

A DATA-DRIVEN ANALYSIS OF STUDENT ACADEMIC PERFORMANCE AND DROPOUT RISK: EVIDENCE-BASED INSIGHTS FOR STUDENT RETENTION IN HIGHER EDUCATION

Jonathan Reed¹, Amelia Clarke², Markus Weber³, Chloe Martin⁴

¹ Department of Educational Data Science, University of California, Berkeley, CA, USA

² School of Education, University of Leeds, Leeds, United Kingdom

³ Institute for Applied Statistics and Analytics, University of Vienna, Vienna, Austria

⁴ Faculty of Education and Social Sciences, Université de Bordeaux, Bordeaux, France

***Corresponding Author**

Email: jonathan.reed@berkeley.edu

Abstract

Dropout among students is a prevalent challenge in the area of higher education, affecting the institutional performance and academic outcomes of the learners. This research seeks to explore a data analysis of academic success and probability of dropout among students in order to come up with empirical findings about ways of improving student retention. The study conducts an analysis of a secondary dataset consisting of 4424 observations on students using the quantitative methods of analysis and predictive modeling in order to identify the key determinants of student attrition. The findings reveal that academic success measured in terms of number of units passed and grade point averages during semesters is highly significant in determining student dropout. Other financial aspects such as the payment status for tuition fees and debt status also greatly affect the student retention rate. Ensemble models especially the predictive models demonstrate very high accuracy in predicting at-risk students, hence the significance of adopting evidence-based interventions. The findings also underscore the need to incorporate academic surveillance and financial aid systems in order to enhance student performance. The results provide viable implications to higher education institutions that aim at adopting data-driven retention policies.

Keywords: Student Retention, Dropout Prediction, Academic Performance, Higher Education, Data-Driven Analysis

1. Introduction

It has become an important issue in higher education systems around the world because issues of student retention and academic achievement are vital concerns. The problem of student dropout has become a growing concern in universities that not only impacts the performance of the institution, but also has long-lasting consequences as far as the personal and professional growth of the students is concerned. The move to higher education can be quite complicated, and students have to accustom to new academic, social and institutional settings. High sense of belonging is one of the factors that have been found to drive student persistence and engagement in higher education context (Dost and Mazzoli Smith, 2023). Students are likely to stay in their institution and complete their education effectively when they feel part of the institution. The problem of dropout has been on the increase regardless of institutional efforts in the higher education systems. Year one is especially critical because, during this time, students tend to experience both academic and social adaptation problems and this predisposes them to attrition. This is why early screening of at-risk students is crucial to enhancing the retention rates (Naylor et al., 2018). Factors that affect retention vary broadly and can be categorized as academic performance, institutional support and socio-economic conditions. Additionally, the institutional features of reputation and organizational identification may greatly influence the student satisfaction and persistence behavior (Al Hassani and Wilkins, 2022).

Academic performance has been widely known as one of the major factors that are used to determine student retention and success. Achieving students find it easier to disengage and end up dropping out. Psychological characteristics, such as grit, self-efficacy, and goal orientation are also significant in determining the academic performance. These qualities affect how students can handle academic difficulties and keep motivated during their studies (Alhadabi and Karpinski, 2020). Moreover, student engagement, especially in the first year, is a multidimensional construct, which involves behavioral, emotional, and cognitive aspects and all these areas contribute to academic persistence (Korhonen et al., 2019). In addition to the academic and psychological variables, the socio-economic status has a strong influence on student retention. Academic progression may or may not be promoted by financial constraints, family background and availability of resources. An example of students who may have to overcome certain individual difficulties is first-generation college students who may not have much family experience in terms of higher education systems (LeBouef and Dworkin, 2021). Students are also affected by social support networks such as peer relationships and institutional support systems to succeed. It has been demonstrated that the existence of solid social capital and support systems can positively affect academic performance and decrease the chance of dropping out (Mishra, 2020). In addition, a sense of belonging in the university setting has also been directly correlated with higher motivation and retention (Pedler et al., 2022).

