Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Understanding the Prediction of Student Retention Behavior during covid-19 Using Effective Data Mining Techniques

Rezwanul Parvez (rezwanul.parvez@colostate.edu)
Colorado State University
Alysha Tarantino
Colorado State University
Syed Imran Ali Meerza
Arkansas Tech University

Research Article

Keywords: Data mining techniques, student retention, machine learning methods, SMOTE

Posted Date: May 22nd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2948727/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Understanding the Prediction of Student Retention Behavior during covid-19 Using Effective Data Mining Techniques

Rezwanul Parvez[1], Alysha Tarantino[2], and Syed Imran Ali Meerza[3]
[1]Colorado State University, Fort Collins, CO 80526 USA
[2]Colorado State University, Fort Collins, CO 80526 USA
[3]Arkansas Tech University, Russellville, AR 72801 USA

Corresponding author: Rezwanul Parvez (e-mail: rezwanul.parvez@colostate.edu).

This work was supported by Colorado State University through the Office of Engagement and Extension.

ABSTRACT Student success, including enrollment, degree completion, and improved retention, has been adversely affected due to the covid-19 pandemic across the higher academic landscape in the United States. It is quite a challenge to accurately predict what factors impact student retention in an academic institution most. This research aims to develop and empirically test a comprehensive list of factors contributing to student retention behavior using data mining techniques (DMT). This study examines retention behavior prediction from the dataset (pre-pandemic and post-pandemic) of 18,000+ students enrolled at an academic institution. Here, we deploy six different machine learning (ML) models (e.g., Logistic Regression (LR), Support Vector Machine (Linear Classifier), Support Vector Machine (Radial Basis Function), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN) to predict student retention. Synthetic Minority Oversampling Technique (SMOTE) is also being used to fix the imbalance class problem and improve the performance of data mining algorithms. Empirical results showed that completed credits, grade point average (GPA), college entrance age, and attempted credits were vital factors in student retention behavior. This research also highlighted that RF algorithms outperformed other DMTs' and achieved the highest accuracy (0.86) with SMOTE for retention prediction. The critical results discussed here should apply to similar higher academic institutions worldwide. Academic institutions would benefit from launching preventative measures to avoid dropouts and improve retention based on the retention metrics reported here.

INDEX TERMS Data mining techniques, student retention, machine learning methods, SMOTE

I. INTRODUCTION

The current global pandemic has severely affected student enrollment and success across academic institutions in the nation. Both students and parents are rethinking their higher education plans due to safety concerns [32]. From an academic institution's perspective, student retention is getting more attention than life before the pandemic. Many academic institutions have expressed concerns about keeping enrollment high and are facing a donor shortage to improve institutions' endowments. Also, some academic institutions were forced to close their programs due to pandemics [14]. Students' mental health, as well as academic performance, have been seriously affected by the pandemic. Thousands of students across academic institutions faced challenges in maintaining good academic standing while dealing with the Covid-19 pandemic. Within a short time, institutions had to allocate resources to provide additional support for students' well-being. Not all institutions have been successful in doing that. An example would be moving support services and advising to the virtual space, which was challenging for many educators and students. Academic institutions had to adopt a large-scale online learning transition [38]. As a result, we see a declining trend in enrollment, higher student dropout, and lower alums/donor giving rates. The global pandemic has seriously impacted students' mental health and resulted in "a period of confusion" and significant stress impacting higher education. Recently published literature indicated an elevated level of academic stress among students due to the pandemic [33]. Mental stress also leads to negative student satisfaction and student retention. Due to this, postsecondary institutions are dealing with cross-disciplinary challenges to identify and implement effective measures to retain more students [33], [19], [21]. One way to tackle this declining trend in enrollment is to allocate resources to retain more students and understand the underlying factors.

As defined by Integrated Postsecondary Education Data System (IPEDS), the retention rate is a measurement (expressed as a percentage) at which students persist in their educational program [27]. In other words, retention rates are calculated as the percentage of first-time undergraduate students who return to the same institution the following fall. "Higher Education Act" identifies retention and graduation as key measures to assess institutional effectiveness [11].

From an institution's perspective, retaining students becomes the top priority to minimize the damage caused by the pandemic. Retaining students is being identified as one of the critical measures of an institution's overall health and an integral part of its strategic planning process. As student retention conversations gain popularity these days, institutions have started allocating more money and resources to identify student retention metrics. This includes an investigation into and addressing dropouts, and students' unwillingness to return to the same institution. Recently published and existing education literature indicates a downward trend of college retention and graduation rates across academic institutions [27], [10]. As per the National Center for education statistics [27], retention does vary across institutions (59% to 96% at public 4-year institutions while 63% to 81% at 4-year private institutions). The trend is even more alarming for 2-year degree-granting institutions, including community colleges (61% for public 2-year and 68% for private 2-year). These trends of declining retention across institutions could be more challenging for the economy, raising concerns for the public interest, and reflect a significant loss of time and monetary capital at all (individual, state, and national) levels. As published in the literature, the challenge of retaining more students is evident across academia [16], [20], [26]. As a result, a comprehensive assessment of exploring the matrices affecting student retention behavior is worth pursuing.

The metrics of student retention are considered one of the contributing factors in assessing an institution's effectiveness and financial stability. Higher retention which leads to higher student success and graduation rates indicates a healthy institution. From the public policymakers' perspective, retention is critical to evaluating institutions' accountability and institutional effectiveness [11]. Also, academic institutions must ensure students' successful college experience by providing quality communication, instructions, and better student engagement. Surprisingly prior to the pandemic, institutions tend to consider retention as a subject of lesser importance [43]. Each academic institution has different management practices for addressing student retention behavior. Understanding the theory and matrices behind a student's decision to stay or leave an academic institution is critical. According to Tinto [41], [42], student development theory (student integration model) defines how students progress through various stages. Multiple factors (e.g., group study, social and academic integration, family background, abilities, and former schooling) influence students' success of whether a student will complete their program of study. A retention study [42] presents a relationship between pre-college entry factors, institutional characteristics, integration with goals and outcomes, and institutional incidents (e.g., faculty interactions, activities on campus, and peer interactions). Another theory (psychological model of retention) suggests environmental variables and students' intentions as retention drivers [18], [12]. Additionally, Astin [3], [4] reported "student involvement theory (SIT)" and "input-environment-outcome model (IEO)" theories explaining retention. These theories include macro-level data across hundreds of academic institutions. The SIT theory suggests the dependence of student retention on student demographics (e.g., age, race, gender) and institutional characteristics (e.g., location, size, level of academic and social involvement). However, the IEO theory predicts retention using relationships between outputs (e.g., degrees earned, number of graduates) and inputs (e.g., gender, age, student ability, primary areas of study). Besides this, Bean's [17] model of student attrition highlights the fact that student satisfaction depends on services offered at the institution (e.g., grades, values, peer support, and rewards) [25], [37], [40]. Including key institutional data points (e.g., registered credits, campus location, award recipients, academic performance, parental education, student demographics) along with output variables is necessary to examine the critical matrices of student retention. Existing literature has attempted to identify metrics driving student retention using empirical methods [30], [16], [13], [26]. Windham *et al.* [26] adopted a post-facto quasi-experimental method to assess factors (ethnicity, age, gender, socio-economic status, ACT score, and skills course participation) affecting student retention at an academic institution. Literature also indicated parental education as a critical driver of student retention, as first-generation students are less likely to persist [24], [29]. Further, a chi-square test of the independence and logit models was employed to measure the success of course effectiveness on student performance [20]. Additionally, Barbatis [28] adopted a constant comparative method to examine the effectiveness of student engagement on student persistence and retention at a large urban academic institution. However, existing retention models that investigated retention behavior have limitations (e.g., failing to include the importance score of predictor variables, needing help finding effective DMT to boost accuracy, and improving model performance). The authors of this paper feel that all of these theories are impactful when you look at the whole individual and the whole student.

