Integrating categorical and continuous data in a cluster-then-classify methodology for predicting undergraduate student success

John Fields
Computer Science
Marquette University
Milwaukee, USA
https://orcid.org/0000-0001-5153-0376

2nd Kevin Chovanec
Computer Science
Marquette University
Milwaukee, USA
kevin.chovanec@marquette.edu

3rd Praveen Madiraju
Computer Science
Marquette University
Milwaukee, USA
praveen.madiraju@marquette.edu

***Abstract*—Student retention in higher education remains a significant challenge despite decades of research. This study introduces a novel cluster-label-classify methodology to predict at-risk students and identify common characteristics among those who drop out. Using data from a small private Midwestern university in the United States, we first applied the K-Prototypes algorithm to cluster non-retained students into five groups based on both numeric and categorical variables. We then labeled these clusters and the retained students, creating a multi-class classification problem. Finally, we used a Gradient Boosting Classifier and XGBoost for classification, achieving F1 scores of 0.82 to 0.89 for predicting non-retained students and 0.96 for retained students after addressing class imbalance with SMOTE. This approach allows for customized labeling specific to each institution and enables more targeted interventions for at-risk students. Our methodology combines demographic, academic, and socioeconomic factors to provide a comprehensive view of student retention, potentially offering new insights into this longstanding issue in higher education. The paper also discusses algorithmic bias, examining potential fairness issues in the predictive models and their implications for different student populations. Finally, a discussion on Privacy Preserving Machine Learning (PPML) provides future strategies for testing how these technologies generalize to other institutions while enhancing the privacy of student data.**

***Index Terms*—machine learning, unsupervised learning, predictive models, classification algorithms, education**

I. Introduction

Student retention in higher education has been a persistent challenge, with over 70 years of active research [1], [2]. Despite these efforts, significant improvements in retention rates remain elusive; approximately one-third of undergraduate students in the United States fail to complete their degrees [5]. In recent years, researchers have leveraged machine learning techniques to identify at-risk students [6], [7]. While each university faces unique challenges, the decision to leave is typically influenced by various factors, including academic performance, financial constraints, social integration, medical issues, and other personal circumstances [3].

This trend aligns with the emerging field of learning analytics in higher education, which focuses on the measurement, collection, analysis, and reporting of data to improve student outcomes [37]. Learning analytics provides a framework for utilizing the vast amount of data generated in educational settings to gain insights into student behavior, performance, and potential risks.

This study introduces a novel approach that combines clustering and classification algorithms to predict at-risk students. We first employed a clustering algorithm to identify common characteristics among students who dropped out, then utilized a classification algorithm to predict at-risk students. This method yielded F1 scores from .82 to .89 in a multi-class classification for non-retained students. The key advantage of this approach is its adaptability; labeling can be customized to each institution's specific context, allowing for more targeted interventions to at-risk students.

II. Related Works

The seminal work of Tinto [4] in 1975 identified key factors contributing to student dropout, including family background, individual attributes, pre-college schooling, academic integration, social integration, goals, and commitment. Subsequent research in the 1980's expanded on the theoretical frameworks explaining student attrition.

Astin's Theory of Student Involvement comprises three elements: inputs, environment, and outputs. It posits that "student involvement refers to the amount of physical and psychological energy that the student devotes to the academic experience" [13]. Bean's Causal Model of Student Attrition identified institutional commitment, institutional quality, routinization, satisfaction, and communication [14].

Kuh [8] developed the National Survey of Student Engagement (NSSE) in 2000 to assess undergraduate education quality. Their findings emphasized that deeper learning and skill development in areas like critical thinking, problem solving, and effective communication are strongly correlated with the extent to which students study, practice, and receive feedback.

Duckworth and Carlson [9] highlighted self-regulation–encompassing abilities such as self-control, delay of gratification, and persistence–as a robust predictor of various

academic outcomes. Their research suggests that these traits often outperform IQ in predicting high school graduation rates, grades, and to some extent, standardized tests scores.

The 21st century has seen the emergence of statistical methods and machine learning techniques applied to student success prediction. A 2019 systematic review by [15] analyzed 67 papers and identified 112 factors affecting dropout across five dimensions: Personal, Academic, Economic, Social, and Institutional. The study concluded that no single method consistently produces superior results, emphasizing the importance of context and data in determining the most effective approach.

Predicting success using machine learning techniques has gained traction in recent years. [27] employed a Random Forest machine learning algorithm to forecast student performance and identify those at risk of academic difficulties. Their study demonstrated the potential of these advanced analytical methods in providing early interventions and support to students who may be struggling. Similarly, [28] utilized a Random Forest algorithm to develop an early warning system for identifying at-risk students, allowing for timely interventions and improved retention rates. This approach also included clustering but this occurred after the classification where our method proposes clustering prior to classification.

