A Comparative Study of Machine Learning Techniques for College Student Success Prediction

Zaiyong Tang
Salem State University

Anurag Jain
Salem State University

Fernando E. Colina
Salem State University

The study aims to compare the performance of various machine learning models for student persistence prediction. The research starts with a historical review of student retention studies and the evolution of predictive models in the field. It highlights the importance of predicting student persistence for educational institutions and individuals. It then describes a dataset from ResearchGate, consisting of anonymized undergraduate student data collected between 2008 and 2018, with 37 features and 4,424 records. Ten machine learning algorithms are considered, with two popular machine learning algorithms, Logistic Regression, and Random Forest classification, being compared in more detail for their performance in predicting student persistence. Evaluation metrics such as prediction accuracy, precision, recall, and F1-score are used. Results show that the Random Forest model outperforms Logistic Regression in predicting student outcomes, particularly when using the synthetic minority oversampling technique (SMOTE) to address the class imbalance. Overall, this study contributes to student retention research and provides insights for developing targeted support measures to enhance student success in higher education.

Keywords: student success, prediction, model comparison, logistic regression, random forest

INTRODUCTION

Student persistence is the ability and willingness of students to continue their educational journey and persevere despite challenges and obstacles they may encounter. Research in this domain of student success has consistently been involved in the business of predicting success rates. Over time, multiple areas of challenges faced by students have been discovered. These challenges include factors such as academic difficulties, financial pressures, personal and family issues, and social /cultural pressures that may impact a student's ability to traverse the educational system and succeed academically.

Student persistence studies have been a critical area of research for decades. The National Center for Education Statistics (Kuh et al., 2006) undertook a comprehensive review. The focus of research efforts in this area is to understand the existing and new factors that influence whether a student successfully

completes their program. The study of student persistence is important because it has significant implications for the effectiveness and efficiency of educational institutions and, in some measures, for the social and economic outcomes of individuals and society.

The study of student persistence in higher education began in earnest in the late 60s and early 70s, with research focused on understanding why students leave college before completing their degrees (Bean & Metzner, 1985). Earlier research on student success identified several factors contributing to student attrition, including academic preparedness, financial need, and institutional characteristics. Tinto's (1975) seminal work on student retention, which proposed a theoretical model of student persistence based on social and academic integration, has been highly influential in the field of higher education research. Early instances of using statistical techniques such as factor analysis were deployed to verify and validate Tinto's model (Pascarella & Terenzini, 1980). Other testing models, such as multiple regression (a classic machine learning model), were deployed (Bean, 1980). These aforementioned studies are considered path-setting works.

With the availability of multiple machine learning models, large data sets, computation speed, and advancements in data science and analytics, there has been a growing interest in using machine learning to develop predictive models to study and improve student success. Various types of machine learning models, such as logistic regression, decision trees, random forests, and unsupervised techniques, such as neural networks, are being used to identify various risk factors associated with dropout rates and to develop predictive models to develop intervention strategies and support services. Earlier influential studies used regression models. For instance, Bean and Metzner (1985) migrated from their earlier research of using regression-based causal models to using logistic regression to develop a predictive model of student attrition. The research has been widely cited and built upon in subsequent research on student persistence. These various models have been used to analyze a wide range of data beyond academic performance, including demographics, behavioral data, and engagement with support services. Collectively, those techniques are referred to as predictive models.

Regarding student persistence, predictive modeling can identify at-risk students better than previous methods and assist in developing intervention strategies (Marbouti et al., 2016). Predictive modeling has several advantages for predicting student persistence. First, it has the capability to analyze large and complex data sets. Second, it combines data from multiple sources such as student demographics, academic history, various engagement, and social-economic metrics. Third, predictive modeling can detect patterns in the data that may be difficult for humans to identify through traditional methods, given the volume of data and its complexity. Tinto (2012) makes a case for a data-driven approach to predicting student success. Similarly, others have also made a strong case for the payoff of using machine learning models. For instance, Smith et al. (2012) suggest that predictive modeling is a powerful apparatus for identifying at-risk students and providing targeted interventions to improve their success, especially in online learning. Other studies have supported the application of machine learning techniques for superior student success prediction (Alyahyan & Düştegör, 2020; Ojajuni et al., 2021).

One of the most widely used machine learning libraries for predictive modeling is Scikit-learn (Hackeling, 2017). By utilizing these tools, researchers can better understand the factors that impact student persistence, including student intention, academic achievement, social integration, and institutional policies. Early identification of students at risk of dropping out is crucial to offer assistance and interventions to help them succeed.

In this comparative study of machine learning methods for student persistence prediction, we will first review previous research in the field. This will include an examination of the evolution of student retention studies, from its origins in the 1600s to the integration of predictive models in the last two decades. We will also discuss the challenges faced by higher education institutions, particularly with the rise of online education, and the significance of predicting student persistence for both individual students and institutions. Next, we will describe the data we will use in our study. This will include information on the sources of the data, its scope and coverage, as well as any pre-processing or cleaning steps taken to prepare it for analysis.

We compared the predictive power of ten popular machine-learning models measured by precision, recall, F1-score, and accuracy. Since the sample class is unbalanced, we used the minority class resampling via the synthetic minority oversampling technique (SMOTE). Following the presentation of the experiment results, we discuss the contributions and limitations of the current research and identify several directions for further research.

LITERATURE REVIEW

Interest in student success prediction has remained strong since the Tinto model was introduced (Tinto, 1975). In this seminal paper, Tinto proposes a model of student persistence in higher education that is widely referenced. It lays the foundation for research on student dropout using statistical models. Tinto's theoretical framework was based on Durkheim's work on individual alienation from society (Kerby, 2015). Research interest in this area remains unabated. Several studies have been undertaken in recent years that conducted reviews on the state of student persistence studies that include various retention models, meta-analysis, online, and machine learning models. Some of these recent studies are briefly summarized below.

