EPH-International Journal of Applied Science

Volume - 12, Issue - 01
January- 2026
DOI-https://doi.org/10.69980/tndfzk62

PREDICTING STUDENT DROPOUT AND ACADEMIC SUCCESS THROUGH EDUCATIONAL ANALYTICS

Dr Ramaswami Subramony

Professor, Departmentof Englishand Foreign Language, Guru Ghasidas Vishwavidyalaya(a Central University), Koni, Bilaspur, Chhattisgarh, Email- ramaswamysubramony@gmail.com

Received:- 11-12-2025, Revised:- 15-01-2026, Accepted:- 23-01-2026, Published:- 30-01-2026

Abstract

Students drop out is one of the major concerns in higher education as it impacts learners, institutions and educational resources. The objective of this study was to use educational analytics and supervised machine learning for predicting the academic outcomes: dropping out, continued enrolment and graduation. The following variables were analyzed: demographic, socioeconomic, admission-related, semester-level academic, and macroeconomic. Logistic Regression, Decision Tree, Random Forest, XGBoost, Support Vector Machine and K-Nearest Neighbours algorithms were used for classification and compared. Performance of the models was assessed by accuracy, balanced accuracy, precision, recall, F1-score, confusion matrices and multiclass ROC-AUC. Random Forest had the highest class balanced accuracy (70.8%) and highest macro F1 score (0.708), and the highest accuracy (75.9%). The overall accuracy of XGBoost was 76.2% with the highest macro-ROC-AUC of 0.890. The most dominant predictors were identified through explainable artificial intelligence analysis, and were approved curricular units in semester 2, course, approved units in semester 1, tuition-fee status, age at enrolment and semester grades. Both the graduations and the dropouts were better classified than the enrolled students. The results show that interpretable machine learning models can be used to assist early-warning systems and guide the implementation of specific academic, financial and counselling services and support to students identified as being at risk.

Keywords: *educational analytics; student dropout; academic success; machine learning; explainable artificial intelligence*

Introduction

Problems with student dropout in higher education are a long-standing issue that has an impact on students' academic progression, their institutions' performance, the utilization of resources, and the long-term social and economic prospects of the student. Schools are making greater use of data-informed approaches to detect learners at risk and to provide timely and appropriate support. The value of machine learning in this context is that it can analyse complex relationships that cannot be fully described using traditional statistical methods, including relationships between demographic, socioeconomic, admission-related and academic factor. Numerous students' performance prediction classification algorithms were found in the systematic review, with tree-based and ensemble algorithms being prevalent (Albreiki et al., 2021). The ability to have access to the institutional record of students has bolstered the efforts to develop predictive models for student outcomes. A research project in the Mexican higher-education sector showed the benefits of using a combination of academic and background information to predict dropout (Alvarado-Uribe et al., 2022).

An early study demonstrated that administrative and enrolment data can also be used to make accurate predictions of students who are likely to drop out of the university before graduating (Aulck et al., 2016). This work has been continued in subsequent studies by distinguishing continuing students from those who may be at risk of dropping out using supervised learning techniques (Del Bonifro et al., 2020). Institutional, economic and national context play a role in predicting dropout. From Italy, evidence shows that students' attributes are linked to their educational trajectory, as is the broader structural context in which they study (Delogu et al., 2024). As observed in the real world at universities, dropout rates can even be forecast at various stages of study, based on the available academic information (Fernández-García et al., 2021). Additionally, in an online learning context, indicators of participation and performance have been adopted to forecast outcomes of examinations and to determine early risk of dropout (Figueroa-Cañas & Sancho-Vinuesa, 2020). In recent years, more research has been conducted on the development of early-warning systems to enable universities to act before disengagement is irreversible (Goren et al., 2024).

Academic, socio-economic and equity considerations have been incorporated with machine learning and survival analysis to enhance institutional decision-making (Gutierrez-Pachas et al., 2023). Comparative tests have shown that different algorithms and data structures yield different results of the models, which means that it is necessary to test several classification methods (Kabathova & Drlik, 2021). Multiclass approaches are of great interest because the choice of student outcome is not binary, but may be a range of outcomes such as dropout, continued enrolment or successful graduation (Martins et al., 2023). An important predictor of academic risk later in life is often identified as previous educational achievement (Nagy & Molontay, 2018). First-year academic indicators also have been demonstrated to be able to offer helpful evidence of student risk for university withdrawal (Opazo et al., 2021).

In Chile, data-mining research has shown the potential of institutional data to assist in a retention-oriented approach to knowledge discovery and student support planning (Palacios et al., 2021). Furthermore, predictive studies carried out earlier in the educational trajectory also suggest that the risk of dropping out can manifest much sooner than the last year of school (Psyridou et al., 2024). Educational analytics projects on a large scale have shown that at-risk students can be identified in national school systems (Queiroga et al., 2022). Recently optimisation-based prediction systems (Rebelo Marcolino et al., 2025) have been integrated with digital learning traces (such as those of the online learning platform Moodle). There are also methodological frameworks suggested for designing, validating and evaluating dropout models in real educational contexts (Rodríguez et al., 2023).