A cluster-then-label methodology has been proposed by several researchers [12], [32]–[35]. These approaches have primarily found application in medical research and analysis of unlabeled images. The cluster-then-label approach is particularly valuable when labeled training data is unavailable, a common scenario in educational datasets. By first clustering the data based on inherent patterns and then assigning labels to these clusters, this method allows for the creation of a labeled dataset from initially unlabeled data, enabling subsequent classification tasks. This can be especially beneficial in higher education settings, where the complex interplay of factors affecting student retention may not be easily captured by predetermined labels. The cluster-label-classify methodology proposed for this study represents the first known application of this technique in higher education, offering a novel approach to analyzing student data and potentially uncovering previously unrecognized patterns in retention risk factors.

A significant challenge in the field of educational data mining is the generalizability of predictive models across different universities or educational contexts. [29] highlighted that many studies rely on a single dataset for analysis, underscoring the difficulties in applying models developed at one institution to another. This limitation stems from variations in student populations, academic programs, and institutional factors. Different schools may have unique factors influencing student achievement, so applying predictive models from one educational context to another should be done carefully.

This underscores the need for ongoing research and development in the field of student success prediction, with a focus on creating flexible and adaptable models that can account for the diverse factors influencing academic performance across various educational settings. In Section VI, we discuss privacy preserving machine learning (PPML) as a potential solution.

III. Methodology

This study utilized undergraduate student data from a small private university in the Midwestern United States. The dataset comprised 3109 students (1664 female, 1441 male, 4 non-responses) from the Fall 2021 cohort, with 765 students (24.6%) not retained. The dataset included 40 variables: 27 continuous and 13 categorical.

Prior to the analysis, the data was de-identified by removing the student ID, first name, and last name. It is worth noting that the university is in the process of incorporating National Student Clearinghouse data to determine whether non-retained students transferred to another U.S. institution or dropped out of higher education entirely.

It is also important to acknowledge that the COVID-19 pandemic could be a significant confounding variable in this study. The Fall 2021 cohort may have experienced unique challenges related to the pandemic, such as disrupted high school education, changes in campus life, and increased stress, which could have impacted retention rates in ways that differ from pre-pandemic cohorts.

Our cluster-label-classify methodology consisted of three main steps:

1) Clustering: We employed the K-Prototypes algorithm [10], [11] to group non-retained students (n = 765) into five clusters. The number of clusters was chosen based on previous work with another university's data where groupings of 3/5/7 clusters were analyzed. An elbow test was also conducted but these always seem to return a cluster size of two or three which was not granular enough for this test. The K-Prototypes algorithm was also chosen for its ability to handle both numeric and categorical data. Other clustering algorithms such as fuzzy K-means have shown improved results over K-Prototypes but most do not support categorical and continuous numeric variables [40].
2) Labeling: We created a new variable labeling the clusters as Class 0 through Class 4 for non-retained students and Class 5 for retained students. These labels were added to the complete dataset (n = 3109).
3) Classification: Using the XGBoost package in Python, we first performed a binary classification to predict retention on the retained/not retained variable. We then trained a model for multi-class classification to predict each of the six labels created in Step 2 using Gradient Boosting and XG Boost algorithms.

Due to class imbalance in the dataset, initial results showed poor performance for predicting non-retained students (F1 scores of 0.00 to 0.13) compared to retained students (F1 score of 0.86). To address this, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to create a more balanced dataset [38].

By addressing the class imbalance issue and acknowledging potential confounding factors like the COVID-19 pandemic, we aim to provide a more nuanced understanding of the

factors influencing student retention. This methodology has the potential to offer valuable insights for higher education institutions seeking to improve their retention strategies and support student success.

IV. Results

After applying K-Prototypes to group the non-retained students, we performed a Principal Components Analysis to determine the variables with the greatest feature importance in clustering. These results are shown in Figure 1 and descriptions of the features are in Table VI.

We then analyzed each cluster to identify its dominant characteristics and **bold-faced type** is used below in each group to highlight the dominant characteristics of each class.

