.scipubhouse.com

https://www.doi.org/10.56830/IJSOL062025

A Data-Driven Framework for Classifying Student Trajectories in Higher Education Using Machine Learning

Walaa H. Elashmawi [1] **Mohamed Rashad** [2] **Yasmin Alkady**[3]

[1,3] Misr International University, Faculty of Computer Science, Cairo, Egypt
[2] Pathfinder International, Cairo, Egypt
The Corresponding author: walaa.hassan@miuegypt.edu.eg

A Data-Driven Framework
doi.org/10.56830/IJSOL06202501

Received:15/03/2025
Revised: 10/05/2025
Accepted 25/06/2025

Abstract

The high rates of dropout from higher education, which range from 30% to 40% globally, pose significant challenges to institutions and societies. Conventional binary classification models (graduate versus dropout) fail to identify enrolled students at risk of academic or personal struggles, hindering proactive interventions. This study proposes a data-driven framework based on machine learning (ML) for classifying student trajectories into three distinct categories: graduate, enrolled, and dropout, providing a nuanced understanding of student progression. Leveraging a Kaggle dataset of 4424 instances with students' demographic backgrounds, academic histories, and personal context features. Three machine learning classifiers are utilized: Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The framework is composed of various phases, including data preprocessing, feature extraction of the topmost significant features, and evaluation of the utilized ML models. The RF model demonstrated superior performance, achieving 73.22% accuracy, 71.19% precision, 73.22% recall and 71.26% F1 score, with critical predictors through a feature importance analysis. This multiclass approach enables early identification of at-risk enrolled students, facilitating targeted interventions such as tailored academic advising and retention strategies. By providing interpretable data-driven insights, the framework empowers institutions to optimize resource allocation and improve student success.

Keywords: Higher Education, Machine Learning, Multiclass classification, Predictive analytics, Random Forest classifier.

1. Introduction

In recent years, higher education institutions worldwide have faced increasing pressure to address student attrition and academic underperformance, with global dropout rates ranging from 30% to 40% in regions such as the United States, Australia, and Europe (Development, 2023); (National, 2022). In the U.S., only 60% of first-time, full-time students at four-year institutions graduate within six years, while part-time students achieve a mere 25% completion rate (National Center for Education Statistics, 2022) These trends result in significant financial losses for universities and broader societal impacts, including reduced workforce readiness and increased educational inequality. Factors such as academic and social integration, financial stress, and low self-efficacy significantly contribute to these challenges (Bean, 1989); (Tinto, 1993); (Rossman & Boscardin, 2021).

To address these issues, institutions have implemented strategies like welcome programs, academic advising, tutoring, and mentoring, increasingly supported by predictive analytics. Machine learning (ML) has emerged as a transformative tool in this domain, enabling the early identification of at-risk students through sophisticated data analysis. Unlike traditional methods, ML models can process complex, multidimensional datasets encompassing academic performance, behavioral engagement, and socio-economic factors to predict student outcomes with high accuracy (Kotsiantis, Zaharakis, & Pintelas, 2007). However, many existing approaches rely on binary classification models (graduate vs. dropout), which fail to capture the dynamic nature of student trajectories, particularly for enrolled students facing academic or personal challenges (Vincent & Tinto, 2015). For instance, students with marginal GPAs due to work obligations or inconsistent attendance due to personal stressors may not be flagged in binary systems, despite needing intervention (Smith & Jones, 2020).

This study proposes a multiclass ML framework to classify student trajectories into three categories: graduate, enrolled, and dropout, using a Kaggle dataset of 4,424 instances with 35 characteristics, including demographic backgrounds, academic histories, and personal context (Realinho, Machado, Baptista, & Martins, 2021). From a machine learning point of view, this is a classification task in which the target variable might assume three different values, depending on the academic situation of the student at the end of the regular period for finishing the degree: either the student finished the degree program, is still enrolled, or has already dropped out.

The framework employs Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (kNN) classifiers to leverage complementary strengths: RF's robustness and interpretability, SVM's generalization in high-dimensional spaces, and kNN's ability to model local data structures.

By integrating ML-driven predictive analytics with institutional support mechanisms, this approach enhances early warning systems, fostering proactive, data-informed strategies to improve student retention and success. The rest of this paper is organized as follows: Section 2 reviews related work on student outcome prediction and early-warning systems. Section 3 presents utilized dataset along with the proposed model architecture and ML learning algorithms. Section 4 demonstrates the effectiveness of the proposed ML models in terms of various evaluation criteria. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Literature Review

2.1. Student Retention Challenges

High dropout rates, ranging from 30% to 40% globally, pose significant challenges for higher education institutions (Organization for Economic Co-operation and Development, 2023). In the U.S., only 60% of first-time, full-time students graduate within six years, with part-time students achieving a 25% completion rate (National Center for Education Statistics, 2022). These trends result in institutional financial losses and societal impacts, such as reduced workforce readiness and increased educational disparities. Retention is influenced by multiple factors. (Tinto, 1993) highlighted the role of academic and social integration, noting that disconnected students are more likely to withdraw. (Bean, 1989) emphasized external pressures, such as financial stress, which elevate dropout risk by 1.5 times for students working over 20 hours weekly. Psychological factors, like low self-efficacy, further exacerbate attrition (Rossman & Boscardin, 2021).

