Predicting Students' Attrition for the Information Systems Program Using Machine Learning Algorithms

Elfadil A. Mohamed [1,*], Mirna Nachouki [1], Riyadh Mehdi [1], and Yara Mohammad [2]

[1] Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman, United Arab Emirates
[2] Department of Accounting & Information Systems, School of Business Administration, American University of Sharjah, Sharjah, United Arab Emirates
Email: elfadil.abdalla@ajman.ac.ae (E.A.M.); mirna@ajman.ac.ae (M.N.); r.mehdi@ajman.ac.ae (R.M.); ymohammad@aus.edu (Y.M.)
*Corresponding author

***Abstract*—Student attrition, a crucial factor in determining institutions' financial resources, is also a measure of an institution's academic standing, with a direct impact on its revenue. This study, which examined the factors influencing students' departure from the Information Systems (IS) program, has significant implications for higher education. The research was conducted in 2 stages. Initially, we experimented with 13 machine learning algorithms to identify the most accurate method for predicting student attrition. Our findings revealed that Extra Trees, Light Gradient Boosting Machine, and Random Forests are the top 3 predictors in that order. In the second stage, we used these algorithms to identify the most influential factors driving student attrition. Our results indicated that the 3 classifiers identified the Last Grade Point Average (Last GPA) as the most significant factor affecting students' decision to leave the IS program. Performance in mathematics and gender are the other 2 factors affecting dropout. We have found that, as a percentage, more male than female students are leaving the program. Moreover, female students are primarily transferring to more academically demanding majors, such as data analytics and information technology, while male students are transferring to less technology-oriented majors, such as management and law. This finding is reinforced by the observation that female students' mathematics performance among students leaving the IS program is significantly higher than that of male students. Our results can have significant policy implications for admission procedures and academic advising, positively affecting the IS program's attrition rate.**

***Keywords*—machine learning, random forests, student attrition, extra trees**

I. Introduction

Student dropout has long been a critical issue in higher education institutions, with low retention rates impacting students' academic and financial plans [1]. Student dropout risk and its early prediction have become among the most focused areas for educational institutions in recent years, and researchers worldwide are paying significant attention to them [2].

Therefore, addressing these critical issues through science-based strategies and plans is essential. Additionally, it is important for educators and policymakers seeking to address this problem to understand the history of student retention and the primary reasons students drop out of their courses. Over the past 4 decades, educators and researchers have examined various theoretical models and empirical studies on this topic in the literature [1]. Their findings provide practical insights for improving student retention rates.

Aljohani [1] pointed 3 major conceptual models influenced early research on student retention, Durkheim's [3] theory of sociological suicide and the social anthropological theory of rites of passage in tribal societies [4]. However, later studies interpreted student retention from various theoretical perspectives, including physiological, psychological, sociological, cultural, organizational, environmental, interactional, and economic views [1].

Durkheim's [3] famous work on suicide, shaped most of the models developed after 1970 regarding student retention. Durkheim [3] believed that a person's inability to integrate socially and intellectually into society was the reason behind suicide. According to early models of student retention, suicidal behavior and student attrition are linked [5–7]. Tinto [8] notes that there are similarities between the process of suicide and dropping out of college, even though they are not the same as failure. Both actions can be seen as intentional withdrawals from a particular society.

Bean [9] argued that student attrition can be compared to employee attrition, and both leave for similar reasons. In universities and organizations, student and employer satisfaction play a vital role in predicting retention. For

Manuscript received December 1, 2025; revised February 11, 2026; accepted March 20, 2026; published June 22, 2026.

example, Bean [9] asserted that whereas the "pay" factor is one of the most significant indicators of employee turnover in workplaces, the educational system's corresponding predictors include student Grade Point Average (GPA), development, institutional quality, and practical value.

The literature categorizes theoretical models and studies on student retention into various classifications based on the perspectives considered, such as psychological, sociological, organizational, environmental, interactional, and economic [8, 10–12]. Psychological and sociological factors mainly dominate the field and are often the general concepts to which most student retention models belong [1]. Organizational factors also influence student retention in higher education institutions in various ways, including faculty numbers, infrastructure, administrative systems, and resources [8]. These factors have been examined by Bean [9, 13, 14]. Additionally, economic factors play a crucial role in students' assessments of the costs of their study programs relative to the potential benefits [8, 12]. Habley *et al.* [12] state that these costs include the indirect costs of "the time and energy" students dedicate to external and college commitments. Examples of this factor are reported by Manski *et al.* [15] and John [16]. The main theoretical models that underpin many recent empirical studies are: Undergraduate Dropout Process Model [5, 6], Institutional Departure Model [7, 8], Student Attrition Model [9, 13], Student-Faculty Informal Contact Model [17], Non-traditional Student Attrition Model [18], and Student Retention Integrated Model [19].

Aljohani [1] concluded that the main limitation of student retention studies, regardless of the model used, is their generalizability, as their findings are specific to the program and student population studied. The Information Systems (IS) program at our institution has a relatively high attrition rate compared with other programs. Therefore, to examine how individual students' attributes affect their decision to leave the program, we conducted this study to improve student retention rates.

The objectives of this paper are twofold: first, to identify the most effective machine learning technique for predicting IS student program attrition, and second, to examine the factors influencing IS student program attrition.

This paper is organized as follows: Section II reviews empirical studies on student retention factors and the machine-learning algorithms used to predict students at risk of dropping out of an academic program. Section III describes the research methods employed, the dataset, and the technical specifications of the various classifiers analyzed. Section IV discusses the models' performance and assesses the ranking of predictor variables. Finally, Section V presents the conclusions and limitations of the study.

II. Literature Review

Student dropout or attrition prediction is typically viewed as a supervised classification task, where historical student data are used to identify individuals at higher risk of leaving an academic program [20]. Similar predictive techniques have been applied to customer churn prediction in e-commerce and customer relationship management [21–23], employee turnover prediction in human resources [24], and patient retention analysis in healthcare systems [25, 26]. Recent studies highlight that predictive models should not only forecast dropout but also identify its main causes, using methods such as feature importance analysis and explainable artificial intelligence. This enables institutions to develop targeted interventions rather than merely generating retrospective reports [23]. Reviews also suggest that predictive analytics, Machine Learning (ML), and decision-support systems are increasingly employed in higher education to predict dropout rates and improve retention efforts, especially when multiple academic and socio-demographic factors are incorporated into the models [27].

Machine learning-based attrition prediction has become a crucial tool for institutional decision-making. These models assist universities in identifying at-risk students early and implementing targeted retention strategies [28–30]. Research reviews highlight that the effectiveness of predictive models largely relies on data quality, feature engineering, and thorough evaluation procedures. Moreover, modern attrition prediction models are designed not only to forecast dropout outcomes but also to pinpoint influential predictors that support actionable institutional policies [27, 29, 30].

Many empirical studies benchmark different supervised learning algorithms to evaluate predictive performance. Among these, Random Forest often demonstrates strong results. For example, Kondo *et al.* [31] reported that Random Forest outperformed Logistic Regression and Support Vector Machines (SVM) when predicting "at-risk" student status based on GPA-related features. Likewise, Adnan *et al.* [32] compared several algorithms—including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), AdaBoost, Gradient Boosting, Artificial Neural Networks (ANN), and Deep Feed-Forward Neural Networks, and found that Random Forest achieved the highest predictive accuracy (91%) for early-stage dropout risk. Additional studies confirm the effectiveness of ensemble models, Asogwa *et al.* [33] achieved 79.45% accuracy with a Random Forest model to identify computer science students at risk of leaving their programs.