A. CLASS 0

- Count = 163
- **Campus A** (100%)
- **Athlete** (88%)
- **Male** (68%)
- Pell Eligible (41%)
- Religious Affiliation of the University (11%)
- First Generation (29%)
- TERMGPA = 2.5
- Credits Attempted = 15

B. CLASS 1

- Count = 141
- **Campus B** (100%)
- **Athlete** (99%)
- **Male** (79%)
- Pell Eligible (19%)
- Religious Affiliation of the University (25%)
- First Generation (11%)
- TERMGPA = 2.7
- Credits Attempted = 15

C. CLASS 2

- Count = 123
- **Campus B** (95%)
- **Not an Athlete** (72%)
- (Female 46%, Male 54%)
- **Pell Eligible** (85%)
- Religious Affiliation of the University (6%)
- **First Generation** (62%)
- **TERMGPA = 1.0**
- **Credits Attempted** = 13

D. CLASS 3

- Count = 203
- **Campus B** (95%)
- **Not an Athlete** (99%)
- **Female** (67%)
- Pell Eligible (19%)
- **Religious Affiliation of the University** (54%)
- First Generation (10%)
- **TERMGPA = 3.0**
- Credits Attempted = 15

E. CLASS 4

- Count = 135
- **Campus A** (98%)
- **Athlete** (89%)
- **Male** (62%)
- Pell Eligible (44%)
- Religious Affiliation of the University (19%)
- First Generation (24%)
- **TERMGPA** = 2.0
- **Credits Attempted** = 13

The result of this binary classification with XG Boost was an F1 score of 0.87 for not retained students and 0.96 for retained students as shown in Table I. This is similar to the results in [15].

The results of the multi-class classification using SMOTE were significantly improved and the not retained F1 scores ranged from .82 to .89 and the retained student F1 was .96 for both methods. These multi-class results in Tables IV/V are similar to the simple binary classification described above and shown in Table I.

The significance of this lies in the contrast between binary and multi-class tasks. In binary classification, random guessing yields a 50% accuracy, while in a 6-class problem, it drops to about 16.7% (1/6). This substantial gap underscores the heightened challenge in multi-class classification for this scenario. The model faces a more demanding task, requiring a choice with a narrower margin for error.

TABLE I
Binary Classification

Student Class	Binary Classification with XG Boost		
	Precision	*Recall*	*F1 Score*
0 (not retained)	0.93	0.81	0.87
1 (retained)	0.94	0.98	0.96

TABLE II
Multi-Class Classification

Student Class	Gradient Boosting Classifier		
	Precision	*Recall*	*F1 Score*
0	0.20	0.05	0.08
1	0.00	0.00	0.00
2	0.00	0.00	0.00
3	0.33	0.08	0.13
4	0.00	0.00	0.00
5 (retained)	0.78	0.98	0.86

Here are two examples to illustrate how these targeted predictions can be used to proactively reach out to students who have a higher likelihood of dropping out:

- Class 3 has students at Campus B, not athletes, majority female, religious affiliation of the university, and high term GPA.
- Class 0 and Class 4 are similar except that Group 4 has lower grades and credits attempted. Class 0 may be leaving to play their sport at another university where