A critical gap lies in detecting enrolled but struggling students who exhibit risks, such as inconsistent engagement, but remain active. (Vincent & Tinto, 2015) noted that binary models (graduate vs. dropout) fail to identify these students, hindering timely interventions like advising. Figure 1 illustrates the dynamic pathways from enrollment to graduate, enrolled-at-risk, or dropout, emphasizing the need for multi-class classification to enable targeted support.

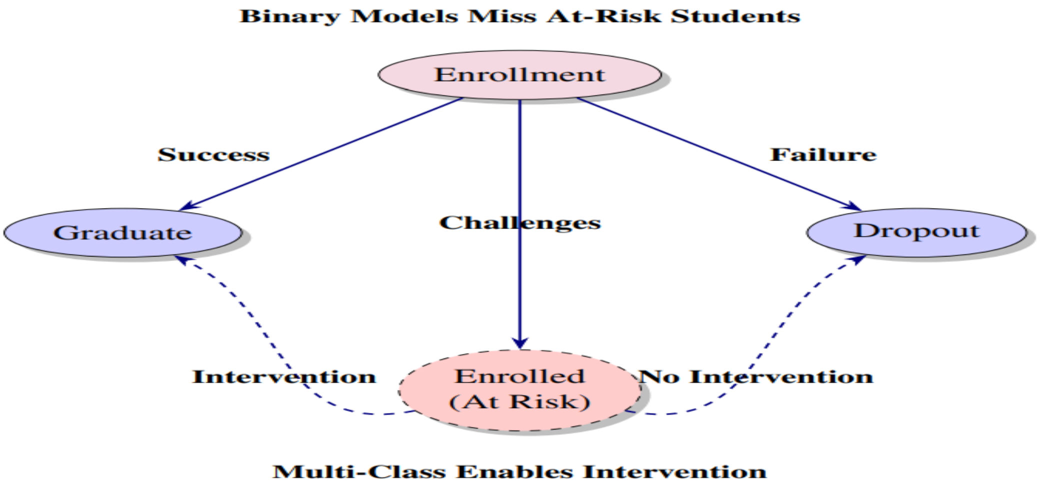

Figure 1: The student trajectories highlighting the enrolled-at-risk group. Multi-class classification enables interventions to shift students toward graduation.

2.2. Machine Learning in Education

Machine learning (ML) has transformed educational prediction by modeling complex student data. Early models, like logistic regression, used GPA and attendance to predict dropout but were limited by linear assumptions (Pascarella & Terenzini, 1980). The authors (Kotsiantis, Zaharakis, & Pintelas, 2007) showed that decision trees, capturing non-linear interactions, achieved over 80% accuracy. Recent advancements leverage temporal data. (Kim & Lee, 2024) used Gradient Boosting to predict retention with 91% accuracy, incorporating engagement patterns over semesters. However, such models often require significant computational resources, limiting scalability. However, in (Park & Kim, 2023), the authors found that weekly login frequency improved dropout prediction by 12%, highlighting the value of real-time data. Ensemble models like Random Forests offer interpretability and robustness, handling imbalanced datasets where dropouts are a minority (Gray, McGuinness, & Owende, 2018). (Kim & Lee, 2024) reported an F1-score of 0.90 using Random Forests, with course participation as a top predictor, enabling actionable insights for advisors.

Binary classification oversimplifies student trajectories by ignoring enrolled-at-risk students, who may show declining engagement without immediate withdrawal (Smith & Jones, 2020). Multi-class models address this by including an “enrolled” category.

A multi-class XGBoost model (success, at-risk, dropout) developed (Lee & Zhang, 2024). It achieves a recall of 0.89 for at-risk students using features like quiz scores and financial status. Their model supported a 13% retention increase via counseling but lacked engagement metrics. However, (Choi & Park, 2023) used a multi-class Random Forest model, achieving 85% accuracy with features like study hours and socio-economic status. Its scalability suited diverse institutions, but it omitted temporal trends. In (Kim & Lee, 2024), a Long Short-Term Memory (LSTM) model was applied to capture longitudinal engagement, achieving an F1-score of 0.88 for enrolled students. However, LSTM’s complexity and limited interpretability hinder practical deployment. Table 1 summarizes some of these comparative studies with their strengths and limitations.

Table 1: Comparative Analysis of Machine Learning Methodologies for Student Outcome Pre-Diction

Study	ML Methodology	Key Features	Performance Metrics	Strengths	Limitations	Class Type
(Pascarella & Terenzini, 1980)	Logistic Regression	GPA, attendance	Not specified	Simple, interpretable	Limited by linear assumptions	Binary
(Kotsiantis, Zaharakis, & Pintelas, 2007)	Decision Trees	Academic performance, non-linear interactions	>80% accuracy	Captures non-linear relationships	Limited scalability, noisy data sensitivity	Binary
(Kim & Lee, 2024)	Gradient Boosting, Random Forest, LSTM	Engagement patterns, course participation, quiz scores	91% accuracy (GB), 0.90 F1-score (RF), 0.88 F1-score (LSTM)	High accuracy, interpretable (RF), temporal trends (LSTM)	High computational cost, LSTM less interpretable	Multiclass
(Choi & Park, 2023)	Not specified	Weekly login frequency	12% improvement in dropout prediction	Real-time data improves prediction	Methodology not detailed	Binary
(Gray, McGuinness, & Owende, 2018)	Random Forest	Imbalanced datasets, engagement metrics	Not specified	Handles imbalanced data, interpretable	Limited performance details	Binary
(Smith & Jones, 2020)	Not specified	Declining engagement	Not applicable	Highlights need for multiclass models	No specific methodology	Multiclass