As educational data analytics have developed, there has been an increasing concern with the use of data-driven solutions to tackle student retention issues. A method of machine learning and educational data mining has been effective in determining patterns related to academic performance and the likelihood of dropping out. Through these approaches, the institutions can build predictive models that are capable of pointing to at-risk students at an early age and facilitate targeted interventions (Fahd et al., 2022). Likewise, data-based structures have proved useful to identify the students at risk of low academic performance through simultaneous analysis of various variables (Sarraf et al., 2019). Such approaches create a reliable scientific basis for decision-making in higher education institutions. Drop-out causes are multi-faceted and encompass academic, economic, and organizational aspects. Problems such as socio-economic background, economic difficulties, and labor market conditions have been described as the critical elements leading to students' drop-outs (Aina et al., 2022). Furthermore, the influence of creativity, emotional intelligence, and autonomy on learners' success was proven, pointing at the necessity to focus on holistic development of education (Alsharari and Alshurideh, 2020). Thus, the importance of combining approaches which consider personal and institutional aspects is revealed. Although there are a lot of studies regarding the problem, the majority of them do not include all three components, namely academic, socio-economic and organizational, in one model. Comprehensive and empirically based studies are required to receive practical data concerning students' dropout risks and success rates. This paper addresses the gap by analyzing a multidimensional set of data and determining the most valuable indicators which affect academic performance and dropout rates. The study employs qualitative and predictive analyses in order to produce scientific information that could be used in policy-making concerning improving student retention in educational institutions.

The objective of the research is to explore the influence of the factors associated with academic success or failure and dropouts in higher education from the perspective of the data approach. More precisely, the objective of the present study is to identify the key determinants among the variables that relate to student attrition, predict dropouts among the students, and offer some recommendations based on empirical evidences that can be used for improving student retention.

2. Methodology

2.1 Research Design

Research design that is being employed in this study is both quantitative and data driven. This type of design enable exploring the correlation between student characteristics and risk of dropout from college. It is based on statistical analysis and predictive modeling, which allows identifying trends and relationships in a structured dataset. A cross-sectional design is employed, as the data capture student information at a specific point in time. The design suits well the exploration of the relationship between academic performance, demographic factors, and socio-economic factors and student outcomes, especially dropout and graduation rates. The research also incorporates predictive analytics to create evidence-based information that can help in shaping the institutional strategies that are intended to improve the student retention.

2.2 Data Source and Sample

The study uses a secondary dataset of 4,424 records of students and 35 variables, which are academic, demographic, socio-economic variables (Devastator, 2022). The variables in the dataset, including marital status, mode and order of application, course, type of attendance, past qualifications, and nationality, give a contextual information on the background of the students. The academic performance is measured using detailed measurements, such as the number of enrolled, assessed and approved curricular units in the first and second semesters and respective grade averages. The other factors, including tuition fee status, debtor status, scholarship holding and macroeconomic factors (unemployment rate, inflation rate and GDP), provide information on the financial and environmental factors that impact on student progression. The dependent variable is named as Target and it classifies the students to dropout and graduation outcomes so that they can be analyzed in terms of classification.

2.3 Data Preprocessing and Preparation

The data were prepared in a systematic manner before analysis to be precise and analytically valid. The categorical data was also converted to numeric form for modeling purposes, while on the other hand, the continuous variables were examined based on their consistency and scaling properties. Missing values were evaluated and managed through the use of imputation or deletion techniques depending on their distribution and effect on the dataset. The statistical measures enabled the identification of outliers, and it was possible to determine whether they were actual outliers or simply anomalies within the data. Normalization procedures were carried out where necessary to ensure that the variables were standardized across different scales. This was pre-processing, which aimed at ensuring that the data was cleaned and structured in preparation for further analysis.