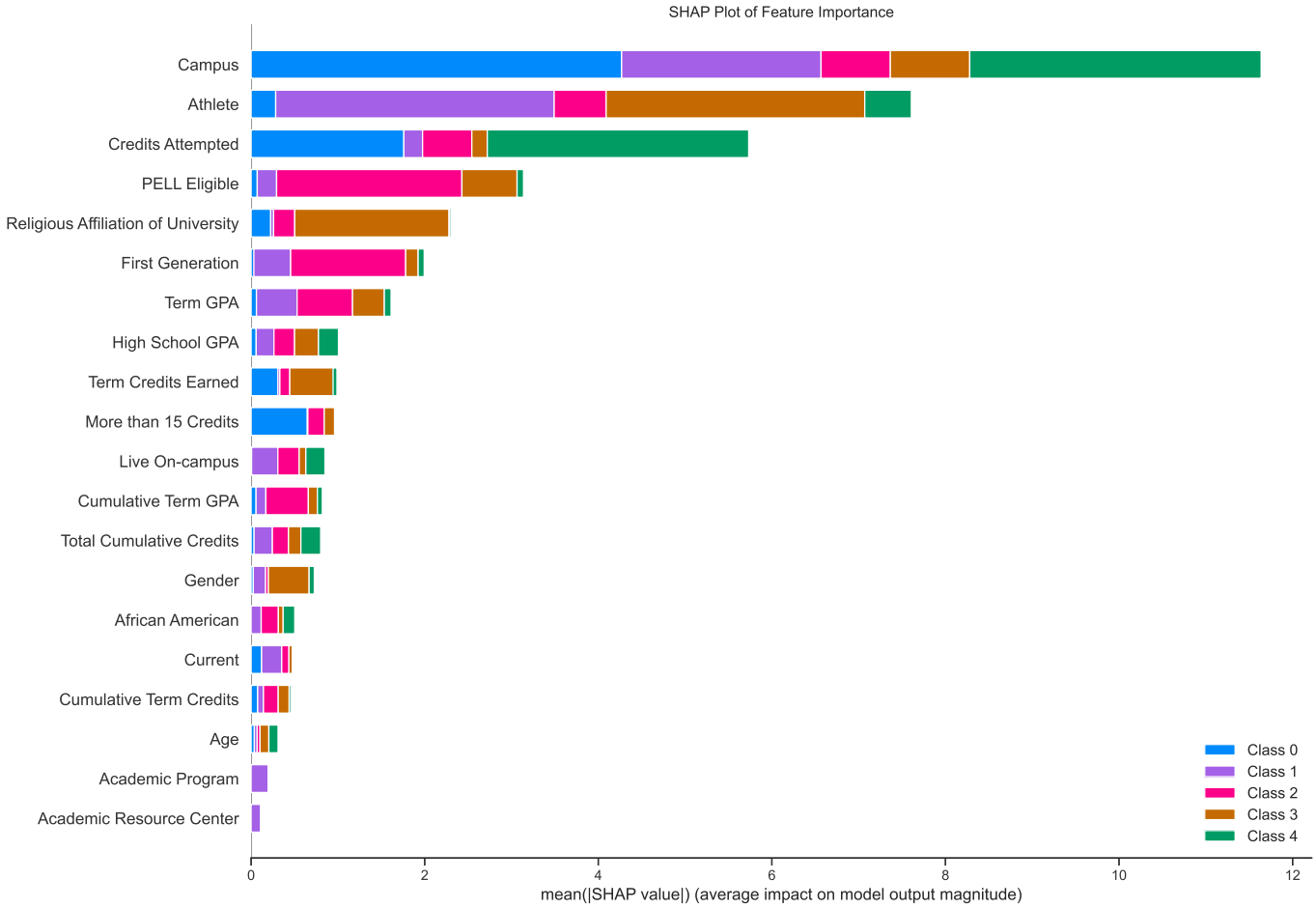

Fig. 1. SHAP Plot of Feature Importance for Clustering

TABLE III
MULTI-CLASS CLASSIFICATION

Student Class	XG Boost		
	Precision	*Recall*	*F1 Score*
0	0.20	0.05	0.08
1	0.00	0.00	0.00
2	0.33	0.07	0.11
3	0.17	0.04	0.06
4	0.00	0.00	0.00
5 (retained)	0.77	0.97	0.86

TABLE IV
MULTI-CLASS CLASSIFICATION WITH SMOTE

Student Class	Gradient Boosting Classifier		
	Precision	*Recall*	*F1 Score*
0	0.97	0.73	0.83
1	0.92	0.77	0.84
2	0.98	0.73	0.83
3	0.98	0.81	0.89
4	0.96	0.78	0.86
5 (retained)	0.93	0.99	0.96

TABLE V
MULTI-CLASS CLASSIFICATION WITH SMOTE

Student Class	XG Boost		
	precision	*Recall*	*F1 Score*
0	.93	.73	0.82
1	.91	.77	0.83
2	.98	.73	0.84
3	.94	.81	0.87
4	.98	.78	0.87
5 (retained)	.93	.99	0.96

Class 4 may be leaving due to academics. This is an example where the NSC data or a qualitative survey would be valuable to better understand the path that students take after leaving a school.

In conclusion, our analysis using K-Prototypes clustering and Principal Component Analysis has revealed distinct groups of non-retained students, each with unique characteristics. The subsequent multi-class classification using SMOTE demonstrated improved F1 scores comparable to binary classification, despite the increased complexity of the task. These findings provide valuable insights into the diverse factors

influencing student retention across different campuses and student populations. By identifying these specific clusters and their dominant characteristics, institutions can develop targeted interventions and support strategies tailored to each group's needs. This nuanced approach to predicting and addressing student retention issues has the potential to significantly improve overall retention rates and student success outcomes.

V. Algorithmic Bias

The use of machine learning models to predict student dropout in higher education also presents ethical concerns regarding potential biases. While much attention has focused on racial bias in such models, it is crucial to consider a broader range of potential biases and sensitive attributes [21], [23], [26]. Gender, religion, socioeconomic status, disability, financial status and other protected characteristics can all potentially introduce bias into predictive models. For example, a model might systematically underestimate the dropout risk for female students in STEM fields due to historical gender imbalances in those disciplines [22]. Similarly, first-generation college students or those from low-income backgrounds may face unique challenges not captured by race alone [24].