3. A Proposed Data-Driven Framework for Classifying Student Trajectory

This section introduces a comprehensive data-driven framework that utilizes multiple machine learning models to classify students into three outcome categories: graduate, enrolled, and dropout. The framework is built on a dataset comprising academic and demographic features. Three widely used classifiers are implemented: Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) to evaluate the classification task from multiple algorithmic perspectives. These models are trained and validated using a consistent pipeline, including preprocessing, feature selection, model training, and performance evaluation across key metrics: accuracy, precision, recall, and F1-score as shown in Figure 2. In addition to classification performance, the framework includes a comparative feature importance and interpretability analysis, providing actionable insights for academic advising, early intervention strategies, and data-informed institutional decision-making.

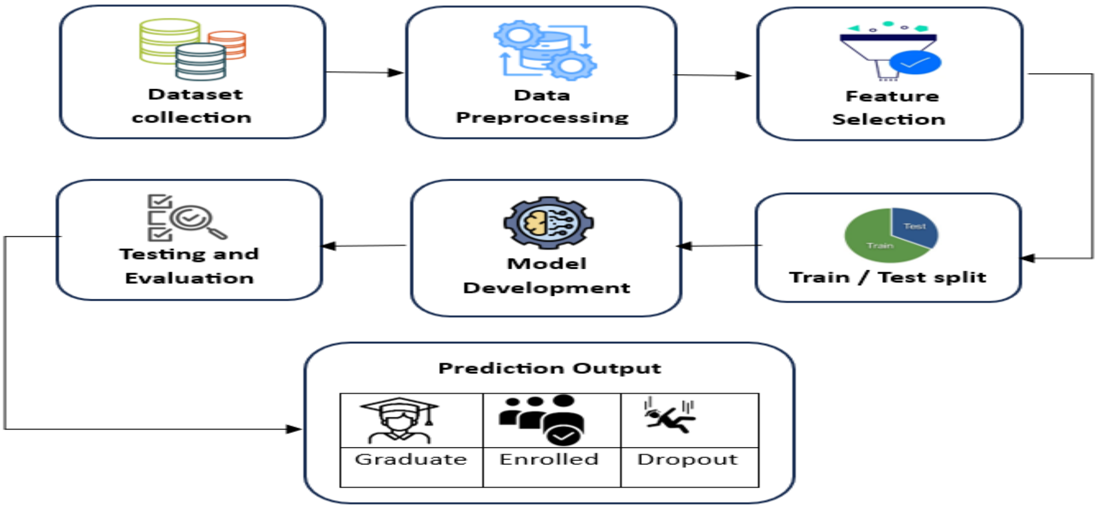

Figure 2: A pipeline of the proposed ML-based framework to classify student trajectory in higher education

3.1. Higher Education Dataset

The dataset employed in this research was obtained from an open-access educational repository on Kaggle (Devastator, 2022) by the authors (Realinho, Machado, Baptista, & Martins, 2021). It encompasses a comprehensive range of attributes reflecting students' demographic backgrounds, academic histories, and personal contexts. Key variables include, but are not limited to, marital status, application mode and order, enrolled course, attendance type (daytime or evening), and prior educational qualifications. Additional features account for socio-economic indicators such as parental education and occupation, nationality, and other relevant descriptors. The dataset is composed of 35 features and 4424 instances. The target variable classifies each student into one of three outcome categories: "Dropout," "Graduate," or "Enrolled," enabling a multi-class classification framework suitable for predictive modeling. Figure 3 shows the target class distribution across the dataset.