Academic communities, including higher learning institutions, continue to confront severe challenges in identifying metrics and implementing effective measures to retain more students [19]. In our research, only a handful of studies discussed intervention programs to mitigate student dropout rates and impact retention [2]. The importance of student retention numbers, the complexity, and the uncertainty of predicting them due to pandemics are well-acknowledged in the literature [1]. Federal and state governments (FSG) have understood the value of student success in building a strong economy. As a result, FSG offers multiple programs to all layers of students across higher academic institutions to enhance student success. Moreover, higher educators and policymakers are genuinely interested in examining the underlying cause of student dropout and continuation and work to impact and change it.

RESEARCH HIGHLIGHT & CONTRIBUTION

This study contributes to the literature in the following ways:

- We deployed an efficient data mining technique to predict student retention behavior accurately.
- We diagnosed the major risk factors, including the significant features of student retention and the importance score of key predictor variables explored from the dataset.
- We conducted a comparative analysis of tree-based, regression-based, and statistical-based models using SMOTE algorithm to predict student retention behavior.

We organized the structure of this research as follows. First, we explain our conceptual framework, followed by a literature review. Second, we present the sample data and variables. Third, we describe the empirical models followed by findings and research contributions. We then discuss concluding remarks and directions for future research.

II. RELATED WORK

The application of data mining techniques, including machine learning to examine student data is evident in recent literature [8], [5], [31]. They also indicated the superiority of DMTs (e.g., neural network, SVM, DT) to statistical models (e.g., logistic regression, discriminant analysis) to predict results. These methods also better handle quantitative and qualitative (categorical data) [36]. For example, Machine learning algorithms (e.g., DT, KNN, LR, naïve bayes, RF, SVM) have been employed and reported an accuracy of more than 80% to predict student retention for the first three years of study in a Chilean case study [5]. Further, Yu *et al.* [6] study reported the application of DMTs to investigate the factors (e.g., transferred hours, residency, and ethnicity) impacting university student retention. Similarly, Other factors (e.g., average mark, course award, race) can predict student retention behavior [44].

This study examines student retention behavior by deploying six different data mining techniques. We also deploy evaluation metrics (e.g., Accuracy, Sensitivity, Specificity, Precision, F1-score, and Kappa) to assess model performance in predicting retention. Furthermore, we include additional key indicator variables, comparing results among DMT and whether the findings are supportive across the literature. Including crucial factors and proposed machine learning techniques might bring better ways to investigate the shifting trends of retention and accurately identify factors impacting student retention. This methodology of selecting predictor variables to predict retention is consistent with existing retention theories. Key findings reported here will provide institutions' personnel with actionable information and best management practices to assess retention predictors. Additionally, using a combination of data (e.g., student engagement, student performance, and student demographics) for retention prediction will help the institution adopt targeted intervention measures to establish knowledge-based and data-valued decision-making. Similar institutions across higher academia can share findings reported here to define factors affecting retention and allocate resources efficiently across the institution.

III. MATERIALS AND METHODS

A. DATASET AND VARIABLE DESCRIPTION

We collect available student data from a 4-year flagship institution (Colorado state university- Fort Collins) in Colorado. The cross-sectional dataset includes student records of 18,000+ students covering the academic year 2019-2021.

TABLE 1. Dataset Specifications

Sr No	Variables	Descriptions	Units
1	retained	The indicator for persistence is EITHER still enrolled at the institution or has graduated.	Y = 1; N=0
2	KEY	Indicator for Participation in Key Learning Community during the first fall	Y = 1; N=0
3	C4E	Indicator for Participation in Community for Excellence during the first fall	Y =1; N=0
4	AAC	Indicator for Participation in Academic Advancement Center during the first fall	Y = 1; N=0
5	Honors	Indicator for whether a student belongs to the University Honors Program during the first fall	Y = 1; N=0
6	Athlete	Indicator for whether a student is an athlete during the first fall	Y = 1; N=0
7	first_gen	Indicator for whether a student is a First-Generation student during the first fall	Y = 1; N=0
8	gpa_cal	Semester grade point average during the first fall	Number
9	age	Age as of census during the first fall	Years
10	Attempted_credits	Total attempted credits during the first fall	Number
11	Completed_credits	Total completed credits during the first fall	Number
12	CO_resident	Indicator for Colorado resident	Y = 1; N=0
13	Grants_scholarships	The total amount of grants and scholarships paid during the first year	Dollar
14	Loans	The total amount of loans paid during the first year	Dollar
15	Gender_female	An indicator of a student is female during the first fall	0 = Male 1= Female
16	Race_white	An indicator of a student is a White	Y = 1; N=0
17	Race_black	An indicator of a student is Black	Y = 1; N=0
18	Race_native_american	An indicator of a student is Native American	Y = 1; N=0

The dataset includes both first-time full-time/part-time degree-seeking students. The primary data sources are the national center for education statistics (NCES), the student unit record data system (SURDS), and the integrated postsecondary education data system (IPEDS). The dependent variable includes student retention behavior (See Table 1).

Explanatory or predictor variables include student demographics (e.g., age, race, first-generation, parental education, residency status, and gender), institution characteristics (e.g., campus location, academic advisor, and primary program indicator), academic achievement, student status, registered credit hours, earned credit hours, grade point average (GPA), enrolled in student support services program (trio advisor), and use of financial aid (see Table 1). The above-mentioned independent variables may influence student retention behavior. All model variables are explained and reported in Table 1.

B. EMPIRICAL ESTIMATION STRATEGIES

This study uses data mining techniques to predict student retention behavior accurately. For example, we employ six different machine learning (ML) models (e.g., Logistic Regression (LR), Support Vector Machine (Linear Classifier), Support Vector Machine (Radial Basis Function), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN) to test the accuracy of each technique in predicting student retention. Also, we employ Synthetic Minority Oversampling Technique (SMOTE) algorithm to make the data relatively balanced between classes. Here, we also train the model on the highest-ranked features indicated by the RF method. As mentioned in the literature, these DMT performs well by decreasing the error in prediction and actual outcomes when there is a presence of non-linear features and complex interactions among predictor variables [15].