While ensemble models often perform well, other algorithms also show competitive results in different situations. Naseem *et al.* [34], for example, used Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, and KNN to predict first-year dropout among computer science students and found that Logistic Regression achieved the best performance, with an Area Under the Curve (AUC) of 0.8902.

Neural network-based models have also gained attention in attrition prediction. Lee *et al.* [35] compared logistic regression, decision tree, Naïve Bayes, and a Multilayer Perceptron (MLP) neural network for dropout prediction and found that the MLP performed the best, with an F-score of 0.87 and an AUC of 0.98. Similarly, Attiya *et al.* [36] evaluated multiple algorithms, including

decision trees, random forest, logistic regression, neural networks, AdaBoost, XGBoost, and Long Short-Term Memory (LSTM), and reported predictive accuracies above 78%. Bayesian methods have also been used in related retention modeling tasks, such as attendance prediction with Bayesian additive regression trees [37, 38].

A recurring challenge in student attrition prediction is class imbalance, since dropout events typically constitute a small proportion of educational datasets. Therefore, studies emphasize the importance of preprocessing techniques and strategies for handling class imbalance to improve model performance and interpretability. Methods like Synthetic Minority Over-sampling Technique (SMOTE), careful feature encoding, and thorough data cleaning are commonly used to address this issue [20, 26]. At the same time, explainability has become a crucial aspect of predictive modeling. Recent work increasingly employs interpretable ML methods and model-agnostic explanation techniques to enhance transparency and to identify actionable factors influencing dropout risk [39]. This trend aligns with broader cross-domain advances in attrition analytics that focus on interpretable decision support [23].

Beyond predictive modeling techniques, a significant body of research examines the factors influencing student attrition. One of the most consistently identified indicators is academic performance, especially during the early stages of university studies. Students who perform well academically are much less likely to withdraw from their programs [40]. Institutional analytics further emphasize the importance of early academic signals: first-semester GPA and the ratio of earned credit hours are among the most influential predictors of student re-enrollment in later semesters [41]. Recent evidence syntheses also show that academic variables, such as course grades, cumulative GPA, and academic standing, are central to early warning systems because they are measurable, actionable, and closely linked to the timing of institutional interventions [42]. Even when broader contextual factors are included in predictive models, academic performance often remains the top predictor. For example, Delen *et al.* [43] developed a deep artificial neural network model that incorporated socio-economic, demographic, educational, and financial variables and reported an accuracy of 0.884 and an AUC of 0.886, with educational performance identified as the most influential predictor of first-year dropout.

In addition to academic performance, demographic and socioeconomic factors also influence student retention. Several studies emphasize gender differences in persistence, with female students often exhibiting higher completion rates than male students [44, 45]. Age and program characteristics are also commonly included in predictive models to analyze differences in persistence patterns among student groups [29, 42]. Family educational background is another important factor, students whose parents did not attain higher education degrees, often referred to as first-generation students, are more likely to drop out [46]. Research on educational inequality further confirms that first-generation students generally have lower completion rates, with variations across fields of study and academic years [47].

Financial conditions also greatly affect retention outcomes. Students working nearly full-time while studying or relying heavily on loans to pay for their education are more likely to drop out [48]. Conversely, students with stable financial support tend to persist and finish their degrees [49]. Recent research also connects financial stress and employment demands to dropout intentions through mechanisms involving institutional support and social capital [50]. In addition to academic and financial factors, scholars highlight the importance of psychosocial variables, such as student motivation, self-efficacy, and sense of belonging, which interact with socioeconomic disadvantages and academic challenges to influence persistence decisions.

Overall, current research indicates that predicting student attrition requires advanced machine learning techniques and multiple explanatory factors. Effective predictive systems rely not only on algorithmic accuracy but also on the integration of academic, demographic, socioeconomic, and psychosocial variables to enable timely, targeted interventions that enhance student retention.

III. Materials and Methods

A. Machine Learning Classification Models

Machine Learning (ML) techniques have been extensively applied in educational data analytics to predict students' academic performance and identify potential attrition risks. The following section outlines several widely adopted ML algorithms that have demonstrated effectiveness in forecasting student dropout and retention outcomes. Light Gradient Boosting Machine (LGBM) is among the most prominent algorithms in predictive modeling, owing to its high predictive accuracy and computational efficiency [51]. The algorithm is based on gradient boosting principles, employing iterative optimization and weight updating to minimize the loss function. Its performance is enhanced by several distinctive features, including a histogram-based approach for determining optimal split points, a leaf-wise tree growth strategy, and an optimized data storage structure, all of which contribute to faster training and improved model accuracy.

Decision tree algorithms are also widely used for prediction tasks because of their conceptual simplicity and strong ability to identify meaningful patterns in both small and large datasets [52]. A decision tree is usually represented as a binary structure where internal nodes set decision thresholds to guide data instances along specific branches. Terminal (leaf) nodes assign the final class labels. This method is especially effective for pinpointing and visualizing key predictors, as variables near the root nodes indicate the most important factors influencing the outcome. Decision trees are generally computationally efficient for small to medium-sized datasets and provide high interpretability, making them particularly useful in educational prediction settings.

Random Forest (RF) is a popular ensemble learning method used for both regression and classification tasks. The algorithm builds multiple independent decision tree classifiers and combines their outputs to make a final prediction. As noted by Díaz-Uriarte and Andrés [53], RF uses bootstrap sampling to create multiple training subsets from the original dataset. During tree construction, instead of considering all input variables at each split, a randomly chosen subset of features is used to identify the best split. This stochastic process increases diversity among the trees and reduces correlation between individual learners. By combining predictions from multiple uncorrelated trees, the random forest model effectively lowers estimator variance and improves generalization. Before training, several hyperparameters need to be set, including the minimum node size, the number of trees in the ensemble, and the number of features sampled at each split. The final prediction is generated by aggregating the outputs of all trees, typically via majority voting in classification problems.

The Extremely Randomized Trees (Extra Trees) algorithm is another ensemble machine learning approach that has gained significant attention because of its strong classification performance. Similar to random forests, this method combines the predictions of multiple uncorrelated decision trees to produce final classification results [34]. However, it mainly differs in how the decision trees are built. In Extra Trees, tree generation starts directly from the training phase without bootstrap resampling. At each node, candidate split thresholds are chosen randomly, and the best split is selected based on predefined mathematical criteria. Additionally, a random subset of k features is assigned to each tree from the available feature space, increasing randomness in the learning process. This added randomness results in greater diversity among trees, less correlation among learners, and improved prediction stability of the ensemble.

Naive Bayes is a straightforward prediction technique that assumes all x values are independent of each other. It computes 2 probabilities: the probability of each class (y) and the conditional probability of y given x. Naive Bayesian classifiers assume that the effect of a feature value on a given class is independent of the values of other features. As described by Pei *et al.* [52], given m classes and a record X, the classifier predicts that X belongs to the class with the highest posterior probability, conditioned on X. In other words, the Naive Bayesian classifier assigns tuple X to class Ci if and only if $P(Ci|X) > P(Cj|X)$ for all $j \neq i$ (where $1 \leqslant j \leqslant$ m). This approach produces accurate results with large training datasets and many predictor variables. It is especially effective when categorical predictor variables are involved in multi-class classification situations.

B. Dataset Selection and Description

In 2022, the Department of Information Technology at a university in the United Arab Emirates (UAE) updated its information systems program curriculum, adding 3 new courses: object-oriented programming, calculus for information technology, and discrete mathematics.