Interestingly, some recent work has suggested that including protected attributes in predictive models may actually lead to fairer and more accurate results in certain contexts. Yu et al. (2020) examined this question specifically for college dropout prediction, finding that models incorporating protected attributes like race and gender outperformed those that excluded such information [20]. The authors argue that by explicitly accounting for these factors, models can better capture the unique challenges and risk factors faced by different student populations.

These findings highlights the complex relationship between bias, fairness, and model performance in educational data mining. While the use of protected attributes in predictive modeling remains controversial, it underscores the need for nuanced approaches that carefully consider the ethical implications and potential benefits of different modeling strategies.

One such approach to addressing bias and improving fairness in predictive models is the Synthetic Minority Oversampling Technique (SMOTE). SMOTE is particularly beneficial for reducing bias against underrepresented groups, which aligns with the broader goal of considering multiple dimensions of identity in model development [25]. It works by creating synthetic examples of the minority class, effectively oversampling this class to balance the dataset. SMOTE operates by selecting a minority class instance and finding its k-nearest neighbors. It then creates new synthetic instances along the line segments joining the selected instance to its neighbors. This approach increases the representation of minority classes in the dataset, providing machine learning models with more balanced training data. By doing so, SMOTE helps mitigate the tendency of models to be biased towards the majority class, which is often a significant source of unfairness in predictive algorithms. This is particularly crucial when dealing with protected attributes or underrepresented groups, as it ensures these groups have a more equitable representation in the model's training process, potentially leading to fairer and more accurate predictions across all segments of the population. Thus, SMOTE offers a practical solution to some of the ethical concerns raised earlier regarding bias in dropout prediction models. However, it is important to note that while SMOTE addresses overall class imbalance, it may not fully resolve bias issues in sub-groups within the minority class [36].

To assess potential bias following the application of SMOTE, we employed the Python Fairlearn [39] package to evaluate the fairness of a key sensitive variable: the student's Pell grant status. This variable indicates whether a student qualifies for the U.S. Federal Pell Grant Program, which is awarded based on financial need. By analyzing this variable, we aimed to ensure that our model did not disproportionately favor or disadvantage students based on their economic background.

The outcome of the Fairlearn analysis of the Pell variable:

1) Demographic Parity Difference: 0.006508491836667685 This metric measures the difference in the probability of a positive outcome between the privileged and unprivileged groups. The value is close to zero, which suggests that there's only a small difference (about 0.65%) in the rate of positive predictions between the two groups. This indicates relatively good fairness in terms of demographic parity.
2) Equalized Odds Difference: 0.09333333333333332 This metric measures the maximum difference in true positive rates and false positive rates between the privileged and unprivileged groups. The value of about 0.093 suggests there's a 9.3% difference in either true positive rates or false positive rates between the groups. This indicates some disparity in the model's performance across groups.
3) Confusion Matrix (Privileged): [[18, 9], [1, 410]]
   - True Negatives: 18
   - False Positives: 9
   - False Negatives: 1
   - True Positives: 410
4) Confusion Matrix (Unprivileged): [[38, 12], [3, 1059]]
   - True Negatives: 38
   - False Positives: 12
   - False Negatives: 3
   - True Positives: 1059

Here is an interpretation of these results:

1) The model shows good demographic parity, with only a small difference in prediction rates between groups.
2) There's a more noticeable difference in equalized odds, indicating some disparity in the model's accuracy across groups.
3) The model seems to perform well for both groups, with high true positive rates and low false negative rates.
4) False positive rates are slightly higher for the privileged group (9 out of 27) compared to the unprivileged group

(12 out of 50).

Overall, while the model shows good fairness in terms of demographic parity, there is room for improvement in equalized odds. Additional research in this area and tests of other sensitive attributes are potential areas to explore in subsequent studies.

VI. Discussion and Future Research

While machine learning methods offer promising advancements in predicting student dropout rates in higher education, as our research demonstrates, significant challenges remain.

Future research in predicting student dropout rates should consider incorporating advanced evaluation of clustering methods such as Partition Coefficient (PC), Partition Entropy (PE), Xie-Beni (XB), and Silhouette Fuzzy (SIL.F) to assess the quality and performance of predictive models [40].

Privacy concerns and regulations such as the Family Educational Rights and Privacy Act (FERPA) pose substantial limitations on data collection and analysis [17]. Universities face restrictions in gathering, storing, and analyzing certain types of information due to the sensitive nature of student data. These constraints can potentially diminish the accuracy and efficacy of dropout prediction models.

For instance, the university studied conducts exit surveys for departing students. While this textual data could potentially enhance our understanding of dropout factors through multimodal classification [31], we were unable to incorporate it into our study. The absence of explicit consent forms for survey participants precluded the use of this valuable information.