The DMT splits data into training and validation subgroups. In this study, we used 80 percent of our observations to train the model and the rest 20 percent for model validation. The random forest method measures the importance of predictor variables in predicting the response variable [9], [23]. For each observation, this method gives two types of out-of-bag predictions: (i) obtained from actual data and (ii) obtained from variable k permuted data. The difference between error rates of predictions is computed for each observation and averaged over all observations [7].

IV. ANALYSIS AND DISCUSSION OF RESULTS

Table 2 shows the descriptive statistics of model variables. As shown by Table 2, students' mean college entrance age was 18.90, with a standard deviation of 2.44. Also, the mean GPA of students was 3.04, with a standard deviation of 0.89. Further, an undergraduate student registered (attempted) 13.92 credits (on average) in their first semester. However, a student completed 13.13 credits which is lower than the registered credits. Sixty-four percent of the total student population are residents of Colorado (mean value of 0.64) while attending college. In other words, 36 percent of students are not Colorado residents.

TABLE 2. Descriptive Statistics

Variables	Mean	S.D.*
KEY	0.084	0.277
C4E	0.071	0.258
AAC	0.018	0.134
Honors	0.076	0.266
Athlete	0.016	0.128
first_gen	0.241	0.427
gpa_cal	3.049	0.892
age	18.900	2.440
Attempted_credits	13.927	2.252
Completed_credits	13.135	3.446
CO_resident	0.645	0.478
Grants_scholarships	5618.871	4395.629
Loans	7375.597	6160.766
Gender_female	0.560	0.496
Race_white	0.890	0.311
Race_black	0.045	0.208
Race_native_american	0.035	0.185

*S.D. refers to Standard deviation

The total number of female students is higher (56 percent) than another gender at this institution. White is this institution's dominant race (89 percent) compared to other ethnic groups. Also, 1 out of 4 (24 percent) students declared themselves first-generation here. The majority of students have registered for classes at the primary campus. Finally, most students declare their primary major while attending the institution and are also eligible to receive financial aid. Students received an average of $5618.87 through grants and scholarships. Similarly, students also received an average of $7375.59 as a loan while studying at this institution. We excluded these variables (e.g., grants_scholarships and loans) from the ML analysis due to more missing values.

This study employs six classification models: Logistic Regression (LR), Support Vector Machine (SVM) Linear Classifier, Support Vector Machine (SVM) Radial Basis Function Classifier, Decision Tree Classifier (DTC), Random Forest (RF) Classifier, and k-nearest neighbors (KNN) Classifier. We also employ supervised machine learning models to assess better the performance of all six models reported here. We divided the dataset into an 80:20 ratio (80% of the observations to train the model and 20% to test the model). This ratio is selected to set the classification tasks and resolve the overfitting issue. We also reported several evaluation metrics to test the performance of the machine learning classifiers. Python environment (V 3.11.2) uses different libraries on a Windows 11 Enterprise 64-Bit, Intel core i7, and 32 GB of RAM to evaluate the ML algorithms.

SYNTHETICMINORITY OVERSAMPLING TECHNIQUE (SMOTE)

SMOTE can be defined as an intelligent alternative to oversampling and act as a solution when there is data imbalanced problem. This technique uses Euclidean distance to help increase the data by creating synthetic data points of the minority class from its nearest neighbor (Ishaq et al., 2021). It also avoids data duplication of the minority class and help creates synthetic data points that are relatively like the original ones. Here, we created a new training dataset using this technique. We fixed the data imbalance problem using SMOTE (Figure 1).

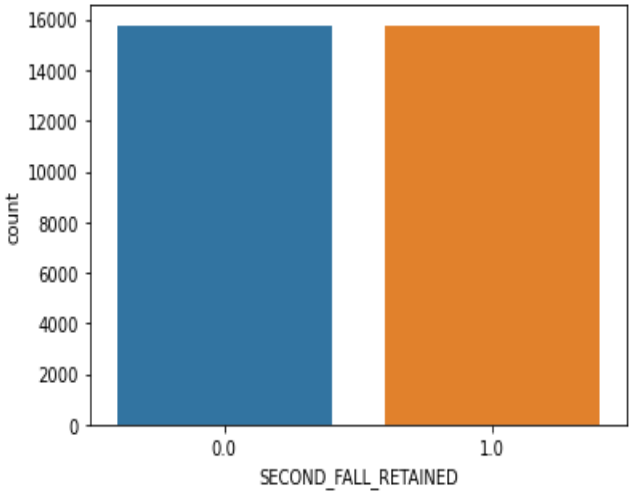


Figure 1. Balanced data using SMOTE.

This study examines six metrics (Accuracy, Sensitivity, Specificity, Precision, F1-score, and Kappa) to present the difference among machine learning-based algorithms. The existing literature supports the application of these metrics to evaluate ML-based algorithms [35]. The confusion matrix is also employed here to calculate all six metrics. Here, the confusion matrix elements are true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Performance evaluation measures are presented in Table 3.

TABLE 3. Performance Evaluation Measures

Name	Formula
Accuracy	Number of correctly classified predictions/ total predictions
Sensitivity/Recall	TP/(TP+FN)
Specificity	TN/(TN+FP)
Precision	TP/(TP+FP)
F1-score	2* (precision. recall/precision+recall)

Kappa	1-(1-observed agreement)/(1-expected agreement)

We present the pattern of the dataset through the form of data visualization. We can better explain student retention behavior by visualizing the factors impacting retention. We employed the random forest classifier to explain feature ranking.

The importance of the feature, as predicted by the RF, is displayed in Figure 2 and Figure 3.

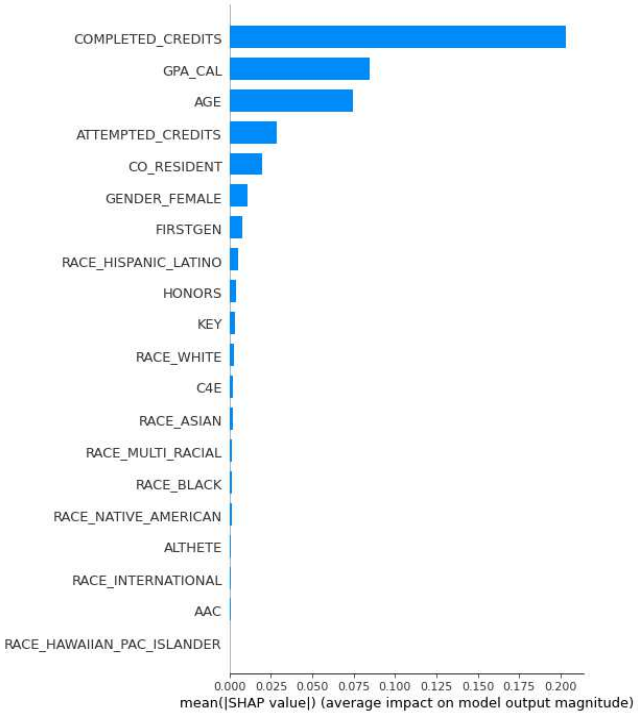


Figure 2. Factors importance of student retention behavior.