The paper examines whether these 3 new courses affect IS student attrition. To do this, data from the transcripts of 163 IS students at the university were collected. The factors analyzed in this research include Last Grade Point Average (last GPA), nationality, High School Score (HSS), high school type, gender, English status, and the grades of introductory programming (INT100), object-oriented programming (INT201), and calculus (INT101), along with dropout status, which is a binary variable indicating whether a student dropped out or continued. Among the 163 students, 96 are male, and 67 are female. Additional statistics show that 133 are Arab students and 30 are non-Arab. Classification by secondary school reveals that 108 earned a UAE high school certificate, while 55 obtained their high school certificates from other international schools, such as British, American, or Indian curricula. Analysis by age indicates that 141 students are under 23, with the remaining over 23.

In this research, the collected dataset is relatively small and reveals class imbalance. To address this issue, SMOTE [54] was applied exclusively to the training subset after the dataset was split. The dataset was first divided into training, validation, and testing sets, ensuring that the validation and test sets retained the original dataset's class distribution.

Specifically, the data processing pipeline follows the sequence below. First, the original collected dataset was cleaned by removing incomplete or inconsistent records. Second, the Dataset was split into training, validation, and testing subsets. Third, the SMOTE method was applied only to the training set to generate new samples for the minor class. This step was intentionally performed after the dataset split to prevent any synthetic samples from influencing the validation or testing datasets. Fourth, the ML models were trained on the SMOTE-balanced dataset. Finally, the models' performance was evaluated on the validation and testing sets without resampling or augmentation, thereby maintaining the original data distribution.

The rationale for choosing Accuracy Ranked Feature Inclusion (ARFI) is that the dataset contains only 6 predictor variables: Last GPA, High School Score, Gender, Nationality, High School Type, and English Status. This relatively small feature space makes ARFI appropriate [55], as the method incrementally evaluates variable contributions based on their ranked predictive importance, rather than relying on high-dimensional search procedures that are typically more sensitive to sample-size limitations. Given the limited number of predictors, ARFI provides a transparent and computationally stable mechanism for identifying an informative subset of variables while reducing the risk of unnecessary model complexity and overfitting.

ARFI is also suitable for datasets containing mixed variable types. In our research, the predictors include continuous variables (e.g., Last GPA and High School Score) and categorical variables with 2 or more categories (e.g., Gender, Nationality, High School Type, and English Status). The method is well-suited for educational datasets of this kind, where categorical background characteristics

are common and may still carry important predictive information.

SMOTE has 5 main parameters that directly influence how synthetic samples are generated. These parameters are Sampling Strategy, Number of Nearest Neighbors, Random State, Number of Generated Samples, and Distance Metric. SMOTE generates synthetic minority samples by interpolating between a minority instance and its *k* nearest minority neighbors. The method is primarily controlled by the oversampling ratio, the number of nearest neighbors (*k*), and the random seed used for sample generation [56]. In this research, we have used the following settings: sampling_strategy = 0.6, k_neighbors = 5, random_state = 42, n_samples = default, and the distance metric is Euclidean.

Python offers the SMOTE () method in the imblearn.over_sampling package. This technique creates a more balanced dataset and improves the performance of the classification model. As a result of applying this method, the number of records increased from 163 to 978.

C. Data Pre-processing

Prior to model development, the students' academic records underwent a comprehensive data pre-processing phase to enhance data quality and ensure reliable predictive modeling. Initially, irrelevant and non-informative attributes, including student identifiers and administrative fields, were removed to prevent data leakage and maintain confidentiality. Duplicate records were identified and eliminated.

Records of missing values were removed. Categorical variables were transformed into numerical representations using label encoding for binary attributes such as gender and attrition status. To address variations in measurement scales, all continuous features were normalized using Min-Max scaling, mapping values into the range [0, 1]. Extreme values resulting from data-entry inconsistencies were corrected in the dataset.

Finally, student attrition, the target variable, was defined as a binary target variable (0 = retained, 1 = dropped out). As class imbalance was observed in the dataset, SMOTE was employed during the training phase to improve model generalization and prediction stability.

We used Pearson's correlation coefficient to assess the relationships among the independent variables.

Table I shows the Pearson correlation coefficient between the predictor variables.

TABLE I. PEARSON CORRELATION COEFFICIENT BETWEEN PREDICTOR FEATURES

Feature	Gender	Last GPA	Nationality	High School Type	High School Score	English Status	INT100	INT201	INT101
Gender	1								
Last GPA	−0.44	1							
Nationality	−0.09	0.17	1						
High School Type	0.01	0.09	0.28	1					
High School Score	−0.22	0.3	0.15	−0.2	1				
English Status	0.16	−0.48	−0.28	−0.09	−0.26	1			
INT100	−0.38	**0.75**	0.11	−0.02	0.23	−0.5	1		
INT201	−0.36	**0.8**	0.19	0.06	0.26	−0.48	**0.85**	1	
INT101	−0.35	**0.7**	0.1	0.02	0.21	−0.52	**0.67**	**0.63**	1

D. Technical Specification

Python's Anaconda 3, an open-source package for deploying ML applications, is employed to build and deploy AI models to predict student attrition status, either dropping out or remaining. PyCaret, an open-source Python library, is used to build and compare multiple classification models simultaneously, speeding up experimentation. In addition, PyCaret uses cross-validation to validate the built models and prevent overfitting.

The Cross-Validation (CV) setting establishes the experimental framework for training, validating, and evaluating machine learning models to achieve an unbiased estimate of their generalization performance. Instead of relying on a single train–test split, cross-validation repeatedly divides the dataset into training and validation subsets, allowing for a more robust performance assessment and reducing sensitivity to data sampling variability.

In *k*-fold cross-validation, the dataset is divided into *k* mutually exclusive folds of approximately equal size.

Assume D_1, D_2, …, D_k represents the dataset divided into *k* mutually exclusive subsets (folds). Where *D* is the entire dataset.

For each iteration *i*:

$$\text{Training set:} \quad D_{train}^{(i)} = U_{j \neq i} \times D_j$$

$$\text{Validation set:} \quad D_{val}^{(i)} = D_i$$

where, *i* represents the current iteration index, where $i = 1, 2, \ldots, k$. D_j represents the *j*-th fold, where $j = 1, 2, 3, \ldots, k$. *j* represents the Index used to refer to all folds except the *i*-th fold. $D_{train}^{(i)}$ represents the training dataset at iteration *i*. $U_{j \neq i}$ represents all folds except *Di* are combined to form the training set. $D_{val}^{(i)}$ represents the validation dataset at iteration *i*. D_i represents the *i*-th fold, used for validation. Each fold is used exactly once for validation.

In this study, we used 5-fold cross-validation. The train_test_split() function divides the dataset into training and testing sets with a 70:30 ratio, using default parameters to prepare the data for modeling. 13 models were developed using the following machine learning algorithms: Light Gradient Boosting Machine (LGBM), Random Forest Classifier, K-nearest neighbors' classifier, Gradient Boosting Classifier, Extra Trees Classifier, Quadratic Discriminant Analysis, Decision Tree Classifier, Linear Discriminant Analysis, Ridge Classifier,

Logistic Regression, Naive Bayes, SVM, and AdaBoost Classifier.

IV. Result and Discussion

A. Model Building

The dataset, which includes 978 records, is split into training and testing sets, with 70% (685 records) and 30% (293 records), respectively. The target variable is the IS student status, a binary indicator of whether a student drops out or stays enrolled.