Furthermore, sharing data between institutions for research or model improvement becomes challenging, hindering cross-institutional studies and model generalization. Privacy Preserving Machine Learning (PPML) emerges as a promising solution to these challenges [18], [19]. PPML techniques allow machine learning models to be trained on sensitive data without directly accessing the raw information. Methods such as Federated Learning, Differential Privacy, Homomorphic Encryption, and Secure Multi-Party Computation offer ways to protect individual privacy while still leveraging valuable data for analysis. By implementing PPML, institutions may be more willing to participate in research, leading to larger and more diverse datasets that can improve model accuracy. Additionally, PPML can help universities comply with FERPA and other privacy regulations while still contributing to valuable research.

The potential benefits of PPML in student success prediction are significant. It could increase data availability, improve model accuracy, enhance privacy protection, and ensure regulatory compliance. However, challenges remain in developing efficient and practical PPML techniques for large-scale educational data analysis. Future research should focus on balancing privacy protection with model performance and interpretability, creating standardized frameworks for implementing PPML in educational contexts, and investigating the effectiveness of PPML-based models compared to traditional approaches. Additionally, the ethical implications of using PPML techniques in educational decision-making processes warrant careful consideration.

As the field of student success prediction continues to evolve, PPML represents a promising avenue for addressing privacy concerns while advancing our understanding of factors contributing to student retention. By embracing these innovative approaches, researchers and institutions can work towards more effective and privacy-conscious methods of identifying and supporting at-risk students, ultimately improving educational outcomes while respecting individual privacy rights.

Note: The author is currently working with OpenMined and has applied for a National Artificial Intelligence Research Resource (NAIRR) grant to support future work on student success retention using PPML [16].

VII. Conclusion

This study introduces a novel cluster-label-classify methodology for predicting student dropout in higher education, offering a more nuanced and adaptable approach to this persistent challenge. By combining unsupervised clustering with supervised classification, our method achieved F1 scores (0.82 to 0.89) for predicting non-retained students after addressing class imbalance through SMOTE. This approach not only improves predictive accuracy but also provides valuable insights into the characteristics of at-risk student groups, enabling more targeted interventions.

Our research highlights the complexities of addressing algorithmic bias in educational data mining. While we've made strides in balancing our dataset and considering multiple dimensions of student identity, the ethical implications of using protected attributes in predictive models remain a critical area for ongoing discussion and research.

Looking ahead, the integration of Privacy Preserving Machine Learning (PPML) techniques offers promising avenues for expanding this work while addressing crucial privacy concerns. PPML could enable cross-institutional collaboration and model generalization without compromising student data privacy, potentially leading to more robust and widely applicable predictive models.

In conclusion, our cluster-label-classify methodology, combined with careful consideration of bias and privacy issues, represents a significant step forward in the field of student retention prediction. As we continue to refine these approaches and explore new technologies like PPML, we move closer to developing more effective, ethical, and privacy-conscious tools for supporting student success in higher education. Future research should focus on implementing and evaluating PPML techniques in educational contexts, further exploring the ethical implications of predictive modeling in education, and continuing to improve the accuracy and fairness of dropout prediction models.

VIII. Acknowledgment

We extend our sincere gratitude to the Student Retention Data Team, Office of Institutional Research/Effectiveness, and the Vice President of Student Success at the studied institution

for their invaluable assistance with this research project. Their support and collaboration were instrumental in the completion of this study.

REFERENCES

[1] Braxton, John M. 2001. "Introduction to Special Issue: Using Theory and Research to Improve College Student Retention." Journal of College Student Retention; London 3 (1): 1–2.

[2] Berger, Joseph B., Gerardo Blanco Ramirez, and Susan Lyons. 2005. "Past to Present." College Student Retention: Formula for Student Success.

[3] Chovanec, Kevin, John Fields, and Praveen Madiraju. 2023. "Combining Demographic Tabular Data with BERT Outputs for Multilabel Text Classification in Higher Education Survey Data." In 2023 IEEE International Conference on Big Data (BigData), 1403–9. IEEE.

[4] Tinto, Vincent. 1975. "Dropout from Higher Education: A Theoretical Synthesis of Recent Research." Review of Educational Research 45 (1): 89–125.

[5] "The NCES Fast Facts Tool Provides Quick Answers to Many Education Questions (National Center for Education Statistics)." National Center for Education Statistics. Accessed October 19, 2022. https://nces.ed.gov/fastfacts/display.asp?id=40.

[6] Dekker, Gerben, Mykola Pechenizkiy, and Jan Vleeshouwers. 2009. "Predicting Students Drop out: A Case Study." Educational Data Mining, July, 41–50.