Comparisons based on statistics and machine learning have also been conducted to explore the association between dropout, scholarship support, and institutional policy (Romero & Liao, 2025). Data-driven systems have been fairly successful in predicting academic grades and withdrawal rates (Rovira et al., 2017). The choice of student programmes can also influence dropout prediction (Segura et al., 2022). A study in Finnish higher education shows how machine learning can be used practically in identifying vulnerable students in various settings (Vaarma & Li, 2024). Supervised

classification algorithms have been found to be effective for the classification of dropout and academic success in comparative studies, provided a suitable set of predictors and evaluation measures are selected (Villar & De Andrade, 2024). It is important to note that there is a trade-off between the predictive performance of the instrument and the interpretability of its results, in order to allow institutional decision makers to understand the factors that influence the instrument outputs (Zanellati et al., 2024). The same applies to online training environments where explainable and actionable predictions are required to facilitate the retention of the learner (Zerkouk et al., 2024).

This research is focused on predicting the dropout rate, continuation in school and academic success using educational analytics. It covers indicators of demographics, socioeconomic, admission, academic and macroeconomic. The goals are to compare Logistic Regression, Decision Tree, Random Forest, XGBoost, Support Vector Machine and K-Nearest Neighbours, to assess their multiclass prediction performance, find the most performing model and describe the influence of key predictors using SHAP analysis.

2. Methodology

2.1 Research Design

The research design used in this research was quantitative predictive research with educational analytics approach. The primary goal was to build and benchmark the student classifications into three categories dropout, enrolled and graduate using machine learning models. The analytical framework included the demographic characteristics, socioeconomic indicators, as well as indicators on admission into the school and students' academic performance, for which risk factors for dropping out of school and success factors were identified.

2.2 Data Description

A secondary dataset with a structured format was used, with information for 4,424 students in higher education. There were 36 predictor variables and one multiclass variable for the students' ultimate academic status (Realinho et al., 2021). Demographic characteristics, family background, socioeconomic conditions, admission information, previous educational performance, first-semester achievement, second-semester achievement and selected macroeconomic indicators were the predictors. The dependent variable was whether the student dropped out of school, enrolled or graduated.

2.3 Study Variables

The dependent variable was the academic performance of the student, which was either dropout, enrolled or graduation. The independent variables were age at enrolment, gender, marital status, nationality, scholarship status, tuition-fee status, debtor status, previous qualification, admission grade, course type, application mode, parental education and parental occupation and semester-level academic indicators. Academic variables were: number of curricular units enrolled, approved, completed and evaluated, as well as the average grades of the first and second semester, and the number of units not evaluated in each semester.

2.4 Data Preprocessing

Data for this dataset was checked in order to identify missing data, duplicate observations, inconsistent labels and incorrect data types. There were no missing values or duplicate data records found. Names of variables were standardized and categorical variables that were coded numerically were recoded into categorical variables. The three classes (dropout, enrolled and graduate) were used to encode the target variable. Where necessary, continuous variables were transformed to have a standardisation (e.g. Logistic Regression, Support Vector Machine, K-Nearest Neighbours). One-hot encoding or similar encoding method was used to transform categorical variables. To avoid information leakage, data preprocessing has been carried out within the modelling framework.

2.5 Exploratory Data Analysis

Exploratory Data Analysis was performed to gain the insight about the distribution and characteristics of the variables. Frequencies and percentages were calculated for categorical

variables, while means, standard deviations, minimum values and maximum values were reported for continuous variables. The distribution of the target classes was examined to assess class imbalance. Continuous variables were analysed for possible correlation and multicollinearity using correlation technique. The differences in academic and socio-economic indicators between dropouts, enrolled and graduates were analysed using box plots, histograms and class-wise comparisons.

2.6 Feature Engineering and Selection

Features that were relevant in terms of educational relevance, predictive contribution and model-based importance were selected. The available curricular-unit variables can be used to derive academic indicators such as approval ratio, evaluation completion ratio and semester performance rate. Model development was preceded by the evaluation of highly correlated or redundant predictors. Feature-selection techniques such as correlation analysis, recursive feature elimination and tree-based feature importance were considered. The final feature set was chosen based on a compromise between predictive accuracy and the interpretability and avoidance of overfitting.

2.7 Data Partitioning

Stratified sampling was used to split the data into two sets of training and testing data to maintain the ratio of dropout students, enrolled students and graduate students in each subset. The ratio of 80:20 was followed, which means 80% of the observations were used for the model training and 20% for final testing. Stratified k-fold cross-validation was used on the training set to select the model and optimum hyperparameters to solve the problem. To prevent data leakage, all the preprocessing were done independently in each training fold.To avoid data leakage, all the preprocessing were done separately within each fold.