Bawa (2016) reviewed the literature on student persistence in online education, identifying critical factors for high attrition rates and potential solutions to improve retention. The key factors identified included misconceptions about cognitive load, social and family factors, technological constraints, limitations in faculty training and understanding of online students, and institution limitations.

Manyanga et al. (2017) present a comprehensive review of undergraduate student retention models over the past 80 years. Alyahyan and Düştegör (2020) conducted a comprehensive literature review on predicting academic success in higher education. They provided a step-by-step guide for researchers and practitioners looking to apply data mining techniques. They provide an overview of the state of predictive modeling for academic success in higher education and identify best practices for using predictive models.

Rastrollo-Guerrero et al. (2020) performed a qualitative research study on 64 recent articles on predicting student success and summarized the objectives and techniques used. The main objectives were to study student dropout and academic performance, with only two articles focused on recommending activities and resources. The main techniques included supervised and unsupervised learning, recommender systems, artificial neural networks, and data mining.

Sekeroglu et al. (2021) conducted a systematic review of student performance prediction studies between 2010 and 2020 and found 297 relevant articles. After removing duplicates and non-compliant publications, they summarized 176 articles. Most studies (83.5%) were conducted in the latter half of the decade, indicating a growing interest in student success research. The studies were grouped by objectives, predictive models, datasets, evaluation metrics, and validation strategies.

Machine learning techniques have found tremendous accord with researchers in student success. Often in the early 2000s and prior, with the rise in the interest in Data Mining, machine learning models were discussed and applied under that umbrella to education study and prediction. Romero, C., & Ventura, S. (2010) undertook a comprehensive review and labeled the application of such models as emerging areas. Similarly, Delen (2010) also discusses various machine learning algorithms under the umbrella of data mining. The conclusion was that overall, the algorithms have a prediction accuracy in the context of student retention was over eighty percent. Recent research suggests a departure from the umbrella of data mining to be nestled under artificial intelligence while still deploying the same machine learning models in education (Salas-Pilco & Yang, 2022; Stadlman et al., 2022).

Recent studies demonstrated that machine learning algorithms more effectively predict student persistence in higher education. In a recent meta-analysis, Fahd et al. (2022) present the dramatic rise in machine learning techniques, especially over the past five years. Their meta-analysis also brings to the forefront the number of features in student performance covering a vast area of demographic and socioeconomic background, pre-university and university academic records, and online learning. Their meta-analysis of eighty-nine prior review studies reveals the dominance of logistic regression and the growing application of random forest to achieve greater prediction accuracy. Furthermore, classification studies still tend to dominate the landscape regarding supervised learning. Overall, their analysis suggests

a greater than 80% prediction accuracy with various machine learning models over the eighty-nine studies analyzed.

Similarly, in the context of big data and higher education, Alkhalil et al. (2021) identified eighty-four papers that used various learning machine learning models. They concluded that most studies demonstrated their affinity for supervised learning models. Interestingly they also suggest that most research still tends to be mostly conference proceedings and evaluating reports. In another comprehensive review of the literature, Albreiki et al. (2021) analyzed seventy-eight significant studies in student performance prediction and machine learning spanning over a decade. They analyzed the studies in a chronological fashion date-wise and spanning countries globally. They conclude that student academic performance is a significant predictive variable of success. While classification and supervised learning were the emergent techniques of choice, no particular machine learning model stood out except logistic regression. This was especially true in the case of performance prediction and identification of students at risk. In another review of ten specific studies, Salloum et al. (2020) assert that machine learning techniques are best suited for their predictive capability regarding student success.

Collectively, the reviews suggest that machine learning models have the potential to improve the accuracy of predicting student outcomes and thereby assist institutions in providing targeted support to at-risk students. With the growth of online learning, there has been an emphasis on studying the phenomenon of student attrition beyond the traditional (Rovai, 2003) in the domain of online education or e-learning. Khanal et al. (2020) reported on the growing study of applying machine learning to online education. Further, they concluded that clustering models such as logistic regression, decision trees, and their variants were more popular.

In supervised learning, traditional logistic regression is a very popular machine-learning approach. Segura et al. (2022) use datasets from among the largest and illustrative of a wider range of academic disciplines. They compare several machine learning models under the domain of supervised and unsupervised. They conclude that there is variance amongst each given feature set and the academic discipline. However, overall logistic regression is a reliable foundational machine learning algorithm. Similarly, other studies compare the accuracy of logistic regression with unsupervised algorithms such as neural networks and suggest that logistic regression can achieve superior predictive accuracy through their findings.

Yağcı (2022) used a range of algorithms, including random forests, logistic regression, and k-nearest neighbor, to predict students' final exam grades using a dataset of 1,854 students. The classification accuracy was in the range of 70-75%. The article provides a comparative analysis of 11 recent papers on student success modeling, including objectives, variables, student level, dataset size, algorithms used, and performance results.

Literature also suggests that Random Forest is more accurate than the traditional decision tree classifiers. Falat and Piscova (2022), based on their study of various features and selected machine learning algorithms, suggest through their finding that while regression models are robust, there is a case to use random forest for prediction studies in the domain of supervised learning models. Random forest allowed for better generalizability.

Martins et al. (2021) compared several machine learning models for predicting academic success, finding that the random forest classifier outperformed other models in terms of prediction accuracy and average F1 score. They also noted the common issue of class imbalance in student success prediction, where the dropout/failure rate is much lower than the success rate, and how deploying the SMOTE technique improves model performance.

Batool et al. (2023) conducted a comprehensive review of 260 studies on student performance prediction. They found that artificial neural networks and random forest classifiers were the most commonly used data mining tools. They also noted that nearly half of the studies used feature selection before model building to improve results and reduce processing time.