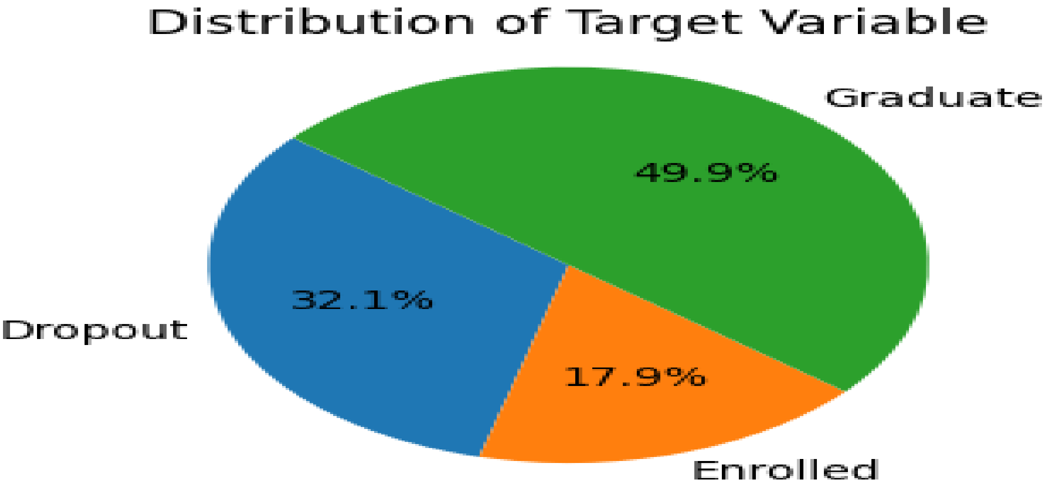

Figure 3: Target class distribution

3.2. Data Preprocessing and Feature Extraction

Data preprocessing represents a foundational phase in the development of effective machine learning models, ensuring that raw inputs are transformed into a structured format suitable for algorithmic processing. This stage encompasses several key operations, including the handling of missing values, the transformation of categorical attributes into numerical forms. In this study, categorical variables were encoded using the LabelEncoder utility from the Scikit-learn library (Pedregosa, et al., 2011), thereby converting nominal data into integer representations aligned with the input requirements of most machine learning algorithms. Following initial preprocessing, a feature extraction procedure was employed to enhance model efficiency and interpretability. A correlation-based analysis was conducted to assess the strength of association between each input feature and the target variable. Based on the computed correlation coefficients, the top ten features exhibiting the strongest positive or negative correlation with student outcomes "Target" were selected for model training (as shown in Figure 4. This focuses on the predictive capacity of the most influential variables.

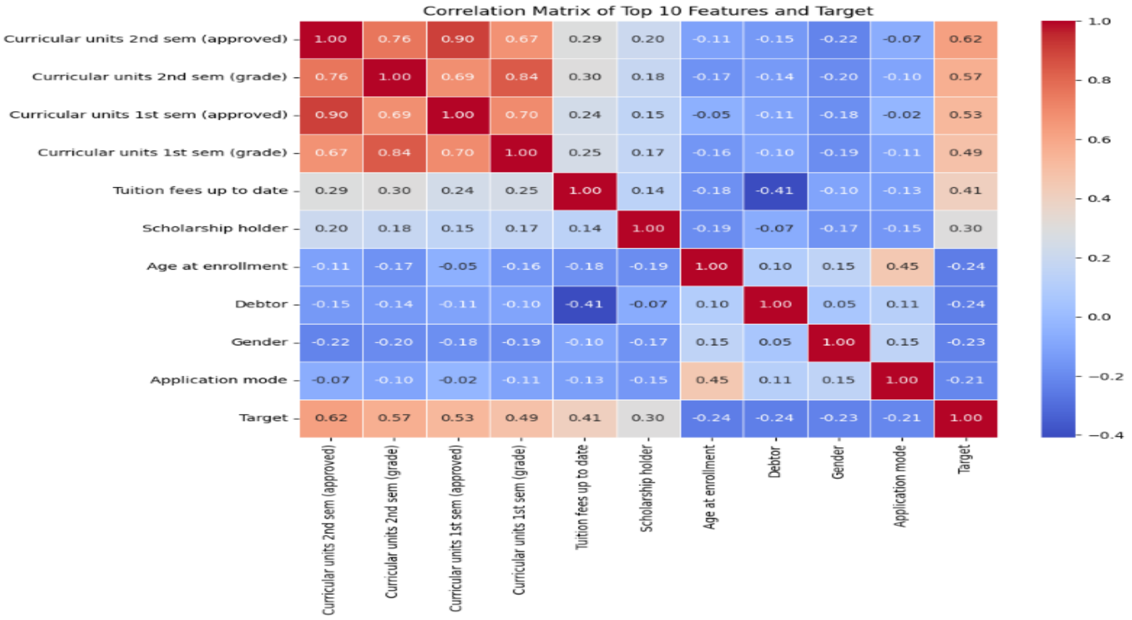

Figure 4: The correlation matrix of the top 10 features with the "Target"

3.3. Machine Learning Models

To capture a range of decision-making strategies in student outcome prediction, this study employed three distinct machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These models were chosen for their complementary strengths — Random Forest offers robustness and interpretability through ensemble learning, SVM provides strong generalization capabilities in high-dimensional spaces, and KNN serves as a non-parametric method that classifies based on proximity in the feature space. By leveraging these diverse algorithmic paradigms, the study aims to evaluate predictive performance from multiple analytical perspectives and ensure a well-rounded understanding of student trajectory classification.

3.3.1 Support Vector Machine (SVM)

Support Vector Machine is a supervised learning algorithm rooted in statistical learning theory, designed to find the optimal hyperplane that maximally separates data points from different classes (Boser, Guyon, & Vapnik, 1992). By focusing on the most informative data points, called support vectors, SVM seeks to minimize classification error while maximizing decision margin as shown in Figure 5. The algorithm can be extended to non-linear problems through the use of kernel functions, such as the Linear or radial basis function (RBF), which projects data into higher-dimensional spaces where linear separation becomes feasible. This ability to model complex decision boundaries makes SVM well-suited for educational prediction tasks involving multidimensional and heterogeneous features.