2.8 Class-Imbalance Management

A pre-modelling assessment of the distribution of the three academic outcome categories was carried out. Therefore, class imbalance was corrected by class-weight adjustment or synthetic oversampling, like SMOTE, in the enrolled group as this group was underrepresented when compared to the graduate and dropout groups. If oversampling was used, it was only applied to the training data set and not the testing data set. By using the recall, precision and F1-score of classes instead of only the overall accuracy, the effectiveness of the imbalance handling methods was assessed.

2.9 Machine Learning Models

Seven supervised machine learning algorithms were developed and compared.

2.9.1 Logistic Regression

An interpretable baseline classifier, Multinomial Logistic Regression was used. The regularisation parameters and class weights for the classification were optimised for better generalisation and imbalance of classes.

2.9.2 Decision Tree

To create the rule-based predictions and to gain insight into the hierarchical relationships between the predictors, a Decision Tree (DT) classifier was used.To generate rule-based predictions and to gain insight into the hierarchical relationships between predictors a Decision Tree (DT) classifier was used. The following parameters were tuned to avoid overfitting: tree depth, minimum samples per split and minimum samples per leaf.

2.9.3 Random Forest

Multiple decision tree ensemble learning, Random Forest, was used. The number of trees, the maximum tree depth, number of features evaluated on each split and minimum observations per leaf were optimized using cross validation.

2.9.4 XGBoost

To find complex and non-linear relationships between student characteristics and academic outcomes, XGBoost has been used. The learning rate, number of estimators, maximum depth, subsampling ratio and column-sampling ratio were tuned.

2.9.5 Support Vector Machine

Linear kernel and radial basis function (RBF) kernel were used to develop a Support Vector Machine (SVM) classifier. The regularisation parameter and kernel coefficient were optimized. Continuous predictors were normalized prior to the model training.

2.9.6 K-Nearest Neighbours

The students were classified by similarity of predictor characteristics using the KNN algorithm. Cross validation was used to select number of neighbours, distance and weighting method. Feature scaling was applied since some features have bigger magnitudes than the other.

2.9.7 Explainable Artificial Intelligence

The best model's predictions have been interpreted using SHapley Additive exPlanations. SHAP values were used to measure the impact of each predictor on the predicted academic outcome. Most influential factors were ranked using Global SHAP analysis and Local explanations were generated to explain why individual students were identified as a dropout, enrolled or graduate.

2.10 Hyperparameter Optimisation

Hyperparameter optimization was done by means of grid search, random search, or Bayesian optimization with stratified cross validation (CV). The main reason for choosing the optimisation criterion was that it is based on macro-averaged F1 score which assigns equal weight to each of the outcome classes. To prevent overfitting and enhance results on unseen observations, the complexity of the model was controlled.

2.11 Model Evaluation

The accuracy, balanced accuracy, precision, recall, specificity and F1-score were adopted as the final models' evaluation criteria. To report for multiclass imbalance, macro-averaged and weighted-averaged were reported. Confusion matrices were employed to analyse correct and incorrect classifications of all academic outcomes. Multi-class receiver operating characteristic curves and area under the curve values were computed using one-versus-rest approach. The final model was chosen using the cross-validation performance, testing-set performance, class-specific predictive ability and interpretability.

2.12 Model Comparison and Robustness Assessment

The same partitioning was used for the machine learning models for training and testing. To check the stability of the model, cross-validation scores were reported as mean ± standard deviation. To look for possible overfitting, the difference between training and testing performance was analyzed. Robustness was evaluated by comparing the results from the models with and without class-balancing methods, and different sets of features. The dropout class was in particular attention as they were not identified and may decrease the usefulness of the prediction system.

2.13 Ethical Considerations

Secondary data and anonymised student-level data was used in the study. No personal data was used for analysis. The predictive models are established for research and educational-support purposes and not for punitive and/or discriminatory decision-making. To enhance the transparency and minimise the risk of using predictions that are not explained, model interpretation was added.

3. Results

3.1 Data Quality and Outcome Distribution

There were 4,424 students and 36 predictor variables in the dataset. There were no missing values or exact duplicate observations found. The academic outcome variable was divided into three categories dropout, enrolled and graduate. The largest group were the graduates, 49.93% of the observations, and the second largest group were the dropout students, 32.12% of the observations, while the currently enrolled students represent the smallest group, 17.95% of the observations. The Table 1 above shows that there was moderate class imbalance as graduates made up 49.93% of the outcomes, 32.12% of the outcomes were dropouts and 17.95% of the outcomes were enrolled students.

Table 1. Distribution of student academic outcomes

Academic outcome	Frequency	Percentage (%)
Dropout	1,421	32.12
Enrolled	794	17.95
Graduate	2,209	49.93
Total	4,424	100.00

3.2 Descriptive Characteristics of the Students

Selected demographic, admission related, academic and macroeconomic characteristics of the students are provided in Table 2. The mean age at enrolment was found to be 23.27 years. The mean score for admission was 126.98 and the mean score for previous qualification was 132.61. The mean of the curricular units approved was 4.71 in the first semester and 4.44 in the second semester. The average grades were 10.64 and 10.23 respectively. Table 2 shows demographic, admission, academic and macroeconomic features of the students, such as age at enrolment, admission grades, students' performance in the first semester, unemployment, inflation, and GDP.