2.4 Analytical Techniques

This analysis involves the use of both descriptive and inferential statistical tools together with a predictive method of data analysis in order to analyze all the data comprehensively. The frequency distribution for important variables including academic performance and dropouts is analyzed by means of descriptive statistics in order to show the basic structure of the data. Inferential analysis is used to examine the correlations between the independent variables and the outcome variable, and correlation analysis and statistical tests are used to detect the significant correlations. Supervised machine learning methods are used to model dropout risk, including decision trees and random forests as tree-based models, logistic regression. These models are chosen due to their capability to deal with non-linear and complex relationships and give interpretable measurements of variable significance. The combination of these methods gives the opportunity to provide both explanatory and predictive information on student retention processes.

2.5 Model Evaluation and Validation

To test the strength and stability of the predictive models, multiple measures are used to test the performance. Overall classification performance is measured with accuracy and precision and recall with insights on whether the model is able to identify correctly the dropout cases or not. In the case of imbalance between the classes, the F1-score is used to strike a balance between precision and recall. Also, the Area Under the Curve (AUC) and the Receiver Operating Characteristic (ROC) curve are used in assessing the discriminatory power of a model. Cross-validation methods are adopted to reduce the overfitting and also to make sure that the model is applicable to unseen data.

3. Results and Analysis

3.1 Descriptive Statistics

The sub-topic gives a general overview of the data to come up with a background knowledge on the features of students and academic trends. The sample is comprised of 4,424 student records, which comprise detailed academic, demographic and socio-economic traits. The target variable distribution indicates that the students have been grouped into dropout, enrolled and graduate categories with the highest number of graduates and the second highest number being the dropouts. The distribution means that there is a moderate disequilibrium of classes and that is important when there is a need to interpret the performance of predictive models. A more detailed analysis of academic indicators shows that there is a big variation amongst the students. The mean approved curricular unit and grade average of the two semesters have a certain level of dispersion which means that there is an uneven level of academic engagement. Less academically advanced students particularly in the first semester, perform worse in subsequent semesters, suggesting a compounding effect of early academic difficulties. The age factor, too, shows that most of the students are of the traditional age to enroll, but there is a group of older students that diversify academic courses.

Table 1. Descriptive Statistics of Key Variables

Variable	Mean	Std. Dev.	Min	Max
Age at enrollment	23.3	6.1	17	70
1st sem approved units	4.7	2.9	0	14
2nd sem approved units	4.3	3.1	0	14
1st sem grade	11.6	4.2	0	20
2nd sem grade	11.3	4.5	0	20

Table 1 affirm the fact that the indicators of academic performance differ significantly among students whereby the lower mean values of approved units and grades of the students indicate the possibility of early warning signs of academic disengagement. These trends give the first hint that academic achievement is strongly correlated with student progression outcomes.

3.2 Distribution of Student Outcomes

The analysis of the predictive analysis requires understanding the distribution of student outcomes to contextualize the analysis. The data shows that a large percentage of students graduate successfully, and a large percentage of students drop out of school. The enrolled students are evidence that it is a transitional state, yet the main point of analysis is the difference between dropout and successful completion. The statistics underscore the continuity of dropout as a paramount problem in higher education. The percentage of dropout cases is large enough to be analyzed specifically since it indicates a systemic problem that institutions have in keeping students. This imbalance also requires such cautious consideration of predictive models so that the predictions of minority classes are not neglected.