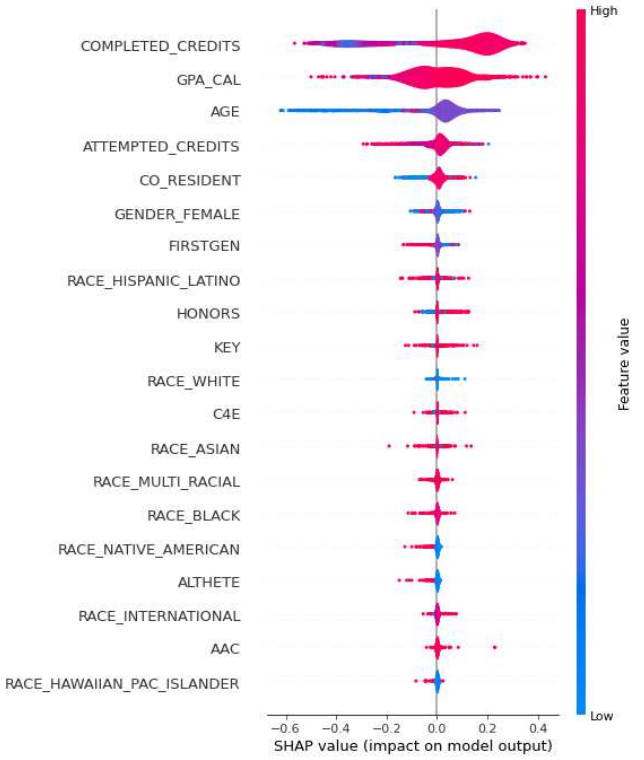


Figure 3. Feature ranking by random forest classifier.

The highly ranked factors that explain student retention are completed credits, GPA, and age, as identified by the RF.

The performance evaluation (testing and training accuracy) of machine learning models on the complete set of variables impacting retention is reported in Table 4.

TABLE 4. Data mining algorithms (testing and training accuracy)

Models	Training accuracy	Testing accuracy
LR	0.687	0.685
SVM(LC)	0.685	0.678
SVM(RBF)	0.748	0.723
DT	0.980	0.838
RF	0.957	0.861
KNN	0.823	0.763

Regarding training accuracy, DT (0.98) and RF (0.95) methods have the best performance compared to other DMTs. LR (0.68) and SVM-LC (0.68) also have the lowest accuracy in the training phase. However, RF has a higher accuracy value (0.86) as compared to DT (accuracy value of 0.83) in the testing phase. As expected, LR (0.68) and SVM-LC (0.67) also have the lowest accuracy value in the testing phase.

FigFig

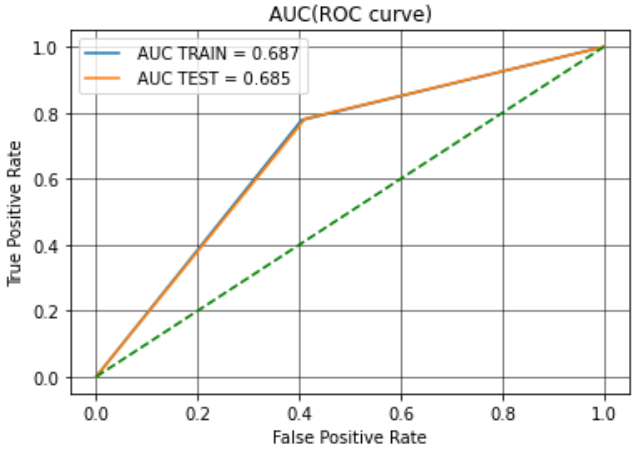


Figure 4. Logistic Regression accuracy (ROC curve)

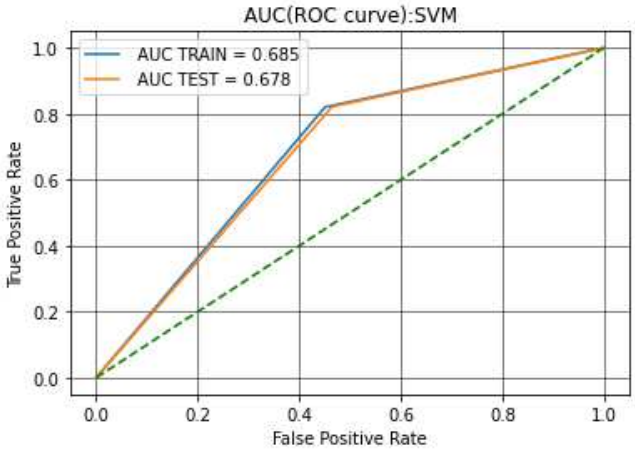


Figure 5. Support vector machine accuracy-LC (ROC curve)

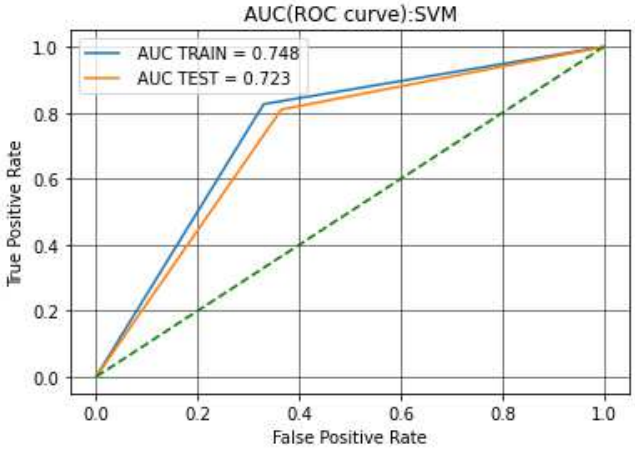


Figure 6. Support vector machine accuracy-RBF (ROC curve)

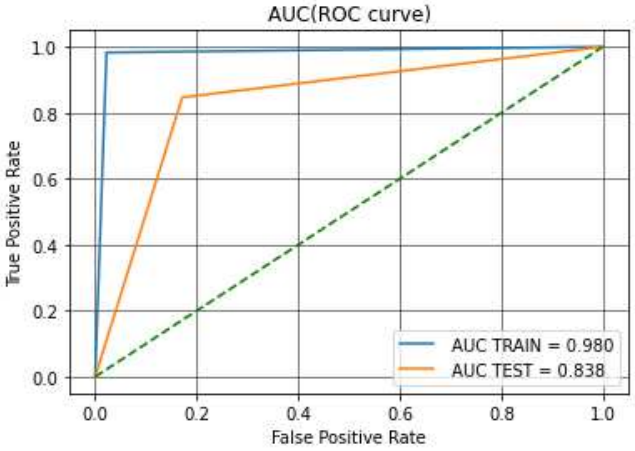


Figure 7. Decision Tree accuracy (ROC curve)