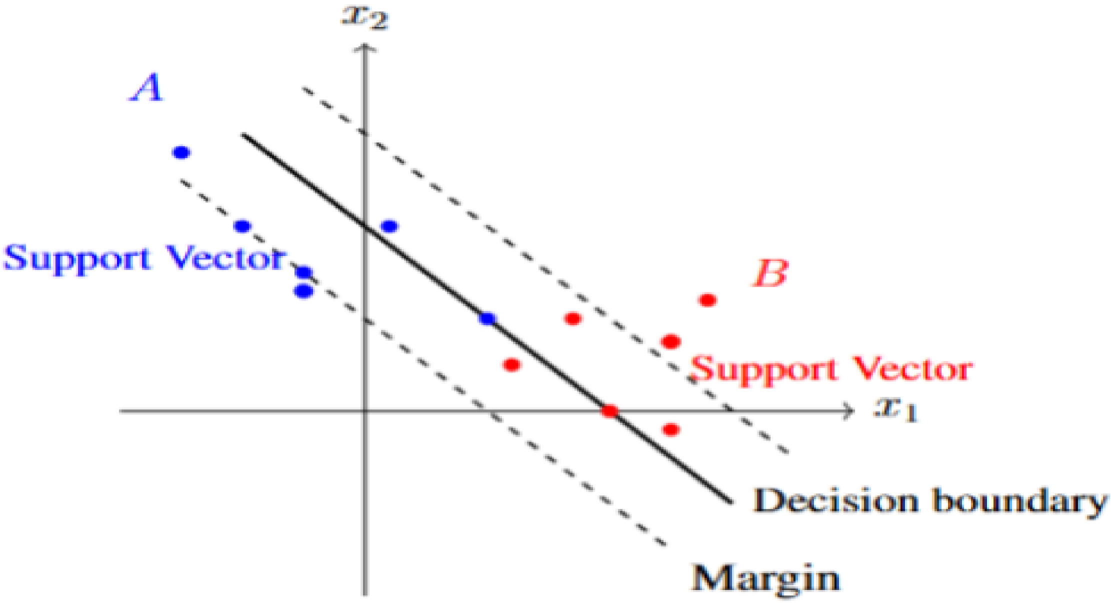

Figure 5: SVM classification with margin and support vectors.

3.3.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric, instance-based learning algorithm that classifies new data points based on the majority class among their k closest neighbors in the feature space (Cover and Hart, 1967). Figure 6 presents an illustrative view of the classification process of KNN. Unlike other algorithms that involve an explicit training phase, KNN defers learning until prediction time, making it simple yet computationally intensive for large datasets. The choice of distance metric (e.g., Euclidean, Manhattan) and the value of k play a critical role in its performance (i.e., in this study, k = 5, which is the default). Despite its simplicity, KNN can effectively model local data structures and is particularly useful in scenarios where decision boundaries are irregular or non-linear, such as in early-stage student outcome classification.

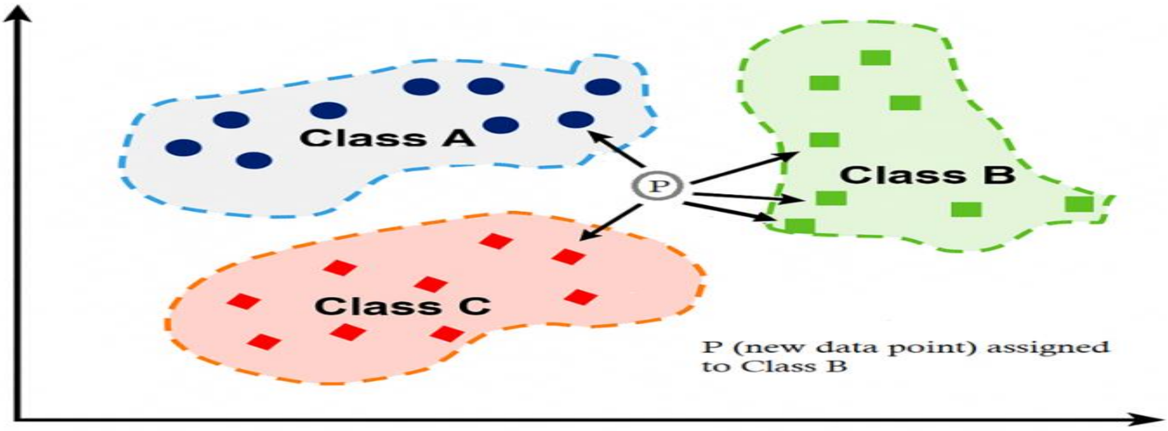

Figure 6: KNN classifier with k = 5.

3.3.3 Random Forest (RF)

The Random Forest algorithm, introduced by Leo Breiman in collaboration with Adele Cutler, represents a robust ensemble learning technique designed to enhance predictive accuracy and model stability (Breiman, 2001). It operates by constructing multiple decision trees, each trained on a bootstrapped sample of the dataset and a random subset of features at each split. The final prediction is derived through an aggregation process, commonly voting with the majority for classification tasks or averaging for regression, thus mitigating the overfitting and variance commonly associated with individual decision trees. Figure 7 depicts the general workflow of the RF classifier.