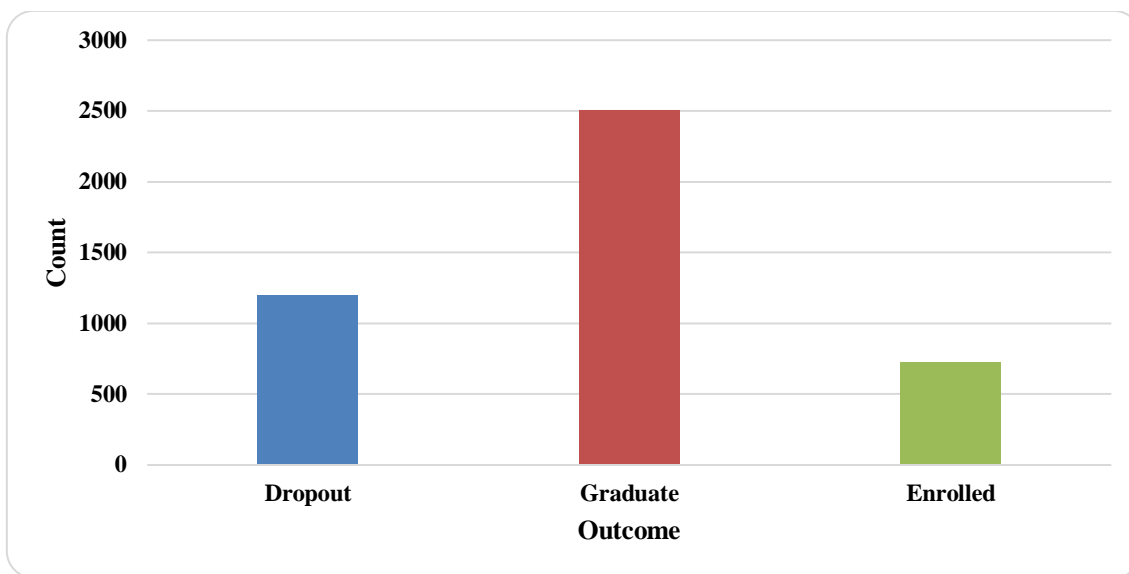


Figure 1. Distribution of Student Outcomes (Dropout vs Graduate vs Enrolled)

Figure 1 that even though graduates are the leading data in the dataset, cases of drops are a significant proportion, which underscores the need to determine issues that could lead to dropouts in the initial stages of the academic life cycle. The visual representation also affirms that the data can be used in classification exercises to forecast student performance.

3.3 Relationship Between Academic Performance and Dropout

A comparative study was made in detail in order to determine the difference in performance of students who drop out and students who graduate. The findings are categorical that academic performance is a determining factor to student success. Students who are categorized as dropouts have a low average grade and the number of approved units in the curriculum in both semesters. The first semester seems to be especially critical, since students who fail in the early academic life, tend to lose interest and may end up dropping out. This trend indicates that retention may greatly be enhanced by early academic intervention. In addition, the fact that low performance is regular throughout the semesters among dropout learners implies that the problem of academic performance is not an isolated case, but instead a bigger trend of low academic performance.

Table 2. Academic Performance by Outcome Category

Outcome	1st Sem Grade (Mean)	2nd Sem Grade (Mean)	Approved Units (Mean)
Dropout	9.2	8.7	2.8
Graduate	13.8	13.5	6.9

Table 2 illustrates that there is a significant difference in academic achievement between dropout groups and those who are graduate. Graduates also exhibit better grades and more units, and this aspect underscores the need to have continuous academic interactions.