Table 2. Descriptive statistics of selected study variables

Variable	Mean	Standard deviation	Minimum	Maximum
Age at enrolment	23.27	7.59	17.00	70.00
Admission grade	126.98	14.48	95.00	190.00
Previous qualification grade	132.61	13.19	95.00	190.00
First-semester approved units	4.71	3.09	0.00	26.00
First-semester grade	10.64	4.84	0.00	18.88
Second-semester approved units	4.44	3.01	0.00	20.00
Second-semester grade	10.23	5.21	0.00	18.57
Unemployment rate	11.57	2.66	7.60	16.20
Inflation rate	1.23	1.38	−0.80	3.70
GDP	0.00	2.27	−4.06	3.51

3.3 Academic Characteristics Across Outcome Groups

There were significant differences among the three groups of students on academic outcomes. The number of approved curricular units and the highest semester grades were recorded by graduate students. However, the academic performance of the drop-out students was significantly lower in both semesters. The results in Table 3 indicate that the grades obtained by the graduate students in the semesters in which they were enrolled were the highest and the curricular units they approved were the highest, while the older students who dropped out of school obtained the lowest grades and the lowest number of accepted curricular units.

Table 3. Mean characteristics according to student outcome

Variable	Dropout	Enrolled	Graduate
Age at enrolment	26.07	22.37	21.78
Admission grade	124.96	125.53	128.79
Previous qualification grade	131.11	131.21	134.08
First-semester approved units	2.55	4.32	6.23
First-semester grade	7.26	11.13	12.64
Second-semester approved units	1.94	4.06	6.18
Second-semester grade	5.90	11.12	12.70
Unemployment rate	11.62	11.27	11.64
Inflation rate	1.28	1.21	1.20
GDP	−0.15	0.05	0.08

The mean age of the dropouts was 26.07 years and that of enrolled students was 22.37 years and 21.78 years for graduates. The average number of approved curricular units completed by the graduate students in the second semester was 6.18, while the number of approved curricular units completed by the dropouts was 1.94. The greatest differences between the groups were found for the

academic variables at the semester level, rather than for macroeconomic indicators. This implies that students' outcomes were more strongly related to academic engagement and progression than to the conditions of the labor market, inflation and GDP.

3.4 Machine Learning Model Performance

Six machine learning algorithms were supervised and tested – Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours and XGBoost. Five-fold cross validation was used to test model performance on the training set, and an independent test set consisting of 885 students. Table 4 provides a comparison of the six machine learning models and indicates that Random Forest obtained the highest values in terms of balanced accuracy and macro-F1 while XGBoost obtained the highest values in terms of overall accuracy and ROC-AUC.

Table 4. Comparative performance of the machine learning models

Model	CV macro-F1	Accuracy	Balanced accuracy	Macro-precision	Macro-recall	Macro-F1	Weighted F1	Macro ROC-AUC
Random Forest	0.712	0.759	0.708	0.713	0.708	0.708	0.762	0.886
Support Vector Machine	0.717	0.736	0.706	0.702	0.706	0.697	0.746	0.881
XGBoost	0.708	0.762	0.685	0.708	0.685	0.693	0.754	0.890
Logistic Regression	0.707	0.723	0.700	0.695	0.700	0.687	0.737	0.878
Decision Tree	0.649	0.668	0.661	0.666	0.661	0.642	0.691	0.829
K-Nearest Neighbours	0.612	0.706	0.591	0.664	0.591	0.600	0.677	0.838

Random Forest showed the best overall balance between the outcome classes with a test accuracy of 75.93%, balanced accuracy of 70.79% and a macro-F1 score of 0.708. It also attained a weighted F1 value of 0.762 and a macro ROC-AUC value of 0.886. XGBoost achieved the best overall accuracy of 76.16% and best macro ROC-AUC value of 0.890. The balanced accuracy and macro-F1 score, however, were slightly lower than those of the other models, Random Forest, suggesting comparatively weak classification of the minority enrolled class. The Support Vector Machine performed best in terms of mean cross validation macro-F1 score (0.717) but testing macro-F1 score dropped to 0.697. The single Decision Tree and K-Nearest Neighbours models were comparatively less accurate with their macro-F1 scores, while Logistic Regression performed competitively. Figure 1 comparative macro-F1 performance of the six classification models, with an emphasis on the best class balanced predictive model which is Random Forest.