The models were built using 13 machine learning algorithms, with PyCaret helping to predict and explain student attrition. The 13 models developed in this study are categorized into different groups: tree-based ensemble models, including Random Forest, Extra Trees, and Light Gradient Boosting Machine (LightGBM); boosting models such as AdaBoost and Gradient Boosting; distance-based models like KNN; probabilistic classifiers, which include Naive Bayes, QDA, and LDA; and linear classifiers, such as Logistic Regression, Ridge, and Linear SVM. Each category of models shares a common set of parameters. For example, the core parameters for Extra Trees Classifier are n_estimators, criterion, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap, class_weight, and random_state. The key parameters for the Random Forest Classifier include n_estimators, criterion, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap, oob_score, class_weight, and random_state. The main parameters for the Light Gradient Boosting Machine are num_leaves, max_depth, learning_rate, n_estimators, min_child_samples, subsample, colsample_bytree, reg_alpha, reg_lambda, objective, boosting_type, and class_weight. The key parameters for the K-Neighbors Classifier (KNN) include n_neighbors, weights, metric, p, algorithm, and leaf_size. The main parameters for the Gradient Boosting Classifier are n_estimators, learning_rate, max_depth, subsample, min_samples_split, min_samples_leaf, max_features, and loss. The key parameters for the Decision Tree Classifier are criterion, max_depth, min_samples_split, min_samples_leaf, max_features, splitter, and class_weight. The main parameters for Quadratic Discriminant Analysis (QDA) include reg_param, store_covariance, and tol. The main parameters for Bernoulli Naive Bayes Classifier are alpha and binarize. The AdaBoost Classifier's key parameters are n_estimators, learning_rate, estimator, and algorithm. The key parameters for Ridge Classifier include alpha, fit_intercept, solver, class_weight, max_iter, and tol. The main parameters for Logistic Regression are penalty, C, solver, max_iter, class_weight, multi_class, and l1_ratio. The key parameters for Linear Discriminant Analysis (LDA) include solver, shrinkage, n_components, tol, and store_covariance. The last model, Support Vector Machine (Linear Kernel), has key parameters such as kernel, loss, penalty, class_weight, max_iter, and dual.

Moreover, the performance of these models was compared using PyCaret. Table II shows the classifiers' ranking based on several metrics: accuracy, AUC, Recall, precision, and F1-Score.

TABLE II. Comparing the Performance of Several Classifiers Models

Model	Accuracy	AUC	Recall	Precision	F1-Score
Extra Trees Classifier	0.87	0.93	0.75	0.86	0.80
Random Forest Classifier	0.85	0.90	0.71	0.82	0.76
Light Gradient Boosting Machine	0.85	0.91	0.69	0.82	0.75
K Neighbors Classifier	0.84	0.90	0.74	0.77	0.75
Gradient Boosting Classifier	0.82	0.87	0.65	0.78	0.71
Decision Tree Classifier	0.79	0.76	0.66	0.70	0.68
Quadratic Discriminant Analysis	0.78	0.80	0.60	0.70	0.64
Naive Bayes	0.76	0.76	0.56	0.67	0.61
Ada Boost Classifier	0.76	0.79	0.54	0.68	0.59
Ridge Classifier	0.76	0.00	0.46	0.73	0.55
Logistic Regression	0.76	0.75	0.45	0.72	0.55
Linear Discriminant Analysis	0.76	0.75	0.47	0.71	0.56
SVM - Linear Kernel	0.75	0.00	0.57	0.67	0.59

Observations from Table II show that the top 3 classifiers are the Extra Trees Classifier, Random Forest Classifier, and LGBM, with accuracies of 0.87, 0.85, and 0.85, respectively.

B. Models Performance

The hyperparameters of the top 3 performing classifiers—Extra Tree Classifier, LGBM, and Random Forest Classifier, were tuned. For the Extra Tree algorithm, the number of trees in the forest (n_estimators) is set to 100. The minimum number of samples needed to split an internal node (min_samples_split) is set to 2, and the minimum number of samples per leaf (min_samples_leaf) is set to 1. Other hyperparameters were set to default.

For the LGBM classifier, the hyperparameters were tuned as follows: the number of trees in the forest (n_estimators) was set to 100, the learning rate (learning_rate) to 0.09, and the maximum depth (max_depth) to 6. The other hyperparameters remain at their default values. For the Random Forest classifiers, hyperparameters were also tuned. The maximum depth of the tree (max_depth) is set to 5, the minimum samples needed to split an internal node (min_samples_split) is set to 2, and the number of trees (n_estimators) is set to 100. Additionally, the out-of-bag score (OOB_score) is enabled to track correctly predicted rows from the out-of-bag sample. The remaining hyperparameters are set to their default values.

Following the tuning of the 3 classifier hyperparameters, 3 models were built. Table III shows the performance of the top 3 models based on accuracy, AUC, recall, precision, and F1-Score measures.

TABLE III. COMPARING THE PERFORMANCE OF THE TOP 3 PERFORMER CLASSIFIERS

Model	Accuracy	AUC	Recall	Precision	F1-Score
Extra Trees Classifier	0.816	0.919	0.731	0.745	0.738
LGBM	0.786	0.875	0.673	0.707	0.690
Random Forest Classifier	0.765	0.850	0.567	0.711	0.631

Fig. 1 shows the Receiver Operating Characteristic (ROC) curves for the top 3 classifiers across all classification thresholds. Table III and Fig. 1 show that the Extra Tree classifier outperforms the others in predicting student status, followed by the LGBM technique.

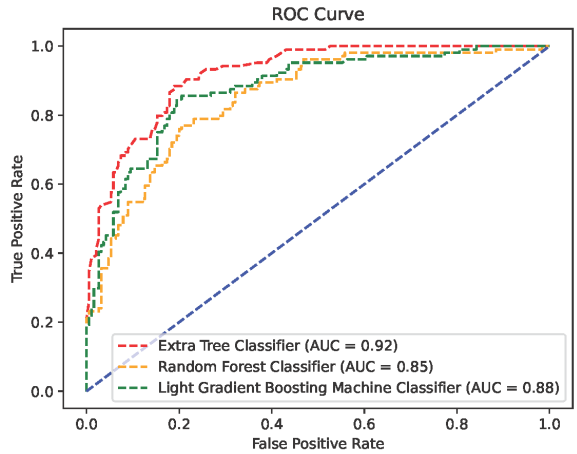

Fig. 1. ROC curve for the top classification prediction models.

C. Ranking of the Predictor Variables

Six variables were used to predict the students' status: whether they are dropping out or remaining. The input variables include Gender, Last GPA, Nationality, High School Type, High School Score, and English Status. The Gender variable is 1 for males and 0 for females. The Last GPA ranges from 0 to 4; nationality values are either Arab or non-Arab. High School Score values range from 50 to 100. The High School Type is a categorical variable with options such as the United Arab Emirates High School Certificate, British High School Certificate, Pakistani High School Certificate, and others. The English Status is either pass (1) or not pass (0).

The importance of feature prediction was estimated using the Accuracy-Ranked Feature Inclusion (ARFI) strategy [57].

The ARFI method is a wrapper-based feature selection and ranking approach that identifies an optimal subset of predictive features by progressively adding them in order of importance and directly assessing model performance. Unlike filter-based methods that rely solely on statistical relevance, ARFI combines feature-importance ranking with iterative model training, enabling the selection of features that enhance predictive accuracy. The ARFI method ranks predictors and selects the best subset of attributes from the ranking. The process involves calculating the model's accuracy with all predictors, then removing one feature at a time to observe the impact on performance. The more essential the feature, the greater the decrease in the model's accuracy when it is removed.