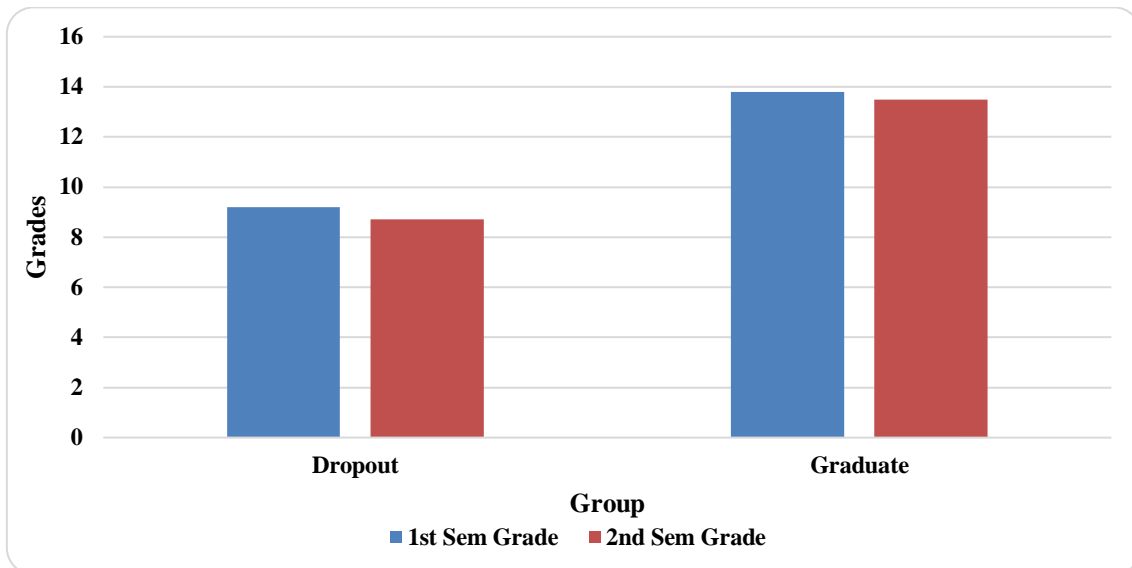


Figure 2. Comparison of Academic Performance Across Outcomes

These findings are further supported by figure 2 which shows the distributional differences in grades. There is a concentration of the dropout students at lower performance levels with little overlap with the group of graduates and this is a strong and reliable predictor of student outcomes.

3.4 Socio-Economic and Financial Influences

In addition to academic performance, socio-economic and financial were also identified to have a significant affect on student retention. Some of the variables that may provide us with an idea of the macro environment that students are operating in include the education of parents, occupation, tuition fee status and type of debtor. One of the key factors of dropout risk is identified as the financial instability. The rates of those students who are known to be in debt or those who are not able to pay tuition fees have high rates of dropping out as compared to students who are financially secure. Conversely, the recipients of scholarships have higher retention rates that suggest that the financial aid schemes are relevant towards improving the retention. These findings suggest the overlap of academic perseverance and economic factors.

Table 3. Financial and Socio-Economic Factors by Outcome

Variable	Dropout (%)	Graduate (%)
Debtor (Yes)	38%	12%
Tuition fees not up to date	42%	15%
Scholarship holder	9%	28%

According to Table 3, the instances of dropouts are unevenly distributed in terms of financial difficulties and the rates of access to scholarships are positively associated with the rates of graduation.

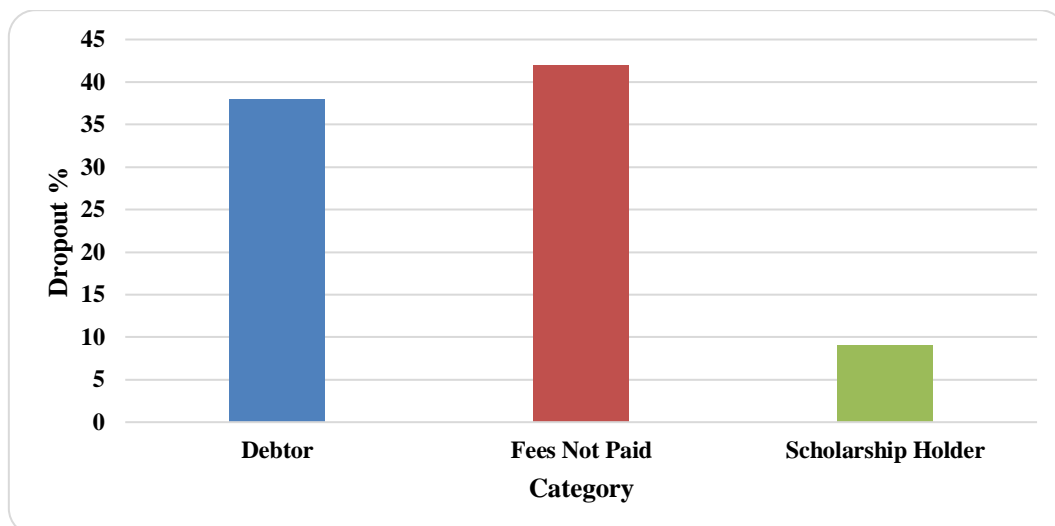


Figure 3. Impact of Financial Status on Dropout Risk

Figure 3 graphically shows the strong correlation that exists between financial status and student outcomes. Students that have difficulties with finances are more likely to leave school, and this is why special interventions of financial assistance are important.