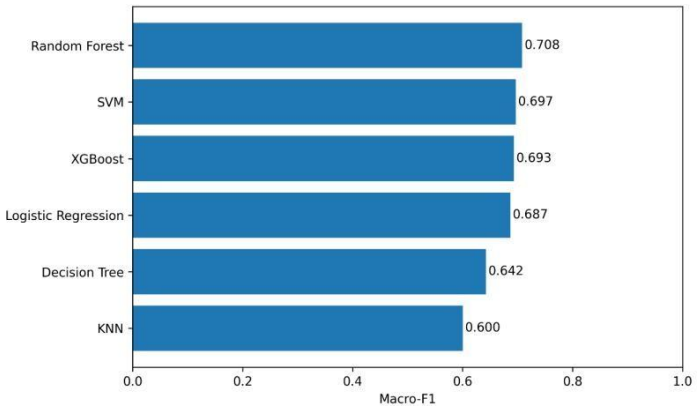

Figure 1. Comparative Performance of Machine Learning Models

3.5 Performance of the Random Forest Model

The final choice of the principal model is Random Forest due to its highest macro-F1 score and balanced accuracy on the independent testing data. The predictive performance of this model is shown in Table 5 by class.Random Forest model predicted the number of graduates and students who dropped out with greater accuracy than those who were enrolled, having the highest F1-score for the graduates class.

Table 5. Class-specific performance of the Random Forest model

Academic outcome	Precision	Recall	F1-score	Test observations
Dropout	0.824	0.708	0.761	284
Enrolled	0.481	0.547	0.512	159
Graduate	0.835	0.869	0.851	442
Macro-average	0.713	0.708	0.708	885
Weighted average	0.768	0.759	0.762	885

The model had the highest precision (0.835), recall (0.869) and F1 score (0.851) for graduate students. The dropout category had also high predictive performance results, Precision was 0.824 and F1 score was 0.761. The performance level for enrolled was relatively lower. Its precision was 0.481, recall was 0.547 and F1 score was 0.512. This may be due to the smaller sample size of this category, as well as the transitional nature of the category, since students enrolled in the program may have academic patterns similar to dropout students as well as similar to graduate students.

3.6 Confusion Matrix Analysis

The 284 dropouts in the test subset were correctly classified as dropouts. There were 49 who were classified as enrolled and 34 who were classified as graduates. Of the 159 students enrolled, 87 were correctly classified, 30 were incorrectly classified as 159 dropouts, and 42 were incorrectly classified as 159 graduates. The total number of graduate students is 442 and 384 students were properly identified. Table 6 displays the actual and predicted academic outcomes 201 of the dropout students and 87 of the enrolled students had been correctly classified, as well as 384 of the graduate students.

Table 6. Confusion matrix of the Random Forest model

Actual outcome	Predicted dropout	Predicted enrolled	Predicted graduate	Total
Dropout	201	49	34	284
Enrolled	30	87	42	159
Graduate	13	45	384	442
Total	244	181	460	885

The confusion matrix shows that the model is able to differentiate between graduate students and drop-out students better than enrolled students. The misclassifications that occurred were most often between enrolled and graduate students. The classification accuracy of the Random Forest model is illustrated in Figure 2, where a higher number of correct predictions is obtained for graduate and dropout students as compared to that for enrolled students.

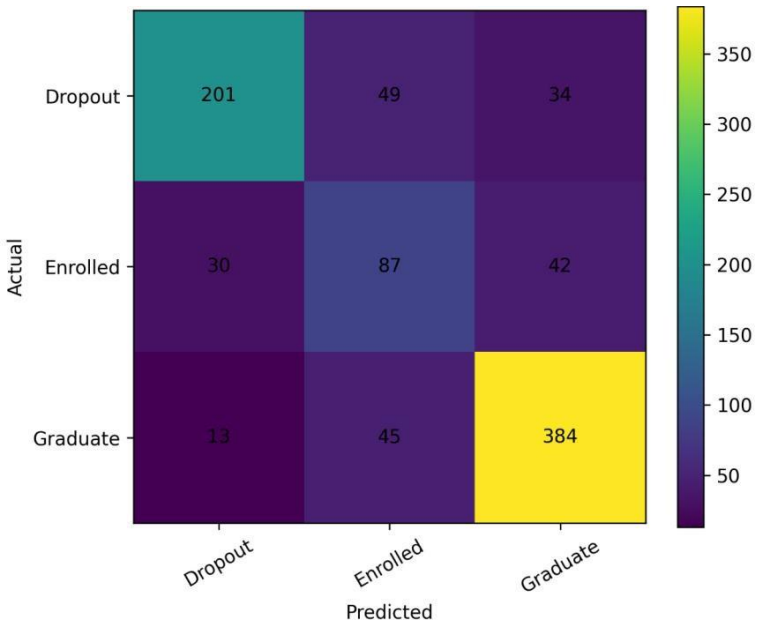

Figure 2. Confusion Matrix of the Random Forest Model

3.7 Receiver Operating Characteristic Analysis

The receiver operating characteristic analysis (ROC) method of a single category versus the rest showed good discrimination between the dropout and graduate categories. Table 7 indicates that the Random Forest model yielded high discriminatory ability for students who have graduated and who had dropped out of school, and a moderate discriminatory ability for enrolled students.

