Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 1, No. 4, January : 2024

PREDICTING STUDENT DROPOUT THROUGH MACHINE INTELLIGENCE: A SURVEY

Ms. Shubhii Shuklla, Assistant Professor, College of Management, IIMT Group of Colleges.
Mr. Ajai Misra, Joint Director, Web Initiatives, Federation of Indian Chambers of Commerce & Industry.

Abstract
In a world driven by data-driven decision-making, machine learning techniques play a crucial role in uncovering insights hidden within the data. This review presents a thorough examination of predicting student dropout utilizing machine learning techniques, offering a systematic literature analysis. By comprehensively reviewing existing research, this paper explores the various methodologies and approaches employed in predicting student dropout. The systematic analysis provides insights into the effectiveness, trends, and challenges associated with machine learning applications in this domain. The findings contribute to a better understanding of the current state of research on student dropout prediction, offering a foundation for future studies and informing the ongoing development of effective intervention strategies in educational settings.

Keywords: Student Dropout Prediction, Machine Learning Applications, Educational Data Analysis, Improvements in Education Analytics, Academic Retention, Data Driven Decision Making.

Introduction

In the realm of higher education, student dropout has emerged as a critical challenge with far-reaching consequences for both individuals and institutions. The attrition of students before completing their academic programs not only represents a loss of potential talent but also poses a significant financial burden on educational institutions. The complex and multifaceted nature of student dropout necessitates innovative approaches to understand, predict, and ultimately mitigate this issue. In recent years, the integration of machine learning (ML) techniques has provided a promising avenue for revolutionizing the way educators and administrators address student dropout.

In educational landscapes, one of the persistent challenges faced by institutions globally is the alarming rate of student dropout. The premature cessation of academic pursuits not only undermines individual aspirations but also poses a substantial obstacle to the overall effectiveness and efficiency of educational systems, as well as having societal and economic implications. To address this pressing issue, machine learning techniques can be employed to predict potential dropouts and intervene early to prevent them. By leveraging historical data on student performance, behavior, and demographic information, machine learning models can identify patterns and risk factors associated with dropout[1]. The models can then be used to flag at-risk students for targeted interventions, such as counselling or academic support programs [2]. The detrimental effects of dropout extend beyond the individual, impacting educational resources, institutional effectiveness, and societal advancement. Recognizing the complexity of factors contributing to student dropout, there is a growing need for innovative and data-driven approaches to predict, understand, and mitigate this phenomenon [3]. While traditional methods of identifying at-risk students rely heavily on subjective judgments and limited data, machine learning algorithms have the potential to provide more accurate and reliable predictions [4]. Machine learning algorithms, such as decision trees, logistic regression, random forest, K-nearest neighbor, and neural networks, have shown promise in accurately predicting student dropout. These algorithms analyze large datasets and identify patterns and correlations that may not be immediately apparent to human observers. By harnessing the power of machine learning, educators and administrators can make informed decisions and allocate resources effectively to support students at risk of dropping out. Despite the potential of machine learning techniques to predict and prevent student dropouts, there exists a debate surrounding their reliance. A central concern involves the ethical implications tied to

utilizing sensitive student data for predicting academic outcomes. The use of historical data on student performance and behavior raises ethical questions regarding privacy and the possibility of biased or unfairly targeting specific groups of students [5].
In conclusion, while machine learning techniques show promise in addressing student dropout, it is essential to carefully consider the ethical implications and limitations of relying solely on these models for predicting and preventing dropout. Balancing the use of machine learning with human judgment and personalized support is crucial to effectively address the complex issue of student attrition[6].