In another study, Lottering, Hans, and Lall (2020) studied dropout rates via several classification models. Their results suggest that random forests have the highest overall accuracy. However, they do not discount the predictability of classical prediction models such as logistic regression. Similarly, a study by

Moreno-Marcos et al. (2020) in the domain of online education suggests that prediction accuracy is best achieved by random forest followed closely by logistic regression, both achieving higher than eighty percent predictive accuracy. In supervised learning, classical machine learning models such as logistic regression and newer used models, especially random forest, enable greater exploration in advanced learning environments such as online modality. While studying student persistence in an online environment, Moreno-Marcos et al. (2020) concluded that in predicting the efficacy of local persistence versus global persistence, random forest had greater prediction accuracy, while logistic regression predictive ability, while not as strong as random forest, is still good enough.

Research at this point had not arrived at a conclusive consensus on whether logistic regression or random forest is a better model. Earlier research supports the deployment of logistic regression (Chai & Gibson, 2015; Mason et al., 2018). On the other hand, newer research supports using the random forest algorithm regarding prediction accuracy (Falát & Piscová, 2022; Hung et al., 2019; Lottering et al., 2020). In comparison, others offer a balanced view where logistic regression is considered a good enough technique even in the comparison space of unsupervised techniques such as neural networks (Segura et al., 2022). Therefore, overall, in the confluence of prediction accuracy, classification, and supervised machine learning, both logistic regression and now random forest are considered reasonable models; each displays their relative strengths given the backdrop of the study.

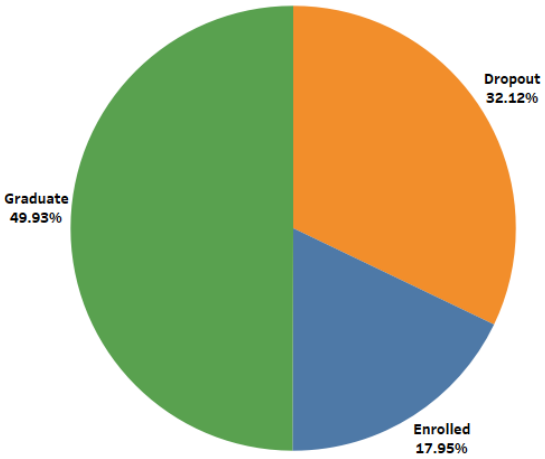
DATA DESCRIPTION

The data set used for this study is obtained from ResearchGate. It can be downloaded via the research entry "Predict students' dropout and academic success" at www.researchgate.net. A subset of the data was used in the publication "Early Prediction of Student's Performance in Higher Education: A Case Study" (Martins et al., 2021).

The data are anonymized undergraduate student data collected between the academic year 2008-2018 at the Polytechnic Institute of Portalegre, Portugal. There are a total of 4,424 records and 37 features/variables. Here is the list of features in alphabetical order: ['Admission grade', 'Age at enrollment', 'Application mode', 'Application order', 'Course', 'Curricular units 1st sem (approved)', 'Curricular units 1st sem (credited)', 'Curricular units 1st sem (enrolled)', 'Curricular units 1st sem (evaluations)', 'Curricular units 1st sem (grade)', 'Curricular units 1st sem (without evaluations)', 'Curricular units 2nd sem (approved)', 'Curricular units 2nd sem (credited)', 'Curricular units 2nd sem (enrolled)', 'Curricular units 2nd sem (evaluations)', 'Curricular units 2nd sem (grade)', 'Curricular units 2nd sem (without evaluations)', 'Daytime/evening attendance\t', 'Debtor', 'Displaced', 'Educational special needs', "Father's occupation", "Father's qualification", 'GDP', 'Gender', 'Inflation rate', 'International', 'Marital status', "Mother's occupation", "Mother's qualification", 'Nationality', 'Previous qualification', 'Previous qualification (grade)', 'Scholarship holder', 'Target', 'Tuition fees up to date', 'Unemployment rate'].

Target is the dependent variable that indicates the outcome of the college students: Graduate, Enrolled, and Dropout. Figure 1 shows that nearly half the students graduated, with about one-third dropping out and nearly one-fifth persisting in enrollment. The case study by Martins et al. (2021) processed the data further to classify the students, based on the time to graduate, into Success, Relative Success, and Failure.