Table 7. Class-specific ROC-AUC values of the Random Forest model

Academic outcome	ROC-AUC
Dropout	0.912
Enrolled	0.819
Graduate	0.928
Macro-average	0.886

The highest ROC-AUC value was found in the graduate category (0.928), followed by dropouts (0.912). A lower, but acceptable ROC-AUC value of 0.819 was achieved in the enrolled category. The results show that the model has good discriminatory ability but it was still difficult to distinguish enrolled students. One-versus-rest ROC curves for the three academic outcomes are presented in Figure 3, which indicate that the model has a good performance in discriminating between graduate and dropout students.

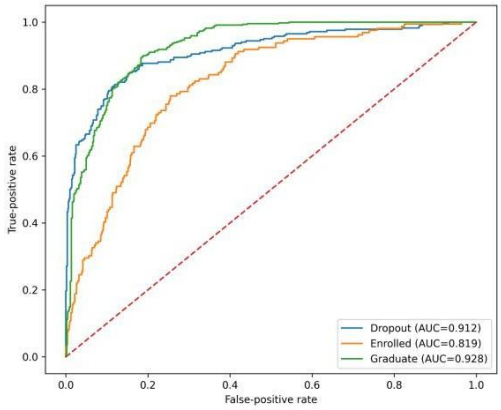

Figure 3. Multiclass Receiver Operating Characteristic Curves

3.8 Explainable Artificial Intelligence Analysis

The global contribution of predictors was computed using SHapley Additive exPlanations that were applied to the XGBoost model. Variables were ranked by their mean absolute SHAP value, indicating the overall importance of the variables to the model predictions. Table 8 lists the predictors by the mean absolute SHAP value and reveals that the second-semester approved curricular units are the most powerful predictor of the academic outcome.

Table 8. Most influential predictors identified through SHAP analysis

Rank	Predictor	Mean absolute SHAP value
1	Second-semester approved curricular units	0.623
2	Course	0.225
3	First-semester approved curricular units	0.216
4	Tuition fees up to date	0.167
5	Age at enrolment	0.128
6	Second-semester grade	0.120
7	Second-semester evaluations	0.117
8	Mother's occupation	0.106
9	First-semester evaluations	0.097
10	Second-semester enrolled units	0.088
11	Scholarship holder	0.087
12	Father's occupation	0.083
13	Admission grade	0.074
14	Application order	0.062
15	Unemployment rate	0.061

The most important predictor was the number of second semester approved curricular units, with a very large mean absolute SHAP value compared with the other predictors. Other factors also played a significant role, such as course, first-semester approved units, tuition-fee status, age at enrolment and second-semester grade. The results of the SHAP were in line with the results of the descriptive analysis, which indicated significant differences between the dropout group, the enrolled group and the graduate group regarding the curriculum units approved and the grades achieved per semester. The predictions were also also based on socio-economic and administrative factors such as tuition-fee status, scholarship status and parental occupation. The relative contribution of the predictors is shown in Figure 4, and it indicates that academic progression (semester level), course, and tuition-fee status, and age at enrolment are the most important predictors.

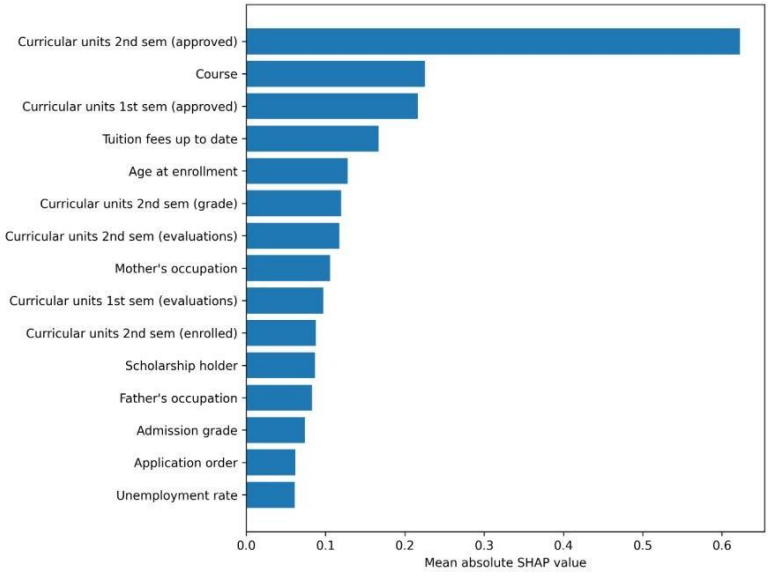

Figure 4. Global SHAP Feature Importance Ranking

3.9 Summary of the Predictive Findings

The results of the machine learning analysis indicated that the student drop-out rate and academic success can be forecasted with satisfactory accuracy with the aid of demographic, socioeconomic, admission related and semester level academic indicators. Random Forest achieved the best class-balanced accuracy and XGBoost has the highest overall accuracy and ROC-AUC value. Academic progression indicators most strongly related to the outcome were the number of approved curricular units completed in each semester of the first and second years. Tuition-fee status, course, age at enrolment and scholarship status and parental occupation also influenced student outcomes classification. Graduates and dropout students were identified with relatively high accuracy, however, prediction of the enrolled category proved to be more difficult due to the small number of students in this category and due to the fact that they had similar features with the other academic outcomes.