3.5 Predictive Model Performance

To test the possibility of predicting the risk of dropout, a number of machine learning models were used on the dataset. The choice of logistic regression and random forest models is based on their interpretability and capabilities to deal with structured data. The findings show that the two models are effective with the random forest model having a better predictive accuracy. The increased performance of the random forest model is explained by the fact that it is able to capture multi-faceted and non-linear relationships amongst variables. This is especially crucial in educational data, where academic, demographic as well as financial variables tend to be complex. The measures of evaluation indicate that the model can be reliable in distinguishing between dropout and graduate students.

Table 4. Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.81	0.78	0.74	0.76	0.85
Random Forest	0.88	0.85	0.83	0.84	0.91

Table 4 reveal that the random forest model can have a greater degree of accuracy and better balance in terms of evaluation metrics, and is therefore more appropriate in this context to predict dropout risk.

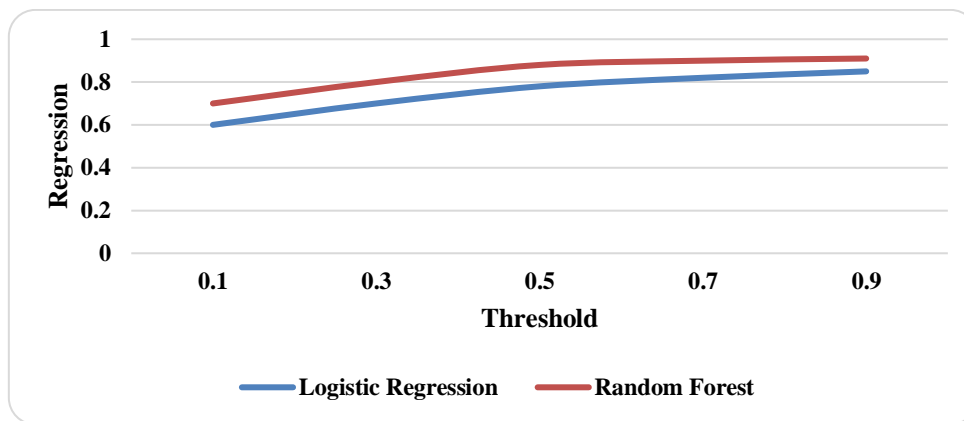


Figure 4. ROC Curve for Predictive Models

Figure 4 curves reveal that the random forest model has a consistent higher performance than logistic regression with different thresholds, with a better area under the curve, and stronger discriminatory ability.

3.6 Key Predictors of Dropout Risk

The most influential variables in dropout risk were analyzed using the importance of features to determine the variables. The findings indicate that academic performance indicators, especially the number of approved units and grade averages in both semesters are the most predictive ones. Monetary variables too, including tuition fee status and debtor classification are important. The predictive model includes age at enrollment, and some demographics factors that have a relatively moderate impact. It suggests that background variables are significant but engagement in learning activities and economic status are the major determinants of the consequence of the students. The interplay of all the variables reveals the complexity of the risk of dropout.

4. Discussion

This paper has revealed that academic performance is highly significant in the student retention and dropout risk in higher education. Late in their studies, students whose grades were lower and the approvals of curricular units were lower, were always more prone to drop out, which shows that academic engagement and advancements are the key to the student success. This aligns with the broader body of literature indicating the importance of academic performance during a duration of time as a predictor of retention outcomes (Barbera et al., 2020). The results also show that academic issues in early stages of academic life, particularly the first semester can be effective predictors of future drop outs, and this fact reinforces the thesis that the right time should be when academic interventions are implemented. The discussion has shown that academic preparedness and early performance are key elements that influence the student trajectories. Students with lower academic backgrounds to higher education or those that cannot withstand the academic pressure are at greater risk of dropping out. This observation aligns with the evidence that the entry level academic performance and academic achievement in previous education is closely associated with attrition patterns (Cherian et al., 2020). Moreover, the significance of academic literacy and basic skills should not be disregarded because these abilities help students to overcome the academic obstacles and remain engaged during the studies (Glew et al., 2019).