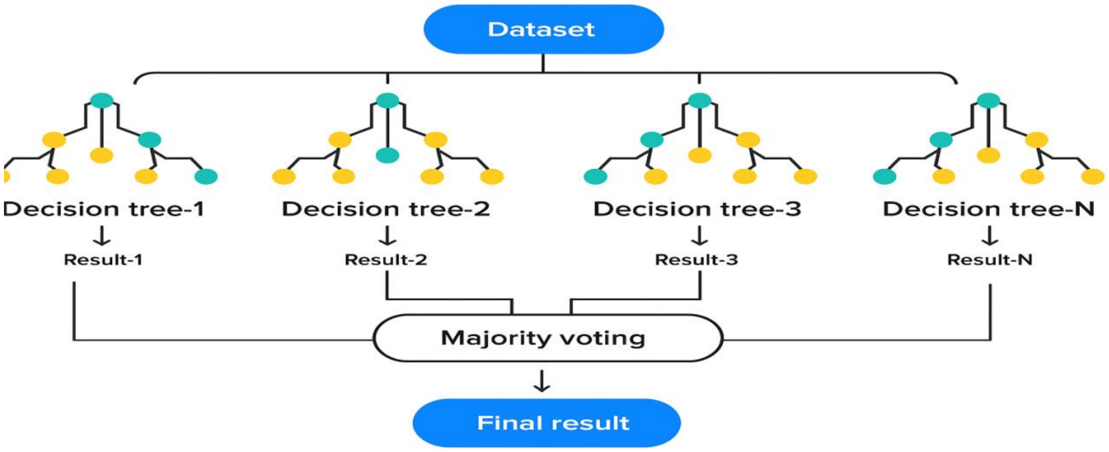

Figure 7: General Workflow of RF classifier

3.4. Training and Evaluation metrics

In this study, the data set was divided into 80% training (i.e., 3539 instances) and 20% testing (i.e., 885 instances) subsets, ensuring that the relative distribution of the three target classes (i.e., Graduate, Enrolled, and Dropout) was consistently maintained across both sets. To ensure fairness and reproducibility, all machine learning models (RF, SVM, and KNN) were trained and evaluated on identical data splits. Evaluating the performance of classification models is essential to assess their predictive accuracy. Model evaluation was conducted using a set of standard classification metrics derived from the confusion matrix terms as listed in Table 2 for multi-class classification. Consider $i \in \{\text{Graduate}, \text{Enrolled}, \text{Dropout}\}$. For each class i, the four terms of the fundamental confusion matrix $((t_i^{+}), (f_i{}^{+}), (f_i{}^{-}), and (t_i^{-}))$ can be computed to evaluate the performance of the classifier with respect to that specific class. Based on These terms, the evaluation metrics are listed below.

Table 1: Confusion Matrix Terms for Class i in Multi-Class Classification

Term	Interpretation in the Context of Student Outcome Classification
t_i^{+} (True Positive)	**The number of students correctly predicted as belonging to class i (e.g., students actually classified as $Graduate$ and predicted as $Graduate$).**
f_i^{+}) (False Positive)	**The number of students incorrectly predicted as class i, but who actually belong to a different class (e.g., students predicted as $Graduate$ but are truly $Enrolled$ or $Dropout$)**
f_i^{-} (False Negative)	**The number of students who actually belong to class i but were misclassified as another class (e.g., a true $Graduate$ predicted as $Dropout$ or $Enrolled$).**
t_i^{-} (True Negative)	**The number of students correctly predicted as not belonging to class i (e.g., a $Dropout$ or $Enrolled$ student correctly not predicted as $Graduate$)**

- Accuracy: The proportion of total correct predictions across all classes. It can be computed according to the following formula

$$Accuracy = \frac{\sum_i t_i^+}{\sum_i (t_i^+ + f_i^+ + f_i^-)}$$

- Precision: The ratio of true positives to all predicted positives for class i as in the following formula.

$$Precision = \frac{t_i^+}{t_i^+ + f_i^+}$$

- Recall: The ratio of true positives to all actual positives for class i as in the following formula.

$$Recall = \frac{t_i^+}{t_i^+ + f_i^-}$$

- F1-Score: The harmonic mean of precision and recall for class i as in the following formula.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These metrics provide a comprehensive evaluation of each classifier’s ability to distinguish between student outcomes and are reported for all three classes to enable detailed performance comparison across models.

4. Results and Discussion

This section presents the results of applying three machine learning models: Random Forest, Support Vector Machine, and K-Nearest Neighbors to predict student outcomes as Graduate, Enrolled, or Dropout. The results are discussed to highlight which model performed best and how these findings can help improve student support and academic planning. Table 3 presents a comparative summary of the performance metrics for the ML models used in this study (i.e., the weighted average is used for the metrics). The highest values for each metric are highlighted in bold to indicate the best-performing model across the evaluated criteria. These results represent weighted averages, computed by taking the metric for each class and averaging them proportionally based on the number of instances in each class. This approach is standard in multi-class classification, ensuring that class imbalances are accounted for when evaluating overall model performance.

Table 3: Performance Comparison of ML Models

ML Model	Accuracy	Precision	Recall	F1-Score
SVM	0.7107	0.7003	0.7107	0.6848
KNN	0.6927	0.6775	0.6927	0.6791
RF	**0.7322**	**0.7119**	**0.7322**	**0.7126**

The results presented in Table 3 indicate that the Random Forest classifier outperformed both SVM and KNN across all evaluation metrics. Specifically, Random Forest achieved the highest accuracy at 73.22%, exceeding SVM by approximately 2.15% and KNN by 4.55%. In terms of precision, Random Forest scored 71.19%, which is 1.16% higher than SVM and 3.44% higher than KNN. Similarly, its recall of 73.22% surpassed SVM by 2.15% and KNN by 4.95%. The F1-score, which balances precision and recall, was also highest for Random Forest at 71.26%, outperforming SVM and KNN by 2.78% and 3.35%, respectively. These results suggest that Random Forest offers more robust and consistent performance in classifying student outcomes.