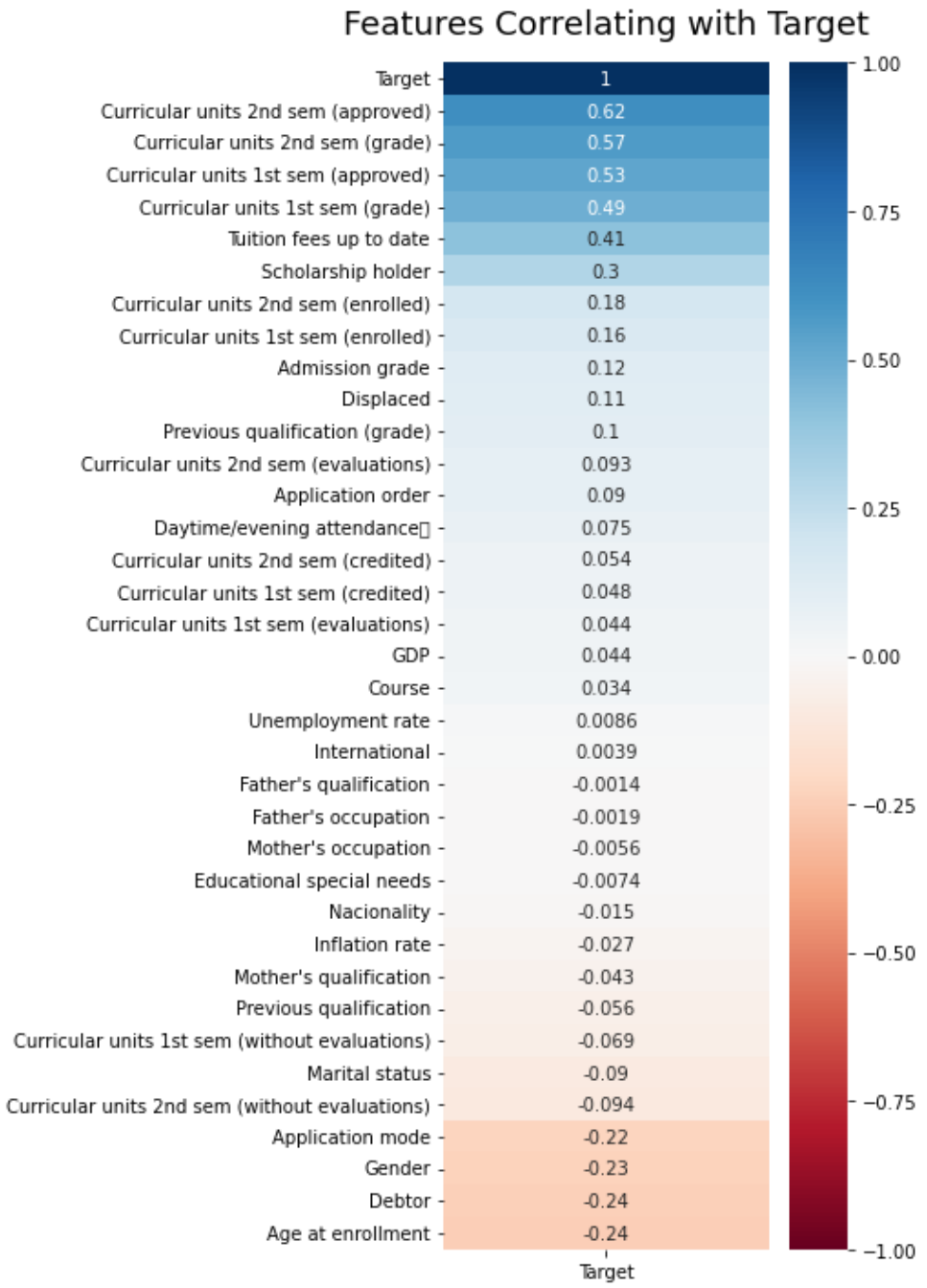
FIGURE 1
STUDENT OUTCOME DISTRIBUTION



The 36 independent variables involve mostly academic and demographic data. There are a few macroeconomic and financial variables. Figure 2 shows the correlation between the target (dependent variable) and the independent variables. Blue indicates a positive correlation, and red indicates a negative correlation. The color shade indicates the strength of the relationship.

Most correlations have intuitive explanations, such as better grades and higher scholarships leading to better outcomes. Some variables, such as parents' occupation and qualification, may not be easy to explain as we do not have details of the data coding. For example, "Father's qualification" has values ranging from 1 to 44. This study will not deal with the interpretation of the variables. We will focus on the predictive power of various machine learning models.

FIGURE 2
CORRELATION BETWEEN TARGET AND INDEPENDENT VARIABLES



Machine Learning Models

We aim to identify the most suitable model for this data set, measured by typical classification metrics: precision, recall, F1 score, and accuracy. Most importantly, we are interested in machine learning techniques to successfully identify at-risk students (dropouts). In terms of college student retention, especially first-year student retention, dropouts are in the minority. However, that is the most important group to focus on because early identification can help prevent this group from increasing across different

year students. This early identification will also provide the necessary support toward student completion and success rate.

We expanded the basket of models and selected from Scikit-learn some popular supervised machine-learning models. Scikit-learn is a free software machine-learning library for the Python programming language. It includes many unsupervised and supervised learning algorithms. Based on preliminary screening, the following models were selected for this study.

1. Bagging Classifier
2. C-Support Vector
3. Decision Tree
4. Extra Tree
5. K-Neighbors
6. Linear Discriminant Analysis
7. Logistic Regression
8. Random Forest
9. Ridge Classifier
10. Stochastic Gradient Descent

We refer the readers to the Scikit-learn website (https://scikit-learn.org/) for documentation, explanations, and sample applications of the machine learning models. The classification metrics are based on the ratios of True Positive (the predicted class is the true class), True Negative (the predicted non-class is the true non-class), False Positive (the predicted class is the true non-class), and False Negative (the predicted non-class is the true class).

Precision = (True Positive) / (True Positive + False Positive). Percentage of correct prediction for the target class.

Accuracy = (True Positive + True Negative) / (Total Sample Size). Accuracy gives overall correct prediction across all classes.

Recall = (True Positive) / (True Positive + False Negative). Percentage of target class overall predicted target class. In other words, recall is the percentage of the class predicted correctly by the model.

F1 score = 2*(Recall * Precision) / (Recall + Precision). - F1 Score is the weighted average of Precision and Recall. It is especially useful when the class sizes are uneven. When the class sizes differ substantially, accuracy as a measure might give a false sense of good performance. Higher F1 scores denote an improved model.

EXPERIMENTAL RESULTS

Ten-fold cross-validation is used to evaluate the performance of the ten machine learning classifiers. The 4,424 record dataset is divided into ten size 442 or 443 subsets. At each run, one subset is reserved for a validation test, while the other nine subsets are used for building the classifier.

First, we gathered the mean and standard deviation of prediction accuracy over the ten cross-validation runs to assess the model's overall performance quickly. The weighted F1 score and accuracy of the classifiers are given in Table 1. The weighted F1 score is computed as the class size weighted average of the F1 scores for all classes.

The Random Forest Classifier has the best F1 score and accuracy (highlighted in the table). Logistic Regression has the second-best accuracy, while Linear Discriminant Analysis has the second-best F1 score. The standard deviation of the accuracy score indicates that the variations of the ten validation runs are relatively small.