The main steps that we have followed are summarized below.

- The choice of the target metric. In this research, we have chosen an accuracy metric.
- We have selected 3 models, which are Extra Tree, Random Forest, and Light Gradient Boosting Machine (LightGBM).
- We have used a 5-fold cross-validation
- We used fixed random_state to reduce the ranking instability.

Fig. 2 illustrates the importance of features for the top 3-performing models, the Extra Tree Classifier, LGBM Classifier, and Random Forest Classifier, which were employed to predict the IS program student status.

Fig. 2 shows that the Extra Tree classifier's performance drops significantly when the Last GPA feature is removed (from 0.816 to 0.752), indicating that Last GPA is the most important feature. The second- and third-most important features are high school scores and Gender, with accuracy rates of 0.782 and 0.806, respectively, compared to 0.816. The least important feature for predicting student status is English Status (0.85 vs. 0.816), suggesting that English proficiency is not crucial in determining whether a student remains enrolled or drops out.

The second classifier depicted in Fig. 2 is LGBM. For LGBM, as with the Extra Tree classifier, the largest drop in model accuracy occurred when the Last GPA feature was dropped (0.748 vs. 0.786), indicating that this feature is most important.

The second key feature is Gender. The model accuracy when the Gender feature is removed is 0.776, compared to 0.786 before removal, a decrease of 0.01. This shows a clear difference in predicting whether IS program students will drop out or stay, based on gender. Additionally, the LGBM classifier identified the High School Type and Nationality features as tied for third place, with an accuracy drop of −0.01. The least significant feature is the High School Score.

Concerning the Random Forest (RF) classifier, Fig. 2 demonstrates that Last GPA is the most important feature for predicting student status in the IS program. The accuracy was 0.765 before removing the Last GPA; it dropped to 0.707 after its removal, indicating that the feature is the most important predictor. Similar to LGBM, Gender and High School Type are ranked second and third in terms of importance for model prediction. The least significant features shown by RF are high school scores and nationality. Tables IV–VI summarize the feature-importance rankings for the 3 classifiers: Extra Tree, LGBM, and RF, respectively. The tables indicate that all 3 classifiers agreed on the importance of the last GPA as the most significant predictor.

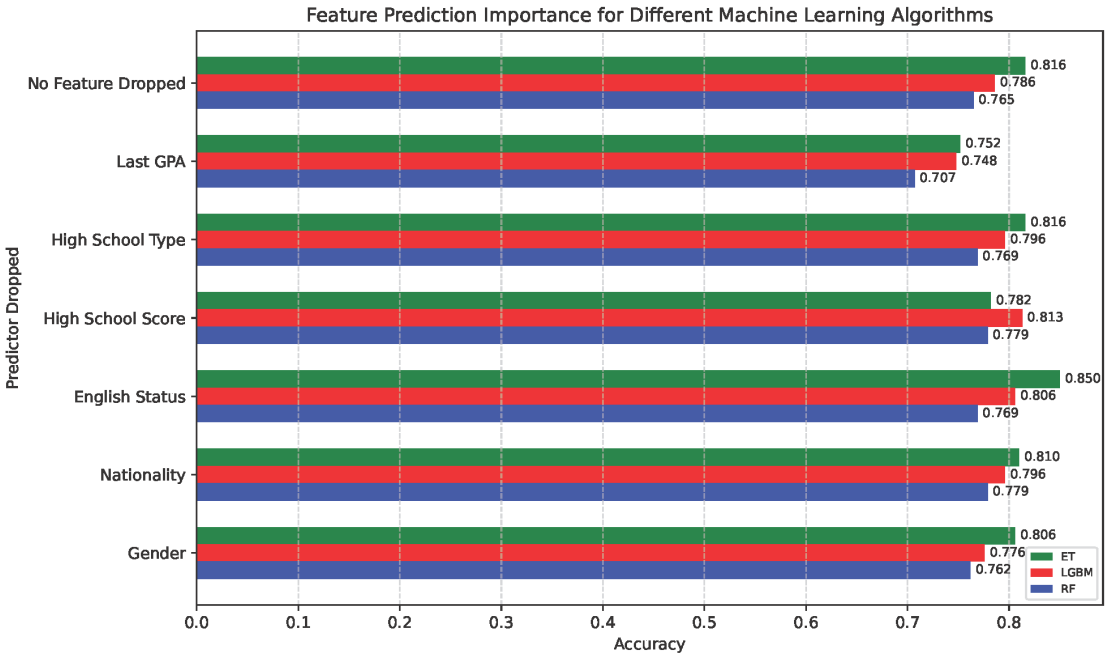

Fig. 2. Feature importance for the 3 top-performing prediction models.

TABLE IV. SUMMARY OF FEATURES IMPORTANCE RANKING: EXTRA TREE CLASSIFIER

Feature Dropped	Accuracy Drop
No Feature Dropped	0
Last GPA	0.064
High School Score	0.034
Gender	0.01
Nationality	0.006
High School Type	0
English Status	−0.034

TABLE V. SUMMARY OF FEATURES IMPORTANCE RANKING: LGBM CLASSIFIER

Feature Dropped	Accuracy Drop
No Feature Dropped	0
Last GPA	0.038
Gender	0.01
High School Type	−0.01
Nationality	−0.01
English Status	−0.02
High School Score	−0.027

TABLE VI. SUMMARY OF FEATURES IMPORTANCE RANKING: RANDOM FOREST CLASSIFIER

Feature Dropped	Accuracy Drop
No Feature Dropped	0
Last GPA	0.058
Gender	0.003
High School Type	−0.004
English Status	−0.004
High School Score	−0.014
Nationality	−0.014

The performance differences among the 3 ensemble models, LIGHTGBM, Random Forest, and Extra Trees, arise from their fundamentally different ensemble learning mechanisms. Random Forest uses bagging with optimal split selection, producing stable but moderately biased predictions. Extra Trees add more randomness in split selection, which greatly reduces variance and enhances generalization in noisy educational datasets. In contrast, LightGBM employs gradient boosting with sequential error correction, enabling it to learn complex feature interactions and subtle early-warning signals for student dropout. As a result, the predictor importance rankings vary across the models.

D. ARFI Significance Testing

To confirm the statistical validity of the selected predictor features, the ARFI strategy incorporates a significance testing mechanism when evaluating the contribution of each feature. ARFI first ranks the candidate features according to their impact on the model's predictive accuracy. Features are then sequentially included in the model following this ranking order. At each iteration, the newly added feature is evaluated by examining whether the observed improvement in predictive performance is statistically significant compared with the model built (ETR, RF, or LGBM) using the previously selected feature subset.

In this research, the statistical significance of each predictor was assessed using *p*-value based hypothesis testing with a significance level of $\alpha = 0.05$, shown in the next section. The null hypothesis assumes that the inclusion of a given predictor does not significantly improve the model's predictive capability. If the *p*-value associated with the predictor is lower than the predefined threshold, the null hypothesis is rejected and the feature is retained in the model. Otherwise, the feature is eliminated from the final predictor subset.

E. P-value Statistics Analysis

Table VII presents the statistical significance of the input features evaluated using the Random Forest model. Feature importance was assessed using permutation-based testing, and the corresponding *p*-values indicate each

variable's contribution to the model's predictive performance. A significance level of $\alpha = 0.05$ was adopted.