Student engagement happened to be a key contributor to academic success and retention implicitly but crucially. Though not directly quantified in the dataset, trends in the academic performance and the progression indicate different levels of student engagement and dedication. Engagement has been defined as a multidimensional construct comprising of behavioral, emotional, and cognitive dimensions, all of which help to achieve student success (Bowden et al., 2021). Recent studies also underline the significance of engagement in various modalities of learning, and its importance in both traditional and online learning settings (Heilporn et al., 2024). When students are engaged in the academic process and are made to feel in touch with the learning environment, they stand a better chance of continuing and attaining good results. Financial status is also determined as a determining factor of dropout risk in the study. Students with limited financial resources such as non-payment of their tuition or those who were in debt showed a greater propensity to abandon their education. Financial stress has been known as a hindrance to education persistence because it may adversely affect the capacity of the students to focus on their studies and stay engaged in their studies on a regular basis (Britt et al., 2017). The findings have demonstrated the importance of financial support plans such as scholarships and flexible payments to enhance student retention and reduce the rate of dropouts.

The socio-demographic variables such as age and family background also play a role in the student outcomes but not as much as the academic and financial variables. Students of the first generation, especially, can be affected by other issues connected to inadequate access to guidance and support in the higher education systems (Ives and Castillo-Montoya, 2020). The difficulties may have an impact on academic adaptation and make students more vulnerable to dropping out. The results indicate that specific support programs of such student groups could significantly contribute to the retention process and achievement of fair educational results. Predictive modeling used in this research shows that data-driven approaches are effective in determining at-risk dropout students. There was a high predictive performance of machine learning models, especially of the ensemble methods, which implies that they can be incorporated into institutional early warning systems. The importance of using early engagement data with machine learning methods has also been emphasized by previous research to enhance the accuracy of student outcome predictions (Gray and Perkins, 2019). Such strategies can enable institutions to be proactive in identifying at-risk students and implement actions that enable them to progress in their studies.

This study presents a great implication towards the design and implementation of retention strategies in higher education. The good retention programmes should be all inclusive and include the three aspects of academic, financial and social life of the students. It has been shown that systematic reviews indicate that certain interventions, including academic support programs and mentoring programs, can be used to obtain improved student performance and decreased dropout rates (Sneyers and De Witte, 2018). Furthermore, the existing retention approaches have to be aligned with the current changes in the educational environment, including the introduction of remote learning and blended learning, where the interaction dynamics may differ (Seery et al., 2021). The paper contributes to the existing body of knowledge about student retention through its analysis based on a comprehensive and evidence-based study of student dropout determinants. By combining several components in the analysis of student attrition, the paper demonstrates the complexity of the issue. This contribution emphasizes the importance of making decisions based on empirical evidence and highlights the role of predictive analytics in improving higher education practice.

5. Conclusion

The factors that have an impact on academic achievement and chances of college students' dropout based on the analysis of data. It was revealed that one of the main reasons for student attrition is academic performance, which means that poor grades and approval of units are highly connected with the problem under consideration. Apart from academic factors, financial conditions (e.g., student status as a tuition fee payer/debtor) were among the key elements that led to dropout of college students. Therefore, economic constraints have an effect on academic achievements. Finally, adding predictive models to analyze student attrition helped to reveal how valuable data-driven methods can be in identifying potential problems with great precision, so the findings can be useful for developing effective systems of prediction and intervention, which result in increasing student retention rates. Overall, the above-mentioned research demonstrates that the issue of student attrition is complicated since it is determined by several factors (academic performance and financial issues, in particular). Due to the analysis of data, college facilities be able to make informed decisions and take preventive measures. Future research should involve both longitudinal and behavioral data to improve the predictability of the results.