TABLE 1
CLASSIFIER ACCURACY COMPARISON

Machine Learning Model	Weighted F1	Accuracy	Std Deviation
Bagging Classifier	0.742	0.750	0.015
C-Support Vector	0.740	0.760	0.014
Decision Tree	0.692	0.689	0.013
Extra Tree	0.653	0.640	0.021
K-Neighbors	0.661	0.673	0.011
Linear Discriminant Analysis	0.749	0.759	0.019
Logistic Regression	0.743	0.764	0.020
Random Forest	**0.763**	**0.781**	0.019
Ridge Classifier	0.712	0.754	0.017
Stochastic Gradient Descent	0.696	0.752	0.021

TABLE 2
PRECISION, RECALL, AND F1 SCORE FOR THREE OUTCOME CLASSES

Machine Learning Model	Metric	Dropout	Enrolled	Graduate
Bagging Classifier	Precision	0.784	0.485	**0.801**
	Recall	0.744	0.393	0.881
	F1 Score	0.763	0.432	0.839
C-Support Vector	Precision	0.851	0.547	0.756
	Recall	0.713	0.303	0.955
	F1 Score	0.774	0.389	0.844
Decision Tree	Precision	0.712	0.384	0.797
	Recall	0.697	**0.412**	0.783
	F1 Score	0.703	0.397	0.790
Extra Tree	Precision	0.673	0.326	0.765
	Recall	0.666	0.349	0.744
	F1 Score	0.669	0.336	0.754
K-Neighbors	Precision	0.731	0.359	0.714
	Recall	0.694	0.256	0.808
	F1 Score	0.711	0.298	0.758
Linear Discriminant Analysis	Precision	**0.884**	0.494	0.769
	Recall	0.698	0.385	0.934
	F1 Score	0.779	0.431	0.843
Logistic Regression	Precision	0.817	0.536	0.774
	Recall	0.765	0.284	0.936
	F1 Score	0.789	0.371	0.847
Random Forest	Precision	0.820	0.589	0.793
	Recall	0.765	0.363	0.935
	F1 Score	**0.791**	**0.446**	**0.858**
Ridge Classifier	Precision	0.819	0.567	0.739
	Recall	0.760	0.139	**0.972**
	F1 Score	0.787	0.220	0.839
Stochastic Gradient Descent	Precision	0.774	**0.619**	0.753
	Recall	**0.790**	0.099	0.956
	F1 Score	0.778	0.149	0.842

The individual metrics in Table 2 reveal that different methods perform better under different conditions. Linear Discriminant Analysis has the best Precision for the Dropout group, while Stochastic Gradient Descent has the best Recall. Random Forest has the best F1 score. For the Enrolled group, Stochastic Gradient Descent has the best Precision, while the Decision Tree has the best Recall. For the Graduate Group, Bagging Classifier has the best Precision, while Ridge Classifier has the best Recall. However, Random Forest has the best F1 Score across all three outcome groups (the best values are highlighted in the table). Since the F1 Score is the weighted average of Precision and Recall, Random Forest is the most balanced model for this dataset.

It is well known that traditional models, such as linear or logistic regression, do not perform well with unbalanced classes in the data set. They tend to do well in predicting the majority class but poorly with the minority class. Since recall is the percentage of the class predicted correctly by the model, the low recall score for the minority class could be a concern if correct identification of the minority class is important. The machine learning community has developed various approaches to address that concern. Thammasiri et al. (2014) report that among the three class balancing methods they tested, the synthetic minority over-sampling technique (SMOTE) outperformed random under-sampling and random over-sampling. We used over-sampling of the minority classes via SMOTE.

Synthetic samples of the minority class are generated to increase the sample size of the minority class. SMOTE selects instances in the minority class and finds the K nearest neighbors in the same class. The pairs of selected instance A and a randomly selected neighbor B are connected via a line in the feature space. The synthetic instances are randomly drawn along the lines. The data set has the following sample distribution: {Dropout: 1421, Enrolled: 794, Graduate: 2209}. After applying SMOTE, the sample distribution is even: {Dropout: 2209, Enrolled: 2209, Graduate: 2209}.

TABLE 3
CASSIFIER ACCURACY WITH MINORITY CLASS RESAMPLING

Machine Learning Model	Accuracy	Accuracy (SMOTE)
Bagging Classifier	0.750	0.810
C-Support Vector Classification	0.760	0.756
Decision Tree Classifier	0.689	0.745
Extra Tree Classifier	0.640	0.742
K-Neighbors Classifier	0.673	0.724
Linear Discriminant Analysis	0.759	0.720
Logistic Regression	0.764	0.728
Random Forest Classifier	**0.781**	**0.840**
Ridge Classifier	0.754	0.706
Stochastic Gradient Descent Classifier	0.752	0.710

Table 3 shows that oversampling the minority classes with SMOTE improves the overage prediction accuracy for 5 of the ten models compared to the accuracy presented in Table 1. The best performer without resampling, Random Forest Classifier, is again the best performer with minority oversampling. The accuracy improvement is noticeable for the model. On the other hand, Logistic Regression, a commonly used classifier, has decreased accuracy with minority oversampling.

The fact that balancing the sample size improves the prediction accuracy of only half of the models tested is unexpected. Realizing that accuracy, which measures overall correct prediction across all classes, does not align well with our goals of identifying the at-risk student (Dropouts), we want to see SMOTE's impact on individual classes. We carried out a more detailed analysis using two models: Logistic Regression and Random Forest.

We selected these two models because a) Logistic Regression is one of the most applied machine learning models for classification, and b) Random Forest has the highest prediction accuracy among the ten

models. It is also a popular method among members of the data science community. Anecdotally, the random Forest is considered to mimic human decision-making closely. Additionally, recent studies and literature reviews suggest the popularity of these methods. By evaluating prediction accuracy, in-class precision, recall, and F1 score, we aim to build an analytical model with high predictive power to identify at-risk students.

We present the individual metrics of the two models in Table 4. The bold-faced values indicate improvement over the results without minority oversampling.