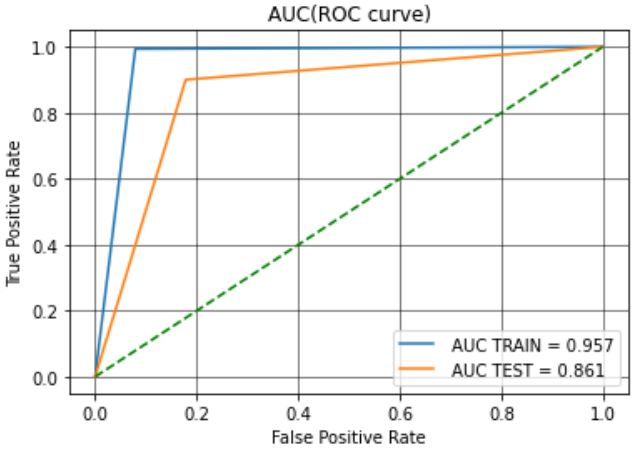


Figure 8. Random forest accuracy (ROC curve)

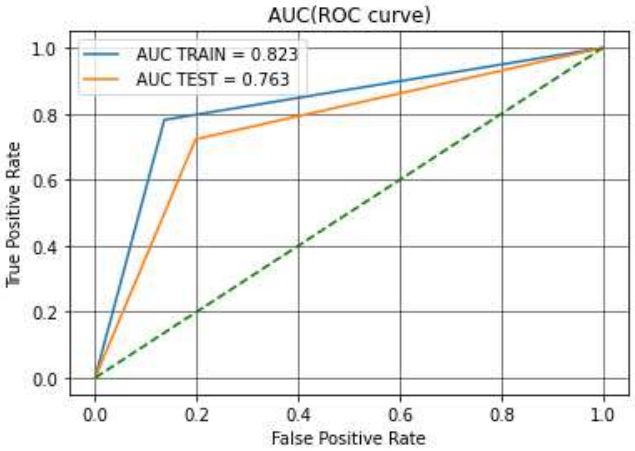


Figure 9. K-nearest neighbor accuracy (ROC curve)

LR models have achieved the highest accuracy (0.86) compared to other DMTs' when run without SMOTE (dataset

with classification problem), as reported in Table 5 and Figure 4-10.

TABLE 5. Classification result without SMOTE

Models	Accuracy*	Sensitivity	Specificity	Precision	F1-score	Kappa
LR	0.867	0.98	0.21	0.21	0.35	0.27
SVM (LC)	0.584	0.99	0.18	0.18	0.30	0.24
SVM(RBF)	0.573	0.99	0.16	0.16	0.27	0.21
DT	0.594	0.87	0.32	0.32	0.47	0.18
RF	0.589	0.96	0.22	0.22	0.35	0.23
KNN	0.570	0.98	0.16	0.16	0.28	0.19

However, with SMOTE (classification problem is solved), DMTs' performed way better than the LR model (table 6).

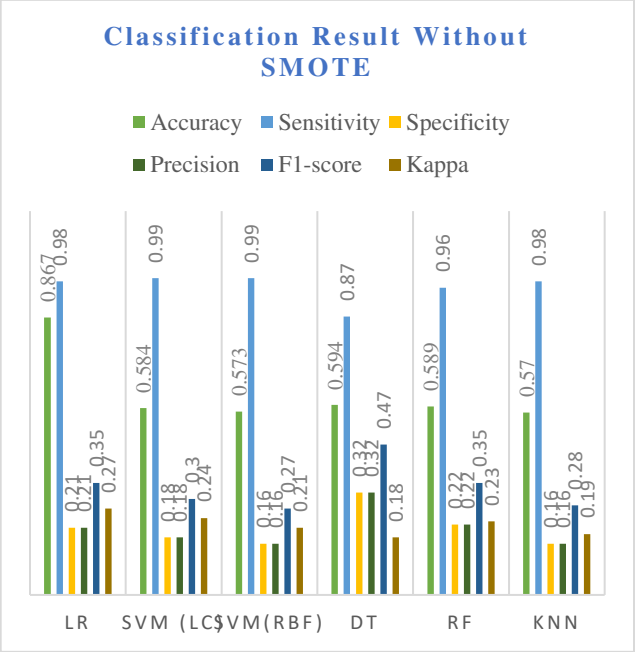


Figure 10. Results without SMOTE.

TABLE 6. Classification result with SMOTE

Models	Accuracy*	Sensitivity	Specificity	Precision	F1-score	Kappa
LR	0.685	0.78	0.59	0.59	0.67	0.37
SVM (LC)	0.678	0.82	0.54	0.54	0.65	0.35
SVM(RBF)	0.723	0.81	0.64	0.64	0.71	0.44
DT	0.838	0.85	0.83	0.83	0.84	0.67
RF	0.861	0.90	0.82	0.82	0.86	0.72
KNN	0.763	0.72	0.80	0.80	0.76	0.52

According to Table 6 and Figure 11, the RF classifier has achieved the highest accuracy with a value of 0.86 compared to other ML models.

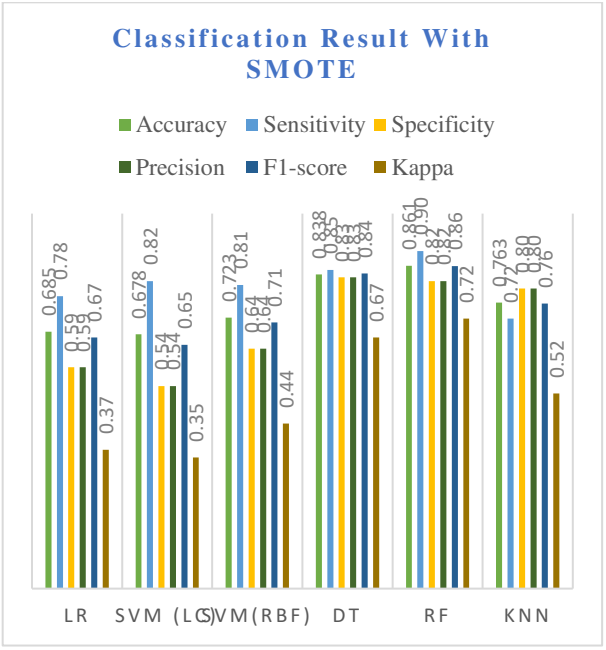


Figure 11. Results with SMOTE.

Also, RF classifier model has the highest value of 0.90 sensitivity, 0.86 F1-score, and 0.72 kappa value. DT model has the second-best accuracy value of 0.83, 0.84 F1-score, and 0.67 of kappa and the highest value with a 0.83 of specificity and precision. However, the regression-based (LR= 0.68) and statistical-based (SVM-LC = 0.67, SVM-RBF = 0.72) models have the lowest prediction performance as compared to other data mining models presented here (table 7 and Figure 12).