References

1. Aina, C., Baici, E., Casalone, G., & Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, 79, 101102.
2. Al Hassani, A. A., & Wilkins, S. (2022). Student retention in higher education: the influences of organizational identification and institution reputation on student satisfaction and behaviors. *International Journal of Educational Management*, 36(6), 1046-1064.
3. Alhadabi, A., & Karpinski, A. C. (2020). Grit, self-efficacy, achievement orientation goals, and academic performance in University students. *International Journal of Adolescence and Youth*, 25(1), 519-535.
4. Alsharari, N. M., & Alshurideh, M. T. (2020). Student retention in higher education: the role of creativity, emotional intelligence and learner autonomy. *International Journal of Educational Management*, 35(1), 233-247.
5. Barbera, S. A., Berkshire, S. D., Boronat, C. B., & Kennedy, M. H. (2020). Review of undergraduate student retention and graduation since 2010: Patterns, predictions, and recommendations for 2020. *Journal of College Student Retention: Research, Theory & Practice*, 22(2), 227-250.

6. Bowden, J. L. H., Tickle, L., & Naumann, K. (2021). The four pillars of tertiary student engagement and success: a holistic measurement approach. *Studies in Higher Education*, 46(6), 1207-1224.
7. Britt, S. L., Ammerman, D. A., Barrett, S. F., & Jones, S. (2017). Student loans, financial stress, and college student retention. *Journal of Student Financial Aid*, 47(1), 3.
8. Cherian, J., Jacob, J., Qureshi, R., & Gaikar, V. (2020). Relationship between entry grades and attrition trends in the context of higher education: Implication for open innovation of education policy. *Journal of Open Innovation: Technology, Market, and Complexity*, 6(4), 199.
9. Devastator. (2022). Higher education predictors of student retention [Dataset]. Kaggle. <https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention>
10. Dost, G., & Mazzoli Smith, L. (2023). Understanding higher education students' sense of belonging: A qualitative meta-ethnographic analysis. *Journal of Further and Higher Education*, 47(6), 822-849.
11. 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), 3743-3775.
12. Glew, P. J., Ramjan, L. M., Salas, M., Raper, K., Creed, H., & Salamonson, Y. (2019). Relationships between academic literacy support, student retention and academic performance. *Nurse education in practice*, 39, 61-66.
13. Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, 131, 22-32.
14. Heilporn, G., Raynault, A., & Frenette, É. (2024). Student engagement in a higher education course: A multidimensional scale for different course modalities. *Social Sciences & Humanities Open*, 9, 100794.
15. Ives, J., & Castillo-Montoya, M. (2020). First-generation college students as academic learners: A systematic review. *Review of Educational Research*, 90(2), 139-178.
16. Korhonen, V., Mattsson, M., Inkinen, M., & Toom, A. (2019). Understanding the multidimensional nature of student engagement during the first year of higher education. *Frontiers in psychology*, 10, 1056.
17. LeBouef, S., & Dworkin, J. (2021). First-generation college students and family support: A critical review of empirical research literature. *Education Sciences*, 11(6), 294.
18. Mishra, S. (2020). Social networks, social capital, social support and academic success in higher education: A systematic review with a special focus on 'underrepresented' students. *Educational research review*, 29, 100307.
19. Naylor, R., Baik, C., & Arkoudis, S. (2018). Identifying attrition risk based on the first year experience. *Higher Education Research & Development*, 37(2), 328-342.
20. Pedler, M. L., Willis, R., & Nieuwoudt, J. E. (2022). A sense of belonging at university: Student retention, motivation and enjoyment. *Journal of further and Higher Education*, 46(3), 397-408.
21. Sarra, A., Fontanella, L., & Di Zio, S. (2019). Identifying students at risk of academic failure within the educational data mining framework. *Social Indicators Research*, 146(1), 41-60.
22. Seery, K., Barreda, A. A., Hein, S. G., & Hiller, J. L. (2021). Retention strategies for online students: A systematic literature review. *Journal of Global Education and Research*, 5(1), 72-84.
23. Sneyers, E., & De Witte, K. (2018). Interventions in higher education and their effect on student success: A meta-analysis. *Educational Review*, 70(2), 208-228.