TABLE 4
PRECISION, RECALL, AND F1 SCORE WITH/WITHOUT MINORITY OVERSAMPLING

Machine Learning Model	Metric	Dropout	Enrolled	Graduate
Logistic Regression (**without SMOTE)**	Precision	0.817	0.536	0.774
	Recall	0.765	0.284	0.936
	F1 Score	0.789	0.371	0.847
Logistic Regression (with SMOTE)	Precision	**0.823**	**0.647**	0.730
	Recall	0.713	**0.669**	0.805
	F1 Score	0.764	**0.657**	0.765
Random Forest (**without SMOTE**)	Precision	0.820	0.589	0.793
	Recall	0.765	0.363	0.935
	F1 Score	0.791	0.446	0.858
Random Forest (with SMOTE)	Precision	**0.897**	**0.803**	**0.829**
	Recall	**0.798**	**0.829**	0.892
	F1 Score	**0.845**	**0.816**	**0.859**

With SMOTE, the Logistic Regression Model does a better job in correctly predicting the minority class, while the performance generally deteriorates slightly in other classes of the outcome. However, the Random Forest Classifier, with minority oversampling, improves its performance across all three outcome classes. Also, the improvement of the metrics in the minority class is very significant.

Figure 3 and Figure 4 are the classification confusion matrices that give the percentage of correct prediction for all three classes. Without minority oversampling, Logistical Regression and Random Forest do poorly with the minority class. The majority class (Graduate) is clearly favored. However, with SMOTE oversampling, both models can predict with high correct ratios for all three classes. The Random Forest outperforms the Logistic Regression with 80% or better correct prediction. Since we used 10-fold cross-validation, all the results are the average over ten runs.

FIGURE 3
CONFUSION MATRIX WITHOUT SMOTE

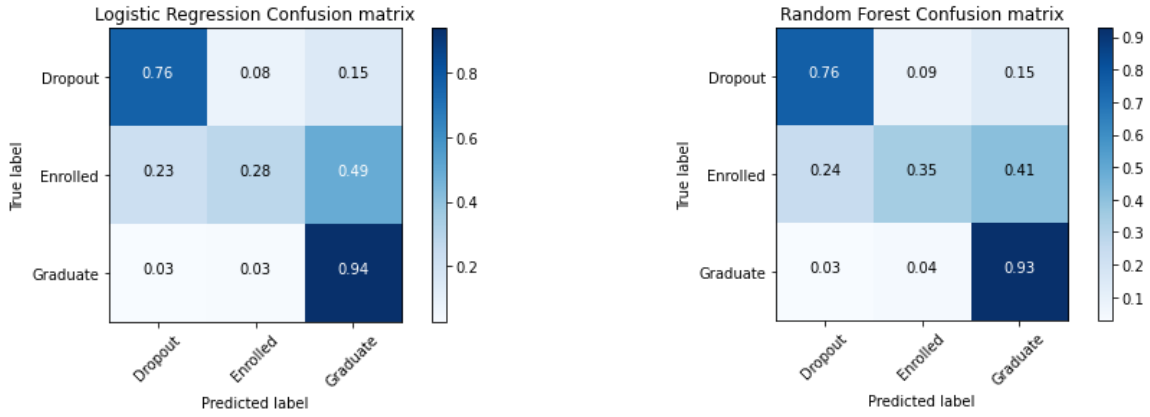
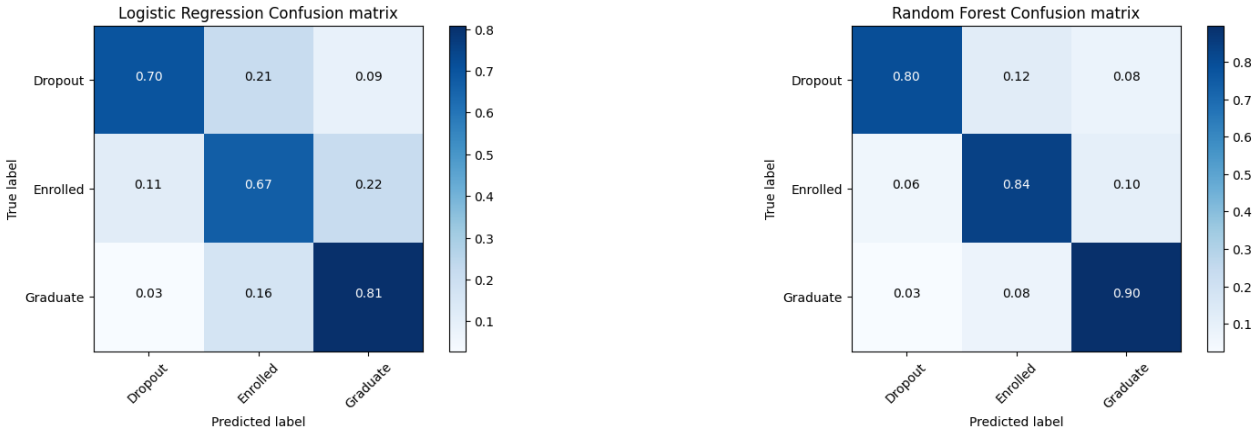


FIGURE 4
CONFUSION MATRIX WITH SMOTE OVERSAMPLING



Comparing the individual class prediction accuracy in Table 3 and Table 4, we see that for the Logistic Regression model, the minority class Enrolled shows a significant improvement in correct prediction rate (from 28% to 67%) when SMOTE oversampling is used. However, the prediction rate decreases for the other two classes. Thus, the value of SMOTE is limited. On the other hand, for the Random Forest mode, SMOTE improves the correct prediction for both the Dropout and Enrolled groups. The decrease in the correct prediction for the Graduate class is relatively small. Furthermore, as shown in Table 4, the prediction accuracy of the Random Forest model is significantly better than that of the Logistic Regression model in all three target classes.