TABLE 7. Accuracy comparison

Models	Accuracy without SMOTE	Accuracy with SMOTE
LR	0.87	0.685
SVM (LC)	0.58	0.678
SVM(RBF)	0.57	0.723
DT	0.59	0.838
RF	0.59	0.861
KNN	0.57	0.763

Thus, RF and DT models have the most robust prediction performance in higher education student retention prediction.

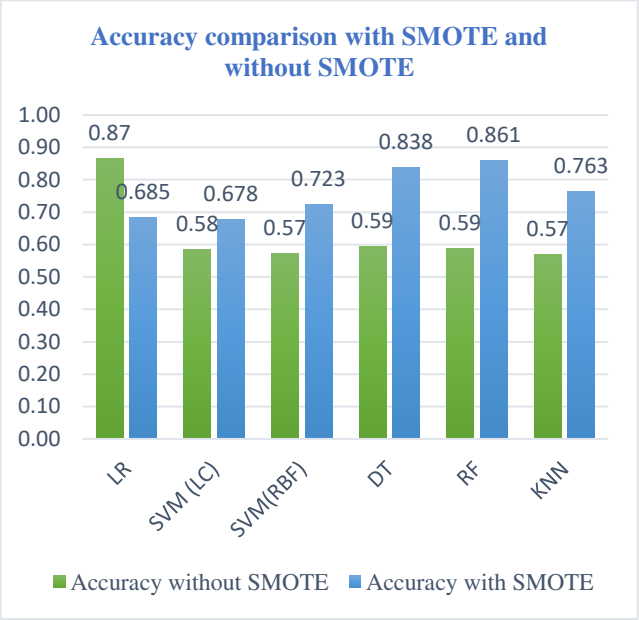


Figure 12. Accuracy comparison.

Key results indicated that supervised ML models efficiently predict student retention scenarios in higher education. Also, tree-based algorithms stated better accuracy on the imbalanced dataset with SMOTE techniques. As per Figure 3, the RF method indicated completed credits, GPA, age, and attempted credits as crucial variables to predict student retention. Additionally, the DT model outperformed all other ML models when SMOTE techniques were employed to balance data. We also presented Heatmaps (Figure 13) to show the density or intensity of retention variables, display patterns, variance, and even anomalies.

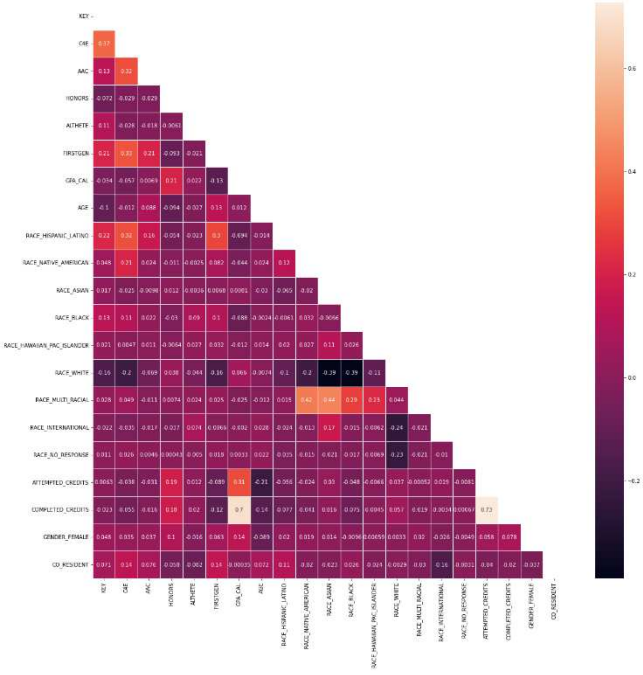


Figure 13. Heatmaps (correlation matrix).

The Heatmaps grid (correlation matrix) displayed the correlation coefficients among student retention variables. The entire distribution of student retention data is represented by the violin plot (figure 14).

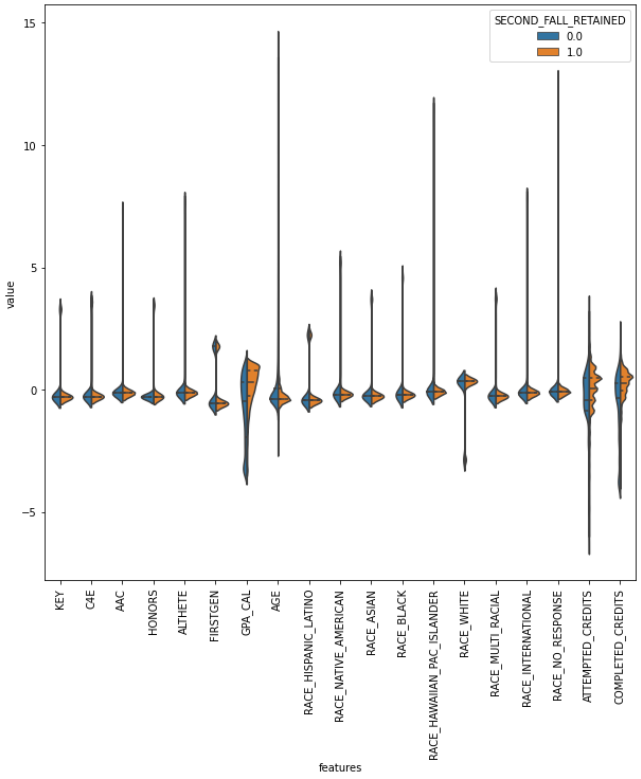


Figure 13. Violin plot.

V. CONCLUSION

ML models have gained popularity in recent years due to their power to understand the complexity and handle the non-linear interaction among predictor variables. The importance of model accuracy in predicting and discussing interpretability is evident in the recently published literature [34]. From an academic institution's perspective, there are significant benefits to accurately predicting retention, and how they can understand how ML models use input features to make predictions. We employed ML techniques to predict retention better, including LR, SVM (LC), SVM (RBF), DT, RF, and KNN. As per the data mining and ML algorithms, predicting student retention was possible with an accuracy that exceeded 80% in most cases. This study helped design educational data mining alternatives for decision-makers to improve retention based on a case study in the U.S. Institutions can adopt best management practices by analyzing factors contributing to student retention. According to RF, the top factors impacting student retention are completed credits, GPA, and age. By allocating proper support, identifying at-risk students, and

teaching methods, an institution can achieve a high retention rate and share stories of better student success.