[7] Delen, Dursun. 2010. "A Comparative Analysis of Machine Learning Techniques for Student Retention Management." Decision Support Systems 49 (4): 498–506.

[8] Kuh, George D. 2001. "Assessing What Really Matters to Student Learning Inside The National Survey of Student Engagement." Change: The Magazine of Higher Learning 33 (3): 10–17.

[9] Duckworth, A., and S. M. Carlson. 2013. "Self-Regulation and School Success." Self-Regulation and Autonomy, November, 208–30.

[10] Kim, Byoungwook. 2017. "A Fast K-Prototypes Algorithm Using Partial Distance Computation." Symmetry 9 (4): 58.

[11] Punhani, Ritu, V. P. S. Arora, and A. Sai Sabitha. 2022. "K-Prototype Algorithm for Clustering Large Data Sets with Categorical Values to Established Product Segmentation." In Proceedings of Data Analytics and Management, 343–53. Springer Nature Singapore.

[12] Peikari, Mohammad, Sherine Salama, Sharon Nofech-Mozes, and Anne L. Martel. 2018. "A Cluster-Then-Label Semi-Supervised Learning Approach for Pathology Image Classification." Scientific Reports 8 (1): 7193.

[13] Astin, A. 1971. "Predicting Academic Performance in College : Selectivity Data for 2300 American Colleges." https://psycnet.apa.org/record/1972-23832-000.

[14] Bean, John P. 1980. "Dropouts and Turnover: The Synthesis and Test of a Causal Model of Student Attrition." Research in Higher Education 12 (2): 155–87.

[15] Alban, Mayra, David Mauricio, Technical University of Cotopaxi, Faculty of Computer Science and Computer Systems, Ecuador;, and National University of San Marcos, Artificial Intelligence Group, Peru; 2019. "Predicting University Dropout Trough Data Mining: A Systematic Literature." Indian Journal of Science and Technology 12 (4): 1–12.

[16] "The National Artificial Intelligence Research Resource (NAIRR)" Accessed July 10, 2024. https://nairrpilot.org.

[17] "Family Educational Rights and Privacy Act (FERPA)." 2021, August. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

[18] Xu, Runhua, Nathalie Baracaldo, and James Joshi. 2021. "Privacy-Preserving Machine Learning: Methods, Challenges and Directions." arXiv [Cs.LG]. arXiv. http://arxiv.org/abs/2108.04417.

[19] Trask, Andrew, Emma Bluemke, Teddy Collins, Ben Garfinkel Eric Drexler, Claudia Ghezzou Cuervas-Mons, Iason Gabriel, Allan Dafoe, and William Isaac. 2020. "Beyond Privacy Trade-Offs with Structured Transparency." arXiv [cs.CR]. arXiv. http://arxiv.org/abs/2012.08347.

[20] Yu, Renzhe, Hansol Lee, and René F. Kizilcec. 2021. "Should College Dropout Prediction Models Include Protected Attributes?" In Proceedings of the Eighth ACM Conference on Learning @ Scale, 91–100. L@S '21. New York, NY, USA: Association for Computing Machinery.

[21] Gándara, Denisa, Hadis Anahideh, Matthew P. Ison, and Lorenzo Picchiarini. 2024. "Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction." AERA Open 10 (January): 23328584241258741.

[22] Vooren, Melvin, Carla Haelermans, Wim Groot, and Henriette Maassen van den Brink. 2022. "Comparing Success of Female Students to Their Male Counterparts in the STEM Fields: An Empirical Analysis from Enrollment until Graduation Using Longitudinal Register Data." International Journal of STEM Education 9 (1): 1.

[23] Bird, Kelli A., Benjamin L. Castleman, and Yifeng Song. 2024. "Are Algorithms Biased in Education? Exploring Racial Bias in Predicting Community College Student Success." Journal of Policy Analysis and Management: [the Journal of the Association for Public Policy Analysis and Management], January. https://doi.org/10.1002/pam.22569.

[24] Chen, Rong, and Stephen L. DesJardins. 2008. "Exploring the Effects of Financial Aid on the Gap in Student Dropout Risks by Income Level." Research in Higher Education 49 (1): 1–18.

[25] Wongvorachan, Tarid, Okan Bulut, Joyce Xinle Liu, and Elisabetta Mazzullo. 2024. "A Comparison of Bias Mitigation Techniques for Educational Classification Tasks Using Supervised Machine Learning." Information. An International Interdisciplinary Journal 15 (6): 326.