TABLE VII. *P*-VALUE ANALYSIS, A = 0.05

Feature	*p*-Value	Significance
Gender	2.32×10^{-2}	Significant
Last GPA	6.80×10^{-10}	Highly significant
Nationality	2.85×10^{-1}	Not significant
Hschool Type	7.00×10^{-2}	Not significant
Hschool Score	4.71×10^{-2}	Significant
English Status	5.22×10^{-3}	Significant

Four variables were found to be statistically significant predictors of student academic performance.

1) Last GPA ($p = 6.80 \times 10^{-10}$)

This feature exhibits the strongest statistical significance among all variables, indicating that prior academic achievement is the most influential predictor in the model.

2) English status ($p = 0.0052$)

English language proficiency is strongly linked to academic performance, highlighting its vital role in instruction, assessment, and learning results.

3) Gender ($p = 0.023$)

Gender shows a statistically significant but relatively moderate impact on performance prediction.

4) High school score ($p = 0.047$)

Pre-university academic achievement significantly contributes to student performance, although its effect is less pronounced than that of Last GPA.

Two features did not demonstrate statistical significance at the 5% level.

5) Nationality ($p = 0.285$)

This variable does not make a statistically significant contribution to the model's predictive ability, indicating that academic performance is not affected by students' nationality within the studied group.

6) High school type ($p = 0.070$)

This variable is not statistically significant, but it nears the threshold and may have a slight effect. Its impact could become clearer with a larger or more varied dataset.

F. Confidence Interval for Attrition Rate

The Confidence Interval (CI) offers a range of plausible values for a population parameter based on sample data. To determine the confidence interval for students' attrition, consider the following assumptions.

$$\text{Let: } Y = \begin{cases} 1, & Dropout \\ 0, & Remain \end{cases},$$

Assume the following statistics:

n = Total number of Information systems students = 163.

x = Number of dropouts = 54. Attrition rate $\alpha = \frac{x}{n} = 0.331288344$. The statistical value for $Z_{0.975}(163) = 1.96$. 95% confidence interval = $\alpha \pm Z_{0.975}\sqrt{\frac{a(1-a)}{n}} = 0.331288344 \pm 1.96\sqrt{\frac{0.331(1-0.331)}{n163}} = 0.331288344 \pm 0.07225781$.

Therefore, the confidence interval is (0.190, 0.473).

We conclude that we are 95% confident that the true dropout rate among students lies between 0.190 and 0.473.

To further strengthen the statistical reliability of the model evaluation, 95% Confidence Intervals (CIs) were estimated for the primary performance indicators, including accuracy, AUC, and F1-Score. These intervals were computed using a bootstrap resampling procedure with 2000 iterations, applied to each model's test predictions. The bootstrap approach provides a non-parametric estimation of the sampling distribution of each metric and is widely used in machine learning studies to assess model stability [58], particularly when datasets are relatively small.

As shown in Table VIII, the results indicate that the Extra Trees classifier achieved the best overall performance with an accuracy of 0.816 (95% CI: 0.772–0.861), AUC of 0.919 (95% CI: 0.889–0.947), and F1-Score of 0.738 (95% CI: 0.670–0.802). The Light Gradient Boosting Machine and Random Forest models achieved slightly lower but comparable performance levels. These confidence intervals demonstrate that the predictive models exhibit consistent and statistically stable performance across resampled datasets.

TABLE VIII. 95% CONFIDENCE INTERVALS FOR MODEL PERFORMANCE FOR THE TOP 3 MODELS

Model	Accuracy	95% CI (Accuracy)	AUC	95% CI (AUC)	F1-Score	95% CI (F1-Score)
Extra Trees	0.816	(0.772–0.861)	0.919	(0.889–0.947)	0.738	(0.670–0.802)
LGBM	0.786	(0.738–0.830)	0.875	(0.833–0.914)	0.690	(0.618–0.756)
Random Forest	0.765	(0.718–0.813)	0.850	(0.807–0.891)	0.631	(0.549–0.708)

We examined the IS program students who dropped out to gain further insights into this finding. Our results show that the Last GPA attribute is the most significant factor influencing students' transfer to other majors that vary in academic difficulty, followed by mathematics performance and gender.

Table IX illustrates the number of students, categorized by gender, who transferred from the IS program to other majors based on their GPA and mathematics performance. Analysis of these results shows that 65% of female students had a GPA over 2.0 and moved to more challenging academic programs, especially in information technology and data analytics. Conversely, 82% of male students had a GPA below 2.0 and switched to less technical majors, such as business administration and law. These findings are supported by a significant difference in mathematics performance between female and male students, as indicated by the average grades for each gender and the student's t-test ($p = 0.0048$), with females performing better than males. These findings have

important implications for enrolling students in the IS program; they highlight the need to advise students on whether the program aligns with their career goals and academic skills.

TABLE IX. STUDENTS' GPA AND PERFORMANCE IN MATHEMATICS BY GENDER

Gender	Student Number			Average GPA	Math	
	Last GPA ≥ 2	Last GPA < 2	Math ≥70		Mean	STD
Male	6	28	9	1.4	59.1	12.83
Female	13	7	14	2.54	68.37	13.05

Gender and students' performance in mathematics are 2 other factors influencing dropout rates. More male students than female students leave the program, which aligns with the findings of Boero *et al.* [44] and Peltier *et al.* [45], who reported that women are more likely to continue their studies than men. Additionally, female students are primarily transferring into more academically demanding majors, such as data analytics and information technology, while male students are moving into less technology-oriented fields, such as management and law. This trend is supported by the observation that female students' performance in mathematics among those leaving the IS program is significantly higher than that of male students, which leads them to shift to information technology or data analytics programs. This migration aligns with Bean's [9] assertion that the "pay" factor is a key indicator of employee turnover, as cybersecurity and data analytics graduates are currently in high demand and earn higher salaries than IS graduates.

Our findings align with previous research, which has consistently reported significant gender differences in academic achievement in higher education. Meta-analytic evidence indicates that female students outperform male students across most disciplines and educational levels, particularly in coursework-based assessments and overall grade point average [59]. These differences are mainly due to behavioural and motivational factors rather than cognitive ability. We have observed that female students spend more time studying.

Our research results also align with educational psychology studies, which demonstrate that female students exhibit higher levels of self-discipline, time management, and self-regulated learning skills, all of which are strong predictors of academic success [60]. Richardson *et al.* [61] further identified effort regulation, engagement, and academic self-efficacy as dominant contributors to university performance, with female students scoring significantly higher on these dimensions.

The successful implementation of a student attrition prediction model requires more than just technical accuracy. It needs systematic integration with institutional systems, clear governance structures, actionable intervention mechanisms, and ongoing performance monitoring. When correctly applied, predictive analytics becomes a valuable institutional tool for improving student success and academic sustainability. The model can be connected with various university systems, such as Student Information Systems (SIS), Learning Management Systems (LMS), and Constituent Relationship Management (CRM) platforms. The proposed model demonstrates how machine-learning based attrition prediction can be transformed from a technical tool into a sustainable decision-support system that promotes student success and academic continuity.

V. CONCLUSION

This research evaluated various machine learning algorithms to predict student attrition in the IS program. The algorithms examined include Extra Trees Classifier, LGBM, Random Forest Classifier, K Neighbors Classifier, Gradient Boosting Classifier, Decision Tree Classifier, Quadratic Discriminant Analysis, AdaBoost Classifier, Logistic Regression, Linear Discriminant Analysis, Ridge Classifier, SVM—Linear Kernel, and Naive Bayes.