DISCUSSION

The growing use of machine learning in student persistence or retention research allows for identifying complex patterns and relationships among multiple variables that impact student persistence. One of the advantages of using machine learning is identifying new factors of student persistence (Chen et al., 2021). Traditional statistical techniques are limited in their ability to identify complex, non-linear relationships among multiple predictors. In contrast, machine learning algorithms can help identify patterns and relationships that are difficult to detect with traditional methods. For example, studies have used machine

learning to identify the impact of a wide range of factors, including student demographics, academic performance, engagement, and institutional characteristics, on student persistence. Another advantage of using machine learning is its ability to predict student outcomes accurately.

Logistic regression can be used to explore the factors related to student persistence, the extent to which students continue their education and complete their degree programs. In this analysis, student persistence is the dependent variable, and the predictor variables can span various influencing domains such as demographic characteristics, academic background, socio-economic status, and other relevant factors. Given various circumstances, the logistic regression model is well suited to estimating the probability of a student persisting in their education. One of the main features of logistic regression is that it allows controlling for confounding variables. For example, if we want to explore the relationship between academic background and persistence, we can control for other factors such as gender, age, race, and socioeconomic status to isolate the effect of academic background on student persistence. Another uniqueness of logistic regression is that it can handle non-parametrized data, i.e., non-linear data. This is significant because the association between variables and persistence may not be linear, and logistic regression can capture this complexity.

Random Forest is a popular machine-learning algorithm. One of the primary advantages of using random forest is its ability to handle large datasets with many variables. Random Forest can generate and combine many decision trees to form a robust predictive model. This uniqueness makes it well-suited for investigating factors predicting student persistence. Our experimental results support the conclusion that the Random Forest model outperforms the Logistic Regression model, especially when the SMOTE oversampling is used.

A significant limitation of the study lies in the interpretability of machine learning models. Both Random Forest and Logistic Regression may be categorized as black boxes. They may produce accurate predictions, but how they do so is not easily interpretable. These approaches make understanding the underlying factors for persistence more difficult to assess, which in turn affects the actions practitioners must take to improve student persistence. For example, some features (e.g., parental occupation and qualification) are not easily interpretable. Indeed, the study offers little guidance on what can be done about these features. The synthetic minority oversampling technique (SMOTE) was used to address the issue of imbalance in relation to student dropouts. The use of the technique should be taken into consideration when analyzing results.

Another major consideration is that the study used academic and demographic data to predict persistence. There are likely socio-economic, cultural, and other factors that affect persistence, but which were not included. A comprehensive understanding of student persistence should take into consideration these external factors. The models, furthermore, do not establish causality between the features and the outcome, meaning that interpretation of the results should be done with care. The dataset was derived from the Polytechnic Institute in Portalegre, Portugal, which may have had a population of students that may not provide generalizable results applicable elsewhere. For example, there may have been a high degree of homogeneity among the students studied, meaning that the sample would not be representative of the population of students for which the algorithm is intended.

Two models were highlighted, Logistic Regression and Random Forest. While these are widely in use, there may be more effective models which may be overshadowed by these two popular models. For example, ensemble models may provide greater usefulness than either of these models alone. Since it is nearly impossible to document how machine learning algorithms can predict with such a high level of accuracy, one possible drawback may be an overfitting of the data that may not be readily apparent, even though tenfold cross-validation is used for building the predictive models. Subsequently, the models may not predict as accurately with different data, limiting the practicability of the algorithms. The study's timeframe was from 2008 to 2018, preceding the COVID-19 pandemic. Subsequent changes in student demand for higher education, the increase in student swirl among different institutions, and changes in academic expectations by students and faculty were not included in the study, raising questions about the generalizability of the models post-COVID-19. Despite these issues, the study presents readers with a

promising outlook on using machine learning to classify student persistence and lays the groundwork for future research.

Extensions to the current study include diving deeper into the prediction results analysis and interpretation. For example, building more parsimonious models that are more efficient and more general through individual feature analysis and selection. This can help researchers identify the most important factors that impact student persistence. Those variables can be used to develop prevention measures to help at-risk students by focusing on areas where they face challenges. Furthermore, the Random Forest model can generate feature contributions to individual prediction results. Thus, we can examine individual students' likelihood of persistence. For those at-risk students, we will be able to identify which factors are causing the dropout risk and develop individualized actionable recommendations for those students. Although it is widely recognized that academic factors such as GPA play a key role in student persistence, many socioeconomic, demographic, and community engagement factors are also important. Future studies may incorporate a wider range of those important factors.

CONCLUSION

Student persistence is an important aspect of educational success and is closely linked to academic achievement, personal development, and lifelong learning. Using machine learning models to study student persistence in higher education has significant potential to identify new predictors and accurately predict student outcomes. We compared ten machine learning models for student success prediction. All ten models show good classification accuracy, but their strengths vary with different prediction measures. Overall, Random Forest produces the best results. The remarkable performance of the Random Forest model is enhanced when the SMOTE oversampling of the minority data class is implemented. Although Logistic Regression is widely used in regression and classification problems and has good performance in student persistence prediction with the dataset we used, we recommend using the Random Forest model for its superior performance and additional capabilities in feature analysis.