Key results indicated that LR models had the highest accuracy (0.86) compared to other DMTs when run without SMOTE. However, with SMOTE, DMTs' model prediction accuracy increased compared to the LR model. For example, the RF classifier has achieved the highest accuracy with a value of 0.86, followed by DT (accuracy value of 0.83). In terms of sensitivity (0.90), F1-score (0.86), and kappa (0.72) value, the RF model performance is the best as compared to others. In terms of specificity (0.83) and precision (0.83) metrics, the DT model performed best. Regarding model accuracy in predicting retention behavior, DT has the second-best value (0.83). However, the regression-based models have the lowest accuracy (LR= 0.68) in predicting student retention, followed by support vector machine models (SVM-LC = 0.67, SVM-RBF = 0.72) compared to other data mining models. Thus, RF and DT models have achieved the highest accuracy in higher education student retention prediction. We also suggested that the DMTs algorithms work best for retention prediction. This study can improve university student retention by analyzing contributing factors. Future research should gear toward combining multiple machine-learning models and also include more diverse datasets across institutions. Future research would also examine how to impact completed credits, GPA rates, and how to support students who enter at an older age (32-41) in a different way to help aid their success and retention. It would be worthwhile to see if the student retention factors stay the same or differed if we include fully online and hybrid students as compared to traditional students taking in-person classes.

Summary of requirements

Declarations:

- Ethics approval and consent to participate

This study has been approved by the Colorado State University Institutional Review Board (IRB) SBER Fort Collins on Monday, July 18th 2022. CSU IRB committee's reference number is # (FWA0000647).

- Consent for publication

"Not applicable".

- Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due to Family Educational Rights and Privacy Act (FERPA) concern but are available from the corresponding author on reasonable request.

- Competing interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

- Funding

The authors did not receive support from any organization for the submitted work.

- Authors' contributions

Rezwanul Parvez wrote the main manuscript text and prepared all the tables mentioned in the manuscript. Alysha Tarantino contributed valuable feedback in writing the manuscript text. Syed Imran Ali Meerza produced all the figures mentioned in the manuscript. All three authors have reviewed the manuscript multiple times.

- Acknowledgements

We would like to acknowledge the contribution of the Office of Institutional Research, Planning and Effectiveness at Colorado state university for helping with the data reported in this study.

- Authors' information (optional)

Rezwanul Parvez is currently working as a Researcher with the Office of Engagement and Extension at Colorado State University (CSU)-Fort Collins. Dr. Parvez has 10+ years of experience working at multiple Institutions as a lead Researcher where he managed multiple projects of the Colorado Department of Higher Education (CDHE), United States Department of Agriculture (USDA), National Resource Inventory (NRI), Environment and Natural Resources Trust Fund (ENRTF), in interdisciplinary settings.

Alysha Tarantino has been in distance/online learning for over 16 years. She currently is the Director of College Relations and Advising with Colorado State University Online and supports the work of Enrollment Services, Course Scheduling, College Relations and Program Strategy (program creation, relationship management, etc.), Dual Enrollment, Industry Relations, Research and Market Analysis, and Academic Advising. She has a bachelor's in Biology and Elementary Education and an M.Ed. in Organizational Learning Performance and Change.

Syed Imran Ali Meerza is currently working as an Assistant Professor at the Department of Agriculture, Arkansas Tech University (ATU). Before joining ATU, he also worked as a Postdoctoral Research Associate in the Department of Agricultural Economics at the University of Nebraska-Lincoln.

REFERENCES

1. Burke, "Student retention models in higher education: A literature review," *Coll. and Univ.,* vol. 94, no. 2, pp. 12–21, 2019.
2. Seidman, "Where we go from here: a retention formula for student success" in *College student retention: Formula for student success,* Rowman & Littlefield Publishers / Amer. Council. Ed., 2005, ch. 11, pp. 295-307.
3. W. Astin, and A. L. Antonio, "Analyzing assessment data" in *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education,* 2nd ed. New York: Macmillan, Rowman & Littlefield Publishers / Amer. Council. Ed., 2012, ch. 6, pp. 101-138.
4. W. Astin, "What matters in college? Four critical years revisited," *The Canad. J. High. Educ.*, Vol. 23, no. 1, pp.133, 1993.
5. A. Palacios, J. A. Reyes-Suarez, L. A. Bearzotti, V. Leiva, and C. Marchant, "Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile," *Entropy.,* vol. 23, no. 485, 2021. https://doi.org/10.3390/e23040485
6. S. Yu, S. DiGangi, A. Jannasch-Pennell, C. Kaprolet. "A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year," *Journal of Data Science.*, Vol. 8, no. 2, pp. 307-325, 2010. doi:10.6339/JDS.2010.08(2).574.
7. Zhang, and Y. Ma, "Random Forests" in *Ensemble machine learning: Methods and Applications*, 1st ed. Springer, Springer New York, Dordrecht Heidelberg London., 2012, pp. 157-175. https://doi.org/10.1007/978-1-4419-9326-7_5
8. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review," *IEEE Access.,* 2022.
9. D. D. Muchilnski, J. H. Siroky, and M. Kocher, "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data," *Political Analysis.,* vol. 24, no. 1, pp. 87-103, 2015.
10. D. F. Allen, and B. Bir, "Academic confidence and summer bridge learning communities: Path analytic linkages to student persistence," *J. Coll. Stu. Ret: Res. Theo. & Prac.,* Vol. 13, no. 4, pp. 519-548, 2012.
11. D. Fike, and R. Fike, "Predictors of first-year student retention in the community college," *Comm. Coll. Rev.,* vol. 36, no. 2, pp. 68-88, 2008.
12. D. Hossler, and J. P. Bean, "The Impact of Student Aid on Recruitment and Retention: What the Research Indicates" in *The Strategic Management of College Enrollments,* San Francisco: Jossey-Bass, 1990, pp. 61-75.
13. D. Raju, and R. Schumacker, "Exploring student characteristics of retention that lead to graduation in higher education using data mining models," *J. Coll. Stu. Ret.*, vol. 16, no. 4, pp. 563-591, 2015.
14. Aspegren, "These colleges survived World Wars, the Spanish flu and more. They couldn't withstand COVID-19 pandemic" *USA Today*, 2021, Retrieved January 12, 2023, from https://www.usatoday.com/story/news/education/2021/01/28/covid-19-colleges-concordia-new-york-education/4302980001/.
15. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," in *IEEE Access.*, vol. 9, pp. 39707-39716, 2021, doi: 10.1109/ACCESS.2021.3064084.
16. J. M. Scott, and R. E. Hoover, "Predictors of First-Year Retention in a Community College," *Comm. Coll. J. Res. and Prac.*, vol. 38, no. 7, pp. 651-660, 2014.
17. J. P. Bean, "Dropouts and turnover: The synthesis and test of a causal model of student attrition," *Res. High. Educ.*, vol. 12, pp. 155-187, 1980.
18. J. P. Bean, "Why students leave: Insights from research." In:D. Hossler, and J. P. Bean (eds.), The Strategic Management of College Enrollments. Jossey-Bass: San Francisco, 1990.
19. K. Hone, and S. Ghada, "Exploring the factors affecting MOOC retention: A survey Study," *J. Comput. & Educ.*, vol. 98, pp. 157-168, 2016.
20. K. Kimbark, M. L. Peters, and T. Richardson, "Effectiveness of the Student Success Course on Persistence, Retention, Academic Achievement, and Student Engagement," *Comm. Coll. J. Res. and Prac.*, vol. 41, no. 2, pp. 124-138, 2017.
21. L. K. Lau, "Institutional factors affecting student retention," *Education.,* vol. 124, no. 1, pp. 126-136, 2003.
22. L. Breiman, "Random forests," *Machine Learning.,* vol. 45, no. 1, pp. 5-32, 2001.
23. M. D. Summers, "Attrition research at community colleges," *Comm. Coll. Rev*., vol. 30, no. 4, pp. 64-84, 2003.
24. M. S. DeBerrad, G. I. Spielmans, and D. C. Julka, "Predictors of academic achievement and retention among college freshmen: A longitudinal study," *Coll. Stu. J.,* vol. 38, no. 1, pp. 66-80, 2004.
25. M. H. Windham, M. C. Rehfuss, C. R. Williams, J. V. Pugh, and L. Tincher-Ladner, "Retention of first-year community college students," *Comm. Coll. J. Res. and Prac.,* vol. 38, no. 5, pp. 466-477, 2014.
26. National Center for Education Statistics. "Undergraduate Retention and Graduation Rates. Condition of Education. U.S. Department of Education, Institute of Education Sciences" Retrieved Jan 21, 2023, from https://nces.ed.gov/programs/coe/indicator/ctr.
27. P. Barbatis, "Underprepared, ethnically diverse community college students: factors contributing to persistence," *J. Dev. Educ*., vol. 33, no. 3 pp. 16-26, 2010.
28. P. Thayer, "Retention of students from first generation and low-income backgrounds," *Opportu. Out. J.,* vol. 3, no. 1, pp. 2-8, 2000.
29. R. D. Reason, "An examination of persistence research through the lens of a comprehensive conceptual framework," *J. Coll. Stu. Dev.*, vol. 50, no. 6, pp. 659-682, 2009.
30. S. Gupta, and A.S. Sabitha, "Deciphering the attributes of student retention in massive open online courses using data mining techniques," *Educ. Inf. Technol.,* Vol. 24, pp. 1973–1994. https://doi.org/10.1007/s10639-018-9829-9, 2019.
31. S. Krishnamurthy, "The future of business education: A commentary in the shadow of the Covid-19 pandemic," *J. Bus. Res.,* vol. 117, pp.1-5, 2020. https://doi.org/10.
32. S. Kunal, W. Wamwara, K. Goodrich, S. Schiller, and J. Dinsmore, "Understanding Business Student Retention During Covid-19: Roles of Service Quality, College Brand, and Academic Satisfaction, and Stress," *Serv. Mark. Quart.*, 2021. doi: 10.1080/15332969.2021.19935591016/j.jbusres.2020.05.034
33. S.M. Lundberg, G. Erion, H. Chen, et al. "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.,* Vol. 2, pp. 56–67, 2020. https://doi.org/10.1038/s42256-019-0138-9
34. S. I. A. Meerza, A. Yiannaka, K.R. Brooks, C.R. Gustafson, "Information avoidance behavior: Does ignorance keep us uninformed?" In Proceedings of the Annual Meeting—Agricultural and Applied Economics Association, Atlanta, Georgia, 21–23 July 2019. Available at https://ageconsearch.umn.edu/record/290757
35. S. Trivedi, "Improving students' retention using machine learning: Impacts and implications," *ScienceOpen Preprints.,* 2022.
36. T. T. Ishitani, and S. L. DesJardins, "A longitudinal investigation of dropouts from college in the United States," *J. Coll. Stu. Ret.*, vol. 4, no. 2, pp. 173-202, 2002.
37. V. Govindarajan, and A. Srivastava, "What the shift to virtual learning could mean for the future of higher ed" *Harv. Bus. Rev.*, March 31, https://hbr.org/2020/03/what-the-shift-to-