The findings indicate that the Extra Trees Classifier performed best, achieving an accuracy of 81.6%. The second-best algorithm is LGBM, with an accuracy of 78.6%. The third top-performing model is Random Forest, which achieved 76.5% accuracy, close to LGBM's, indicating that both models are reliable.

The paper also discussed the factors that influence students' dropout from the IS program. The findings show that the Last GPA is the most significant factor affecting students' status in the three model classifiers, followed by the High School Score for the Extra Tree Classifier. These results are supported by other researchers, who indicate that the cumulative GPA and the Entrance/Admission score (equivalent to the High School Score) are significant and essential predictors of student retention.

This work's results can have important policy implications for admission procedures and academic advising, which could positively influence the attrition rate of the IS program.

The findings of this research offer valuable practical implications for the use of information technology and data analytics in higher education management. By creating machine-learning based models to forecast student attrition in the Information Systems program, this study provides a practical decision-support framework that enables early identification of students at risk of academic withdrawal.

From an operational perspective, the proposed predictive system facilitates proactive student retention efforts. Additionally, academic advisors and program administrators can use the model outputs to identify at-risk students early in their academic journey and implement timely interventions, such as targeted advising, academic mentoring, tutoring, and personalized study plans. This early-warning ability shifts attrition management from a reactive approach to a proactive, data-driven strategy.

Furthermore, the findings also provide actionable insights for program-level academic planning. By quantifying the relative importance of academic performance indicators, engagement metrics, and progression-related variables, the results help Information Systems program coordinators identify structural factors that contribute to student dropout. These insights can inform curriculum refinement, prerequisite alignment,

assessment workload balance, and course sequencing to improve student persistence and the efficiency of progression.

Although the model emphasizes many factors in predicting students at risk of dropping out of the IS program, it is important to note that the results cannot be generalized to other institutions, as each study program has unique features and conditions. These differences make it difficult to generalize the findings. Moving forward, we plan to expand the study by working with other institutions to test the robustness and transferability of the proposed model in different educational settings.

In this study, the instructor's grading style and teaching methodology were not considered, although these factors influence overall course grades. We plan to include these factors in future work. Additionally, we did not consider the students' psychological, socioeconomic, and demographic aspects. Including some or all of these factors could lead to a more accurate prediction of students at risk of changing majors.

ETHICS GUIDELINES

To ensure student privacy, the dataset utilized in this study was thoroughly anonymized before analysis. Personally, identifiable information, such as student names, identification numbers, or contact details, was removed before sharing the dataset with researchers.

The study was conducted in compliance with accepted ethical standards for research involving educational data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

All authors have made significant contributions to the article's conceptualization. EAM, MN, and RM designed the methodology. EAM, MN, RM, and YM participated in the formal analysis. EAM and MN were involved in data collection and preprocessing. RM, MN, and YM drafted the literature review. EAM, MN, and RM contributed to validating the findings and discussing them. All authors participated in the original draft preparation, review, and editing of the final manuscript. All authors have read and approved the published version of the manuscript.

FUNDING

This work was supported by the Deanship of Research and Graduate Studies (DRG) at Ajman University, Ajman, UAE (Grant No.2023-IRG-ENIT-34).

ACKNOWLEDGMENT

The authors wish to thank the Deanship of Research and Graduate Studies (DRG) at Ajman University for its valuable support and encouragement in conducting this study.

REFERENCES

[1] Aljohani, "A comprehensive review of the major studies and theoretical models of student retention in higher education," *Higher Education Studies*, vol. 6, no. 2, pp. 1–18, 2016.
[2] G. Haixiang, L. Yijing, J. Shang *et al.*, "Learning from classimbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.
[3] E. Durkheim, *Suicide*, London, Routledge, 1951.
[4] A. V. Gennep, *The Rites of Passage*, Chicago: University of Chicago Press, 1961.
[5] W. G. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis," *Interchange*, vol. 1, pp. 64–85, 1970.
[6] W. G. Spady, "Dropouts from higher education: Toward an empirical model," *Interchange*, vol. 2, pp. 38–62, 1971.
[7] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, no. 1, pp. 89–125, 1975.
[8] V. Tinto, *Leaving College: Rethinking the Causes and Cures of Student*, Chicago: University of Chicago Press, 1993.
[9] J. P. Bean, "Dropouts and turnover: The synthesis and test of a causal model of student attrition," *Research in Higher Education*, vol. 12, pp. 155–187, 1980.
[10] J. M. Braxton, *Reworking the Student Departure Puzzle*, Nashville: Vanderbilt University Press, 2000. pp. 61–87.
[11] A. Seidman, "Toward reliable knowledge about college student departure," in *College Student Retention: Formula for Student Success*, New York: Praeger, 2005, pp. 107–128.
[12] W. R. Habley, J. L. Bloom, and S. Robbins, *Increasing Persistence: Research-Based Strategies for College Student Success*, San Francisco: Jossey-Bass, 2012.
[13] J. P. Bean, "Conceptual models of student attrition: How theory can help the institutional researcher," *New Directions for Institutional Research*, vol. 1982, no. 36, pp. 17–33, 1982.
[14] J. P. Bean "The application of a model of turnover in work organizations to the student attrition process," *Review of Higher Education*, vol. 6, pp. 129–148, 2017.
[15] C. F. Manski and D. A. Wise, *College Choice in America*, England: Harvard University Press, 1983.
[16] E. P. S. John, *Refinancing the College Dream: Access, Equal Opportunity, and Justice for Taxpayers*, Baltimore: Johns Hopkins University Press, 2003.
[17] E. T. Pascarella and P. T. Terenzini, "Predicting freshman persistence and voluntary dropout decisions from a theoretical model," *The Journal of Higher Education*, vol. 51, no. 1, pp. 60–75, 1980.
[18] J. P. Bean and B. S. Metzner, "A conceptual model of nontraditional undergraduate student attrition," *Review of Educational Research*, vol. 55, no. 4, pp. 485–540, 1985.
[19] A. F. Cabrera. A. Nora, and M. B. Castaneda, "College persistence: Structural equations modeling test of an integrated model of student retention," *The Journal of Higher Education*, vol. 64, no. 2, pp. 123–139, 1993.
[20] A. Villar and C. R. V. d. Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study," *Discover Artificial Intelligence*, vol. 4, 2024.
[21] P. Chou, H. H. C. Chuang, Y. C. Chou *et al.*, "Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning," *European Journal of Operational Research*, vol. 296, no. 2, pp. 635–651, 2022.
[22] C. K. Leung, A. G. M. Pazdor, and J. Souza, "Explainable artificial intelligence for data science on customer churn," in *Proc. 2021 IEEE 8th International Conf. on Data Science and Advanced Analytics (DSAA)*, 2021, pp. 1–10.
[23] V. Chang, K. Hall, Q. A. Xu *et al.*, "Prediction of customer churn behavior in the telecommunication industry using machine learning models," *Algorithms*, vol. 17, no. 6, 231, 2024.
[24] A. Qutub, A. Al-Mehmadi, M. Al-Hssan *et al.*, "Prediction of employee attrition using machine learning and ensemble methods," *Int. J. Mach. Learn. Comput.*, vol. 11, no. 2, pp. 110–114, 2021
[25] C. Kauten, A. Gupta, X. Qin *et al.*, "Predicting blood donors using machine learning techniques," *Information Systems Frontiers*, vol. 24, pp. 1547–1562, 2022.
[26] S. Kanny, G. Post, P. Carbajales-Dale *et al.*, "A comparative approach of machine learning models to predict attrition in a