REFERENCES

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552–579. https://doi.org/10.3390/educsci11090552

Alkhalil, A., Abdallah, M.A.E., Alogali, A., & Aljaloud, A. (2021). Applying big data analytics in higher education: A systematic mapping study. *International Journal of Information and Communication Technology Education (IJICTE)*, *17*(3), 29–51. https://doi.org/10.4018/IJICTE.20210701.oa3

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–21. https://doi.org/10.1186/s41239-020-0177-7

Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), 905–971. https://doi.org/10.1007/s10639-022-11152-y

Bawa, P. (2016). Retention in online courses: Exploring issues and solutions—A literature review. *Sage Open*, *6*(1), 1–11. https://doi.org/10.1177/2158244015621777

Bean, J.P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, *12*(2), 155–187. https://doi.org/10.1007/BF00976194

Bean, J.P., & Metzner, B.S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, *55*(4), 485–540. https://doi.org/10.3102/00346543055004

Chai, K.E., & Gibson, D. (2015). Predicting the risk of attrition for undergraduate students with time-based modelling. In D.G. Sampson, J.M. Spector, D. Ifenthaler, & P. Isaias (Eds.), *Proceedings of cognition and exploratory learning in the digital age* (pp. 109–116). IADIS Press.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Fahd, K., Venkatraman, S., Miah, S.J., & Ahmed, K. (2022). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Education and Information Technologies*, *27*(3), 1–33. https://doi.org/10.1007/s10639-021-10741-7

Falát, L., & Piscová, T. (2022). Predicting GPA of university students with supervised regression machine learning models. *Applied Sciences*, *12*(17), 8403. https://doi.org/10.3390/app12178403

Hackeling, G. (2017). *Mastering machine learning with scikit-learn* (2nd Ed.). Packet Publishing Ltd.

Hung, J., Shelton, B.E., Yang, J., & Du, X. (2019). Improving predictive modeling for at-risk student identification: A multistage approach. *IEEE Transactions on Learning Technologies*, *12*(2), 148–157. https://doi.org/10.1109/TLT.2019.2911072

Kerby, M.B. (2015). Toward a new predictive model of student retention in higher education: An application of classical sociological theory. *Journal of College Student Retention: Research, Theory & Practice*, *17*(2), 138–161. https://doi.org/10.1177/1521025115578229

Khanal, S.S., Prasad, P.W.C., Alsadoon, A., & Maag, A. (2020). A systematic review: Machine learning based recommendation systems for e-learning. *Education and Information Technologies*, *25*(4), 2635–2664. https://doi.org/10.1007/s10639-019-10063-9

Kuh, G.D., Kinzie, J.L., Buckley, J.A., Bridges, B.K., & Hayek, J.C. (2006). *What matters to student success: A review of the literature*. Washington, DC: National Postsecondary Education Cooperative.

Lottering, R., Hans, R., & Lall, M. (2020). A machine learning approach to identifying students at risk of dropout: A case study. *International Journal of Advanced Computer Science and Applications*, *11*(10), 417–422. https://doi.org/10.14569/IJACSA.2020.0111052

Manyanga, F., Sithole, A., & Hanson, S.M. (2017, Spring). Comparison of student retention models in undergraduate education from the past eight decades. *Journal of Applied Learning in Higher Education*, *7*, 30–42. https://doi.org/10.57186/jalhe_2017_v7a3p30-39

Marbouti, F., Diefes-Dux, H.A., & Madhavan, K. (2016, December). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005

Martins, M.V., Tolledo, D., Machado, J., Baptista, L.M., & Realinho, V. (2021). Early prediction of student's performance in higher education: A case study. Paper presented at the *World Conference on Information Systems and Technologies*, pp. 166–175. https://doi.org/10.1007/978-3-030-72657-7_16

Mason, C., Twomey, J., Wright, D., & Whitman, L. (2018). Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. *Research in Higher Education*, *59*(3), 382–400. https://doi.org/10.1007/s11162-017-9473-z

Moreno-Marcos, P.M., Muñoz-Merino, P.J., Alario-Hoyos, C., & Delgado Kloos, C. (2020). Re-defining, analyzing and predicting persistence using student events in online learning. *Applied Sciences*, *10*(5), 1722. https://doi.org/10.3390/app10051722

Ojajuni, O., Ayeni, F., Akodu, O., Ekanoye, F., Adewole, S., Ayo, T., Misra, S., & Mbarika, V. (2021). Predicting student academic performance using machine learning. Paper presented at the *Computational Science and its Applications–ICCSA 2021*, *12957*, 481–491. https://doi.org/10.1007/978-3-030-87013-3_36

Pascarella, E.T., & Terenzini, P.T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The Journal of Higher Education*, *51*(1), 60–75. https://doi.org/10.1080/00221546.1980.11780030

Rastrollo-Guerrero, J., Gómez-Pulido, J.A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences, 10*(3), 1–16. https://doi.org/10.3390/app10031042

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Rovai, A.P. (2003). In search of higher persistence rates in distance education online programs. *The Internet and Higher Education, 6*(1), 1–16. https://doi.org/10.1016/S1096-7516(02)00158-6

Salas-Pilco, S., & Yang, Y. (2022). Artificial intelligence applications in Latin American higher education: A systematic review. *International Journal of Educational Technology in Higher Education, 19*(1). https://doi.org/10.1186/s41239-022-00326-w

Salloum, S.A., Alshurideh, M., Elnagar, A., Shaalan, K., & Tolba, F.M. (2020). Mining in educational data: Review and future directions. Paper presented at the *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), 1153*, 92–102. https://doi.org/10.1007/978-3-030-44289-7_9

Segura, M., Mello, J., & Hernández, A. (2022). Machine learning prediction of university student dropout: Does preference play a key role? *Mathematics, 10*(18), 3359. https://doi.org/10.3390/math10183359

Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., & Idoko, J.B. (2021). Systematic Literature Review on Machine Learning and Student Performance Prediction: Critical Gaps and Possible Remedies. *Applied Sciences, 11*(22), 10907. https://doi.org/10.3390/app112210907

Smith, V.C., Lange, A., & Huston, D.R. (2012). Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of Asynchronous Learning Networks, 16*(3), 51–61. https://doi.org/10.24059/olj.v22i2.1369

Stadlman, M., Salili, S.M., Borgaonkar, A.D., & Miri, A.K. (2022). Artificial intelligence based model for prediction of students' performance: A case study of synchronous online courses during the COVID-19 pandemic. *Journal of STEM Education: Innovations and Research, 23*(2), 39–46.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41*(2), 321–330. https://doi.org/10.1016/j.eswa.2013.07.046

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Tinto, V. (2012). *Completing college: Rethinking institutional action*. The University of Chicago Press.

Yagci, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z