virtual-learning-could-mean-for-the-future-of-higher-ed, Accessed on January 15, 2023.
38. W. D. Berg, E. A. Fisk, P. S. and M. G. Hanson, “A longitudinal cohort study of student motivational factors related to academic success and retention using the college student inventory,” *J. Coll. Stu. Ret.: Res., Theo. and Prac.,* vol. 17, no. 3, pp. 278-302, 2015.
39. V. Tinto, “Limits of theory and practice in student attrition,” *J. High. Edu.*, vol. 53, pp. 687-700, 1982.
40. V. Tinto, “A theory of individual departure from institutions of higher education” in *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). Chicago, IL: University of Chicago Press. 2012.
41. V. Tinto, "Taking student retention seriously: Rethinking the first year of university" In ALTC FYE Curriculum Design Symposium, vol. 5, 2009.
42. Y. Zhang, S. Oussena, T. Clark, and K. Hyensook, “Using data mining to improve student retention in HE: a case study,” In: ICEIS - 12th International Conference on Enterprise Information Systems, 2010, 8-12 June, Portugal.



Rezwanul Parvez is currently working as a Researcher with the Office of Engagement and Extension at Colorado State University (CSU)-Fort Collins. Dr. Parvez has 10+ years of experience working at multiple Institutions as a lead Researcher where he managed multiple projects of the Colorado Department of Higher Education (CDHE), United States Department of Agriculture (USDA), National Resource Inventory (NRI), Environment and Natural Resources Trust Fund (ENRTF), in interdisciplinary settings. He successfully completed a "professional data analytics certification" from GOOGLE in 2022. As a Researcher, Dr. Parvez led in preparing reports for state agencies of Colorado, managed numerous research publications, and spoke about economic issues at multiple conferences during 2011-2023. Academically, he completed Post-doctoral Research work at Oklahoma State University, USA a Ph.D. in Applied Economics from North Dakota State University, USA in 2017, a Master of Science in Economics from South Dakota State University, USA in 2012, and a B.S.S in Economics from Khulna University, Bangladesh in 2007.



Alysha Tarantino has been in distance/online learning for over 16 years. She currently is the Director of College Relations and Advising with Colorado State University Online and supports the work of Enrollment Services, Course Scheduling, College Relations and Program Strategy (program creation, relationship management, etc.), Dual Enrollment, Industry Relations, Research and Market Analysis, and Academic Advising. She has a bachelor's in Biology and Elementary Education and an M.Ed. in Organizational Learning Performance and Change.



Syed Imran Ali Meerza received a B.S. degree in economics from Khulna University, Bangladesh, in 2007 and an M.S. degree in economics from South Dakota State University, SD, USA, in 2013. He earned his Ph.D. degree in agricultural economics from the University of Nebraska, USA, in 2018.
Currently, he is working as an Assistant Professor at the Department of Agriculture, Arkansas Tech University (ATU). Before joining ATU, he also worked as a Postdoctoral Research Associate in the Department of Agricultural Economics at the University of Nebraska-Lincoln. Dr. Meerza’s awards and honors include the Khulna University Gold Medal, Tomek Outstanding Ph.D. Student Award, Outstanding Service to Graduate Students Award, and Mary and Charles Cooper Fellowship.