Furthermore, the confusion matrix for each model is illustrated in Figures 8, 9, and 10, corresponding to the SVM, KNN, and RF classifiers, respectively.

According to Figure 8, the SVM model reveals notable performance disparities across classes. For the Graduate class, the model achieved strong results, correctly classifying 398 out of 418 instances, with only 17 misclassified as Dropout and 3 as Enrolled. In contrast, the Dropout class has 196 correct predictions out of 316, but with significant confusion: 36 were predicted as Enrolled and 84 incorrectly labeled as Graduate. The Enrolled class exhibited the weakest performance, with only 35 of 151 correctly classified instances, while 29 were misidentified as dropout and 87 as graduate.

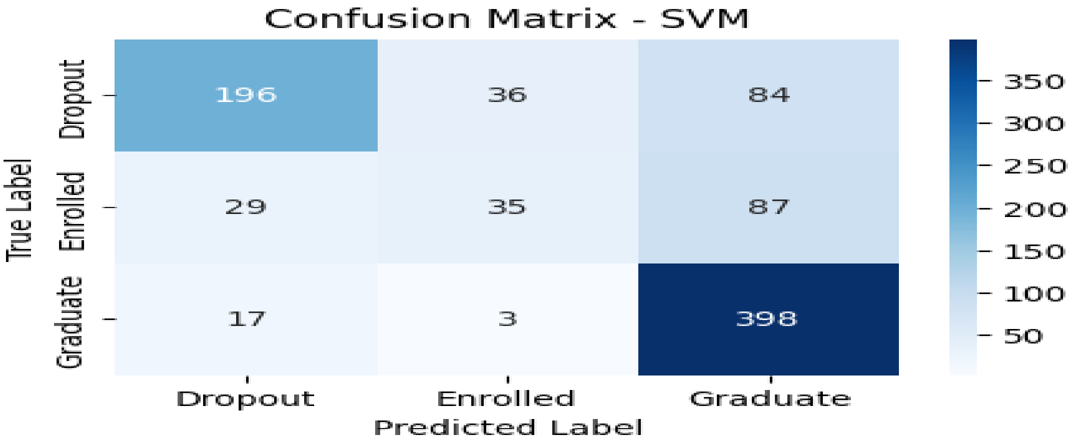

Figure 8: Confusion Matrix for SVM Classifier

The confusion matrix of KNN as shown in Figure 9 demonstrated a moderate performance with varied accuracy across classes. For the Graduate class, the model correctly identified 366 out of 418 students, while 24 were misclassified as Dropout and 28 as Enrolled. In the Dropout class, 207 of 316 instances were correctly classified, but 42 were predicted as Enrolled and 67 as Graduate, reflecting considerable overlap. The Enrolled class once again posed the greatest challenge, with only 40 out of 151 students correctly labeled. A notable 41 were misclassified as Dropout and 70 as Graduate. These results suggest that while KNN performs reasonably well for the Dropout and Graduate categories, it struggles to differentiate Enrolled students, which may be due to overlapping feature spaces and the algorithm's sensitivity to local data distribution.

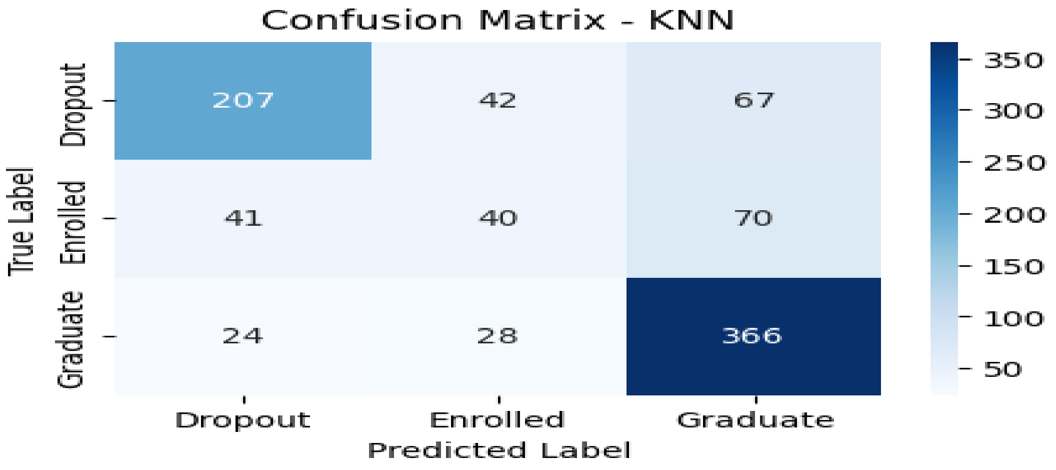

Figure 9: Confusion Matrix for KNN Classifier

However, the Random Forest model indicates strong overall classification performance with notable improvements over other models, particularly in the Dropout and Graduate classes (as shown in Figure 10. For the Graduate group, the model correctly predicted 382 out of 418 instances, misclassifying only 19 as Dropout and 17 as Enrolled. Similarly, 228 out of 316 Dropout cases were correctly classified, with 30 labeled as