4. Discussion

The results show that the combination of educational indicators such as demographic, socioeconomic, admission, and semester level academic student data can predict student academic success and dropout with acceptable accuracy. Random Forest had the highest class-balanced accuracy (70.8%), accuracy (75.9%) and macro-F1 score (0.708). The overall accuracy of XGBoost was slightly better at 76.2%, and the macro-ROC-AUC was the highest at 0.890, but the macro balanced accuracy score showed relatively lower overall performance on the three outcome classes. The findings are consistent with existing work which considers that overall accuracy should not be used as the sole criterion in model selection, especially in the case of poorly balanced student outcome categories (Albreiki et al., 2021). The effectiveness of Random Forest is in line with research that demonstrates that ensemble methods are able to model nonlinear relationships and interactions between student characteristics (Villar & De Andrade, 2024). Also, similar comparative investigations suggest that there is no single algorithm that performs best on all institutional datasets and that multiple measures should be used to evaluate the algorithms (Kabathova & Drlik, 2021). The ability of XGBoost to perform competently also explains the increased popularity of boosting models in educational predictions, especially when incorporating academic and socioeconomic characteristics. (Zanellati et al., 2024).

The findings revealed that the graduates and dropouts were correctly identified, in comparison to the currently enrolled students. The enrolled group had an F1 score of 0.512, whereas the dropout and graduate groups had scores of 0.761 and 0.851, respectively. This may be a result of the transitioning nature of enrolled students, with academic behaviour that can be similar to that of future dropouts or eventual graduates. Studies of multiclass phased prediction have also highlighted that intermediate learning states are challenging to differentiate from final learning outcomes (Martins et al., 2023). Additionally, the fewer observations enrolled may have led to a decrease in the ability of the model to learn stable class specific patterns. Second-semester approved curricular units proved to be the most important factor, followed by course, first-semester approved units, tuition-fee status, age at enrolment, and second-semester grade. These results suggest that characteristics of admission are not sufficient to predict academic outcomes after enrolment, and that academic progression after enrolment offers more predictive power. First-year dropout studies have also emphasised the role of early academic success and credits gained (Opazo et al., 2021). Research on the use of academic records across stages of the university has also yielded results showing that the more advanced one progresses the higher the accuracy of the prediction (Fernández-García et al., 2021).

There is also consistency with evidence that previous and current academic performance are key measures of educational persistence (Nagy & Molontay, 2018) in terms of units and semester grades. Prediction systems using data, like those discussed above, have also been correlated with grade completion, academic engagement, and achievement and dropout (Rovira et al., 2017). Similarly, in online courses, early prediction has shown that performance indicators can identify the risk of students withdrawing before the end of a programme (Figueroa-Cañas & Sancho-Vinuesa, 2020). Other factors also predicted, including tuition-fee status, scholarship status and parental occupation, as well as age at enrolment, which suggests that academic results are not just the consequence of

grades. The inclusion of socioeconomic and equity factors in studies has found that financial and social factors may impact persistence in higher education (Gutierrez-Pachas et al., 2023). There is also research on the impact of financial aid on college dropout that lends support to the relevance of scholarship related variables (Romero & Liao, 2025). Therefore, broader institutional and economic circumstances can influence the likelihood of the retention of academically at-risk students.

The impact of course and application related characteristics indicates that programme "fit" and educational preference could have an impact on persistence. Past studies have revealed that programme preferences are significant in the context of drop-out classification (Segura et al., 2022). Furthermore, across the universities, the importance of predictors and the performance of the model can change (Palacios et al., 2021). Hence, the present model needs to be validated from outside the college environment prior to implementation in other educational context. The results confirm that predictive modelling can serve as early warning systems and that it is not suitable to be used as an automated decision-making system. Other frameworks have suggested incorporating outputs of machine learning with institutional review, counselling, and targeted academic support (Rodríguez et al., 2023). Throughout the country and at the large scale, school-based educational analytics have proven to be useful for identifying students at risk and direct the planning of interventions (Queiroga et al., 2022). But clear explanations are still crucial as false forecasts can also have negative consequences for students or perpetuate inequities.

The use of SHAP makes the analysis more useful, as it reveals why specific results were forecasted. This is consistent with the recommendation to find a balance between predictive accuracy and interpretability when modeling student dropout. Explainable outputs can inform institutions about students who may have financial or demographic vulnerabilities compared to students who are academically struggling. However, secondary, cross-sectional institutional data was used for analysis, and behavioural data (e.g., attendance, learning-platform usage, motivation, psychological well-being, etc.) was not included. The results of recent studies based on the Moodle logs show that digital engagement variables could be further added to enhance the prediction. Thus, future studies should integrate administrative, academic, behavioural, and longitudinal data to create more adaptive and generalizable systems of support for students.