diabetes management program," *Plos Digital Health*, vol. 4, no. 7, e0000930. 2025.
[27] D. M. Córdova-Esparza, J. Terven, J. A.Romero-González *et al.*, "Predicting and preventing school dropout with business intelligence: Insights from a systematic review," *Information*, vol. 16, no. 4, 326, 2025.
[28] C. Fischer, Z. A. Pardos, R. S. Baker *et al.*, "Mining big data in education: Affordances and challenges," *Review of Research in Education*, vol. 44, no. 1, pp. 130–160, 2020.
[29] L. Addison and D. Williams, "Predicting student retention in Higher Education Institutions (HEIs)," *Higher Education Skills and Work-Based Learning*, vol. 13, no. 5, pp. 865–885, 2023.
[30] R. Chen and K. N. Smith, "Effects of federal loans on first-year college student retention, transfer, and dropout," *Journal of College Student Retention: Research Theory & Practice*, vol. 28, no. 1, pp. 245–273, 2026.
[31] N. Kondo, M. Okubo, and T. Hatanaka, "Early detection of at-risk students using machine learning based on LMS log data," in *Proc. 2017 6th IIAI International Congress on Advanced Applied Informatics*, 2017, pp. 198–201.
[32] M. Adnan, A. Habib, J. Ashraf *et al.*, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021.
[33] D. C. Asogwa, E. C. Asogwa, E. C. Mbonu *et al.*, "Student attrition prediction using machine learning techniques," *International Journal of Computer (IJC)*, vol. 49, no. 1, pp. 16–29, 2023.
[34] M. Naseem, K. Chaudhary, and B. Sharma, "Predicting freshmen attrition in computing science using data mining," *Education and Information Technologies*, vol. 27, pp. 9587–9617, 2022.
[35] J. Lee, M. Kim, D. Kim *et al.*, "Evaluation of predictive models for early identification of dropout students," *Journal of Information Processing Systems*, vol. 17, no. 3, pp. 630–644, 2021.
[36] W. M. Attiya and M. B. Shams, "Predicting student retention in higher education using data mining techniques: A literature review," in *Proc. 2023 International Conf. on Cyber Management and Engineering (CyMaEn)*, 2023, pp. 171–177.
[37] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti *et al.*, "Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile," *Entropy*, vol. 23. no. 4, 485, 2021.
[38] E. Nimy and M. Mosia, "Modelling student retention in tutorial classes with uncertainty—A bayesian approach to predicting attendance-based retention," *Education Sciences*, vol. 14, no. 8, 830. 2024.
[39] N. T. Hai, L. Phuong, N. T. T. Thi *et al.*, "A multivariate analysis of the early dropout using classical machine learning and local interpretable model-agnostic explanations," *CTU Journal of Innovation and Sustainable Development*, vol. 16, pp. 98–106, 2024.
[40] M. Bogard, T. Helbig, G. Huff, and·C. James. A comparison of empirical models for predicting student retention. White Paper. Office of Institutional Research, Western Kentucky University. [Online]. Available: https://www.wku.edu/instres/documents/comparison_of_empirical_models.pdf
[41] D. Delen, K. Topuz, and E. Eryarsoy, "Development of a bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition," *European Journal of Operational Research*, vol. 281, no. 3, pp. 575–587, 2020.
[42] G. S. Gonçalves, F. A. R. Serra, J. E. Storópoli *et al.*, "Undergraduate student retention activities: Challenges and research agenda," *Sage Open*, vol. 14, no. 3, pp. 1–21,2024.
[43] D. Delen, B. Davazdahemami, and E. R. Dezfouli, "Predicting and mitigating freshmen student attrition: A local-explainable machine learning framework," *Information Systems Frontiers*, vol. 26, pp. 641–662, 2024.
[44] G. Boero, T. Laureti, and R. Naylor. An econometric analysis of student withdrawal and progression in post-reform Italian universities. *CRENoS: Centre for North South Economic Research, University of Cagliari and Sassari*. [Online]. Available: https://ideas.repec.org/p/cns/cnscwp/200504.html
[45] G. L. Peltier, R. Laden, and M. Matranga, "Student persistence in college: A review of research," *Journal of College Student Retention: Research, Theory & Practice*, vol. 1, no. 4, pp. 357–375, 2000.
[46] T. T. Ishitani, "Studying attrition and degree completion behavior among first-generation college students in the United States," *The Journal of Higher Education*, vol. 77, no. 5, pp. 861–885. 2006.
[47] S. Weisen, T. Do, M. C. Peczuh *et al.*, "How are first-generation students doing throughout their college years? An examination of academic success, retention, and completion rates," *Analyses of Social Issues and Public Policy*, vol. 24, no. 3, pp. 1274–1287, 2024.
[48] S. L. Britt, D. A. Ammerman, S. F. Barrett *et al.*, "Student loans, financial stress, and college student retention," *Journal of Student Financial Aid*, vol. 47, no. 1, 3, 2017.
[49] L. D. S. Jr., "Come and stay a while: Does financial aid effect retention conditioned on enrollment at a large public university?" *Economics of Education Review*, vol. 23, no. 5, pp. 459–471, 2004.
[50] M. A. S. Toyon, "Effect of university social capital on working students' dropout intentions: Insights from Estonia," *European Journal of Investigation in Health Psychology and Education*, vol. 14, no. 8, pp. 2417–2434. 2024.
[51] K. Ramalingam, P. K. Yadalam, P. Ramani *et al.*, "Light gradient boosting-based prediction of quality of life among oral cancer-treated patients," *BMC Oral Health*, vol. 24, 349, 2024.
[52] J. Pei, J. Han, and M. Kamber. (2012). *Data Mining: Concepts and Techniques*. [Online]. Available: https://liacs.leidenuniv.nl/~bakkerem2/dbdm2007/05_dbdm2007_Data%20Mining.pdf
[53] R. Díaz-Uriarte and S. A. de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, 3, 2006.
[54] N. V. Chawla, K. W. Bowyer, L. O. Hall *et al.*, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[55] P. Kokol, M. Kokol, and S. Zagoranski, "Machine learning on small size samples: A synthetic knowledge synthesis," *Science Progress*, vol. 105, no. 1, 2022.
[56] A. H. Putra and A. Salam, "A comparative performance of SMOTE, ADASYN and random oversampling in machine learning models on prostate cancer dataset," *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 603–610, 2025.
[57] G. S. Thejas, R. Garg, S. S. Iyengar *et al.*, "Metric and accuracy ranked feature inclusion: Hybrids of filter and wrapper feature selection approaches," *IEEE Access*, vol. 9, pp. 128687–128701, 2021.
[58] A. A. Huang and S. Y. Huang, "Increasing transparency in machine learning through bootstrap simulation and shapely additive explanations," *PLoS One*, vol. 18, no. 2, e0281922, 2023.
[59] D. Voyer and S. D. Voyer, "Gender differences in scholastic achievement: A meta-analysis," *Psychological Bulletin*, vol. 140, no. 4, pp. 1174–1204, 2014.
[60] J. Broadbent and W. L. Poon, "Self-regulated learning strategies and academic achievement in online higher education," *Internet and Higher Education*, vol. 27, pp. 1–13, 2015.
[61] M. Richardson, C. Abraham, and R. Bond, "Psychological correlates of university students' academic performance: A systematic review and meta-analysis," *Psychological Bulletin*, vol. 138, no. 2, pp. 353–387, 2012.