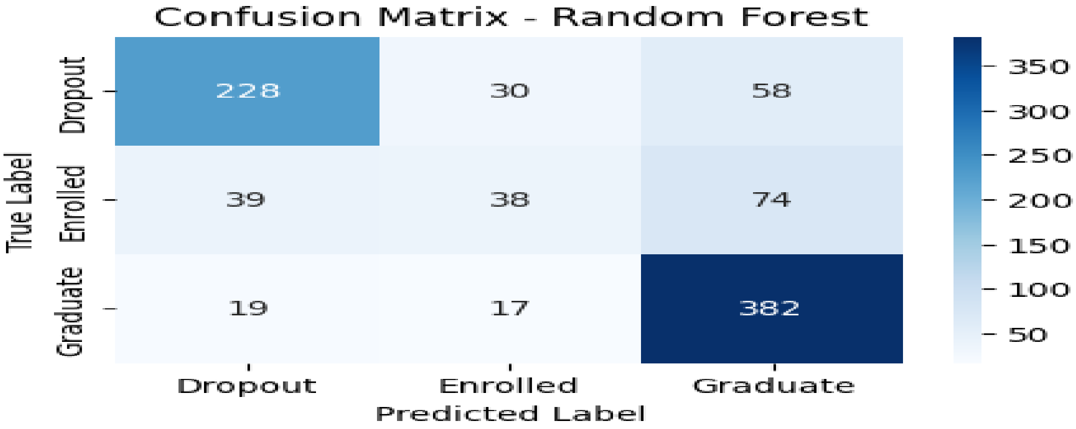

Figure 10: Confusion Matrix for RF Classifier

Enrolled and 58 as Graduate. The Enrolled class continues to be the most challenging, with just 38 out of 151 correctly identified, while 39 were misclassified as Dropout and 74 as Graduate. Despite this, the Random Forest model shows relatively balanced performance and better handling of class distinctions compared to SVM and KNN, especially in minimizing false positives and maintaining high precision for the Graduate class.

Overall, these findings highlight the importance of model selection in multi-class educational prediction tasks and confirm the robustness of ensemble learning methods like Random Forest for supporting early intervention strategies in higher education.

5. Conclusion and Future Works

This study introduced a multi-class machine learning framework to classify student trajectories in higher education, addressing the limitations of binary models that overlook enrolled-at-risk students. Utilizing a Kaggle dataset with 4,424 instances and

37 features, we applied Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) classifiers. RF achieved the highest performance, with 73.22% accuracy, 71.19% precision, 73.22% recall, and 71.26% F1-score, identifying attendance and grades as critical predictors. This framework enables institutions to detect at-risk students early, facilitating targeted interventions such as academic advising and personalized support plans, which are vital given global dropout rates of 30% to 40%. By providing interpretable feature rankings, it empowers data-driven retention strategies and promotes equitable educational outcomes. Future research should integrate longitudinal data to capture dynamic student behaviors over semesters, enhancing predictive accuracy. Exploring deep learning models, such as Long Short-Term Memory networks, could model complex temporal patterns, though interpretability remains a challenge. Validating the framework across diverse institutional datasets will ensure scalability and generalizability. Additionally, developing real-time intervention systems, leveraging engagement metrics like virtual learning interactions, could bridge predictive analytics with practical retention efforts, fostering student success across varied academic contexts.

References:

Bean, J. P. (1989). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12(2):155–187.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *In Haussler, D., editor, COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, New York, NY, USA. ACM.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Choi, M., & Park, H. (2023). Scalable multi-class random forest models for student retention in diverse institutions. *Education and Information Technologies*, 28(6):6789–6805.

Devastator, T. (2022). Higher education predictors of student reten- tion.

Development, O. f.-o. (2023). Education at a glance 2023: Oecd indicators. https://www.oecd.org/education/ education-at-a-glance/.

Gray, G., McGuinness, C., & Owende, P. (2018). An application of classification models to predict learner progression in tertiary education. *International Journal of Educational Technology in Higher Education*, 15(1):1–15.

Kim, H., & Lee, S. (2024). Leveraging temporal engagement data for student retention prediction using gradient boosting and random forests. *IEEE Transactions on Education*, 67(2):123–135.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 21(4- 5):321–344.

Lee, C., & Zhang, W. (2024). Multi-class prediction of student outcomes using xgboost: A case study in higher education. *Computers & Education*, 201:104823.

National, C. f. (2022). The condition of education 2022: Under- graduate retention and graduation rates. https://nces.ed.gov/programs/coe/ indicator/ctr.

Park, J., & Kim, Y. (2023). Enhancing dropout prediction with virtual learning environment engagement metrics. *Journal of Educational Data Mining*, 15(3):78–94.

Pascarella, E. T., & Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The Journal of Higher Education*, 51(1):60–75.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2021). Predict students' dropout and academic success (1.0) [data set]. https://doi.org/10.5281/zenodo.5777340. Accessed: 2024-06-09.

Rossman, D., & Boscardin, C. K. (2021). Psychological factors in student attrition: A review of self-efficacy and motivation. *Higher Education Research & Development*, 40(4):789–805.

Smith, A. J., & Jones, K. L. (2020). Beyond binary: The need for multi-class models in student retention prediction. *Educational Data Mining*, 12(1):45–60.

Tinto, V. (1993). Leaving college: Rethinking the causes and cures of student attrition. *The University of Chicago Press*, 2nd ed.

Vincent, T., & Tinto, V. (2015). Learning to persist: Exploring the role of early alert systems in student retention. *Journal of College Student Retention: Research, Theory & Practice*, , 17(3):346–364.