[26] Lynn, von Winckelmann Stacey. 2023. "Predictive Algorithms and Racial Bias: A Qualitative Descriptive Study on the Perceptions of Algorithm Accuracy in Higher Education." Information and Learning Sciences 124 (9/10): 349–71.

[27] Hutt, Stephen, Margo Gardener, Donald Kamentz, Angela L. Duckworth, and Sidney K. D'Mello. 2018. "Prospectively Predicting 4-Year College Graduation from Student Applications." In Proceedings of the 8th International Conference on Learning Analytics and Knowledge, 280–89. LAK '18. New York, NY, USA: Association for Computing Machinery.

[28] Jayaprakash, Sujith, Sangeetha Krishnan, and V. Jaiganesh. 2020. "Predicting Students Academic Performance Using an Improved Random Forest Classifier." In 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), 238–43. IEEE.

[29] Shafiq, Dalia Abdulkareem, Mohsen Marjani, Riyaz Ahamed Ariyaluran Habeeb, and David Asirvatham. 2022. "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review." IEEE Access. https://doi.org/10.1109/access.2022.3188767.

[30] Delen, Dursun. 2010. "A Comparative Analysis of Machine Learning Techniques for Student Retention Management." Decision Support Systems 49 (4): 498–506.

[31] Fields, John, Kevin Chovanec, and Praveen Madiraju. 2024. "A Survey of Text Classification with Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?" IEEE Access PP (99): 1–1.

[32] Guo, Yifan, Helen X. Mao, Jijun Yin, and Zhi-Hong Mao. 2022. "Cluster-Then-Label Strategy for Sleep Detection U Sing Electroencephalogram (EEG)." In 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA), 37–41. IEEE.

[33] Kumar, Santosh, Xiaoying Gao, and Ian Welch. 2017. "Cluster-than-Label: Semi-Supervised Approach for Domain Adaptation." In 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA), 704–11. IEEE.

[34] Bodapati, J. D., Sajja V. R. K., N. B. Mundukur, and N. Veeranjaneyulu. 2019. "Robust Cluster-Then-Label (RCTL) Approach for Heart Disease Prediction." Ingénierie Des Systèmes D Inf 24 (3): 255–60.

[35] Beil, David, and Andreas Theissler. 2020. "Cluster-Clean-Label: An Interactive Machine Learning Approach for Labeling High-Dimensional Data." In Proceedings of the 13th International Symposium on Visual Information Communication and Interaction, 1–8. Proceedings of Machine Learning Research. New York, NY, USA: ACM.

[36] Afrose, Sharmin, Wenjia Song, Charles B. Nemeroff, Chang Lu, and Danfeng Daphne Yao. 2022. "Subpopulation-Specific Machine Learning Prognosis for Underrepresented Patients with Double Prioritized Bias Correction." Communications Medicine 2 (1): 111.

[37] Viberg, Olga, Mathias Hatakka, Olof Bälter, and Anna Mavroudi. 2018. "The Current Landscape of Learning Analytics in Higher Education." Computers in Human Behavior 89 (December): 98–110.

[38] Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority over-Sampling Technique." The Journal of Artificial Intelligence Research 16 (June): 321–57.

[39] Bird, Sarah, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. "Fairlearn: A Toolkit for Assessing and Improving Fairness in AI." Microsoft, Tech. Rep. MSR-TR-2020-32.

[40] Pasin, Ozge, and Senem Gonenc. 2023. “An Investigation into Epidemiological Situations of COVID-19 with Fuzzy K-Means and K-Prototype Clustering Methods.” Scientific Reports 13 (1): 6255.

TABLE VI
DESCRIPTION OF THE 20 MOST IMPORTANT FEATURES USED FOR CLUSTERING

Variable	Definition
Academic Program	Major
Academic Resource Center	Utilizes ARC resources
African American	African American ethnicity
Age	Age of student
Athlete	Participation in a school sponsored sport
Campus	Campus A or Campus B
Credits Attempted	Course credits attempted overall
Cumulative Term GPA	Cumulative grade point average for the last enrolled semester
Current	Student is current with university account
First Generation	The student is the first in the family to attend college
Gender	Female, male, or did not respond
High School GPA	Average grade point average in high school
Live On-Campus	Does the student live on-campus or off-campus
More than 15 credits	In the last enrolled semester, did the student take more than 15 course credits
PELL eligible	Does the student qualify for a PELL grant
Religious Affiliation of University	Does the student have the religious affiliation of the university
Cumulative Term Credits	Credits earned in the last enrolled semester
Term GPA	Grade Point Average for the last enrolled semester
Term Credits Earned	Credits earned in last enrolled semester
Total Cumulative Credits	Overall credits earned