5. Conclusion

The results of this study showed that the educational analytics and machine learning are effective in predicting student drop out, persistent enrollment, and academic achievement. Random Forest model showed the best class balanced performance results while XGBoost model showed the best overall accuracy and ROC-AUC result among the evaluated models. The results of this study show that ensemble learning techniques can be especially well suited to the identification of complex relationships among demographic, socioeconomic, admission related and academic variables. The most important source of predictive information was semester-level academic progression. Approved curricular units, semester grade, course, tuition-fee status, age at enrolment, scholarship status, and parental occupation were key factors in the categorisation of student outcomes. The classification of the dropout and the graduate students was more successful than the classification of students enrolled in the program, suggesting that the difficulty in classification of states of transition between enrolled and drop out and between dropout and graduate remained high. The findings provide evidence for the need to develop Early Warning Systems (EWS) that aid institutions in identifying students who need academic, financial, and/or counselling assistance. By describing the contribution of individual predictors, the use of SHAP also enhanced the transparency of the model. The findings, however, should be interpreted having taken into consideration the use of secondary and cross sectional data. The findings from this study need to be consolidated in future studies by taking into account attendance data, digital learning behavior, psychological factors, motivation, and longitudinal information for better generalisability and prediction. The study offers a practical and meaningful data-informed framework for supporting student retention strategies in general.

References

1. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. Education Sciences, 11(9),

552.
2. Alvarado-Uribe, J., Mejía-Almada, P., Masetto Herrera, A. L., Molontay, R., Hilliger, I., Hegde, V., ... & Ceballos, H. G. (2022). Student dataset from Tecnologico de Monterrey in Mexico to predict dropout in higher education. Data, 7(9), 119.
3. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.
4. Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020, June). Student dropout prediction. In International conference on artificial intelligence in education (pp. 129-140). Cham: Springer International Publishing.
5. Delogu, M., Lagravinese, R., Paolini, D., & Resce, G. (2024). Predicting dropout from higher education: Evidence from Italy. Economic Modelling, 130, 106583.
6. Fernández-García, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sánchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. IEEE Access, 9, 133076-133090.
7. Figueroa-Cañas, J., & Sancho-Vinuesa, T. (2020). Early prediction of dropout and final exam performance in an online statistics course. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 15(2), 86-94.
8. Goren, O., Cohen, L., & Rubinstein, A. (2024). Early Prediction of Student Dropout in Higher Education Using Machine Learning Models. International Educational Data Mining Society.
9. Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G., & Gomez-Nieto, E. (2023). Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the Latin American context. Education sciences, 13(2), 154.
10. Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. Applied Sciences, 11(7), 3130.
11. Martins, M. V., Baptista, L., Machado, J., & Realinho, V. (2023). Multi-class phased prediction of academic performance and dropout in higher education. Applied Sciences, 13(8), 4702.
12. Nagy, M., & Molontay, R. (2018, June). Predicting dropout in higher education based on secondary school performance. In 2018 IEEE 22nd international conference on intelligent engineering systems (INES) (pp. 000389-000394). IEEE.
13. Opazo, D., Moreno, S., Álvarez-Miranda, E., & Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. Mathematics, 9(20), 2599.
14. Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. Entropy, 23(4), 485.
15. Psyridou, M., Prezja, F., Torppa, M., Lerkkanen, M. K., Poikkeus, A. M., & Vasalampi, K. (2024). Machine learning predicts upper secondary education dropout as early as the end of primary school. Scientific Reports, 14(1), 12956.
16. Queiroga, E. M., Batista Machado, M. F., Paragarino, V. R., Primo, T. T., & Cechinel, C. (2022). Early prediction of at-risk students in secondary education: A countrywide K-12 learning analytics initiative in Uruguay. Information, 13(9), 401.
17. Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict Students' Dropout and Academic Success [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.
18. Rebelo Marcolino, M., Reis Porto, T., Thompsen Primo, T., Targino, R., Ramos, V., Marques Queiroga, E., ... & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: insights from moodle log data. Scientific Reports, 15(1), 9840.
19. Rodríguez, P., Villanueva, A., Dombrovskaia, L., & Valenzuela, J. P. (2023). A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. Education and Information Technologies, 28(8), 10103-10149.
20. Romero, S., & Liao, X. (2025). Statistical and machine learning models for predicting university dropout and scholarship impact. PloS one, 20(6), e0325047.
21. Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and

dropout. PLoS one, 12(2), e0171207.22.Segura, M., Mello, J., & Hernández, A. (2022). Machine learning prediction of university student dropout: Does preference play a key role?. Mathematics, 10(18), 3359.

22. Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. Technology in Society, 76, 102474.
23. Villar, A., & De Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. Discover Artificial Intelligence, 4(1), 2.
24. Zanellati, A., Zingaro, S. P., & Gabbrielli, M. (2024). Balancing performance and explainability in academic dropout prediction. IEEE Transactions on Learning Technologies, 17, 2086-2099.
25. Zerkouk, M., Mihoubi, M., Chikhaoui, B., & Wang, S. (2024). A machine learning based model for student's dropout prediction in online training. Education and Information Technologies, 29(12), 15793-15812.