Literature

[7] addressed the issue of student dropouts in higher education by developing a predictive model and focused on Computer Science undergraduate students from Universiti Teknologi MARA and explored the application of different data mining algorithms to predict student dropout after three years of enrollment. The authors explored various machine learning algorithms and found Logistic Regression model to be the most effective, providing reliable classification accuracy and highlighting specific courses that significantly influenced the likelihood of dropout.
[5] demonstrated that various factors like demographics, cultural background, social support, family background, educational background, socioeconomic status, psychological profile, and academic progress can influence student dropout rates. Authors classified using Random Forest, Decision Tree, Support Vector Machine, and k-nearest Neighbors, with their specific parameters tuned for this task. The performance of these classifiers was evaluated using ROC-AUC scores, and the effects of excluding each type of data (demographic, socioeconomic, macroeconomic, and academic) on the model performance were also analyzed to determine the most influential factors in student dropout. This study identified Random Forest classifier achieved the highest ROC-AUC score amongst the models tested.
[8] developed models to predict the success of university students by using various indicators. The study focused on enhancing the reliability and goodness of fit of these models under different conditions. Four types of models were designed: Chapter-Level Indicator Models, Comparative Models, Diagnostic Analysis Models, and Mathematically Refined Models. These models aimed to improve the accuracy of predicting student success, identify at-risk students, and enhance educational outcomes by providing support and refining practices based on predictive analytics.
[9] demonstrated predictive models which were developed and evaluated using six established metrics. Automatic hyperparameter optimization through random search was employed to enhance algorithm effectiveness. The models, based on data from the initial week of the course, demonstrated high precision in predicting student dropouts. The use of stacked generalization further improved the models' ability to classify potential dropouts. Interestingly, grouped models, utilizing default settings of each algorithm, outperformed individually optimized ones, achieving a predictive accuracy rate exceeding 96%. Additionally, the study highlighted the greater predictive value of behavioral data from student interactions compared to demographic information.
[10]designed a machine learning framework to early predict medical students' performance in high-stakes exams like the Comprehensive Medical Basic Sciences Examination (CMBSE). The framework utilized classic and ensemble machine learning approaches to predict both pass/fail status and scores. It addressed challenges such as imbalanced student numbers, diverse features, and identifying at-risk and high-performing students. Comparative analysis of classic models (logistic regression, support vector machine, k-nearest neighbors) and ensemble models (voting, bagging, random forests, adaptive boosting, extreme gradient boosting, stacking) revealed that ensemble models, especially random forests and stacking, performed optimally. Validation on a real dataset of 1005 medical students over five years demonstrated the framework's robustness. The study concluded that this machine learning framework can effectively predict student performance in licensure exams, offering a resource-saving tool for educational systems to optimize success and adapt to disruptions in traditional testing environments.

[11]explored online entrepreneurship education using modified ensemble machine learning model," the researchers developed a modified ensemble machine learning approach to predict students' adaptability in online entrepreneurship education. The model combines predictions from multiple algorithms to enhance forecast accuracy, considering factors such as demographic data, prior academic performance, learning practices, and interaction patterns. The goal is to assist educators and administrators in identifying students needing additional support, enabling the customization of educational strategies, and designing targeted interventions to improve adaptability and enhance the overall learning experience in online education settings.

[12] researchers developed a predictive model using machine learning, specifically utilizing boosting decision trees, to identify students at risk of dropping out in Chile. The model contributes to public policy by allowing the profiling of schools for qualitative studies, understanding students' dropout trajectories, simulating the impact of events like pandemics, and estimating the return on investment of school retention policies. The model identifies risk factors such as academic lag, high absenteeism, repeated grades, and overage status, offering insights to design targeted interventions and long-term policies to reduce dropout rates, particularly in contexts like Chile where socioeconomic factors can amplify the issue.

[13] developed generalization ensemble model integrated Random Forest, Extreme Gradient Boosting, Gradient Boosting, and a Feed-Forward Neural Network. The two-layer ensemble aimed to enhance prediction accuracy by optimizing parameters through grid search. The model underwent comprehensive performance evaluations, considering metrics such as accuracy, precision, recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve. The system's purpose was to identify students at risk of dropping out, facilitating timely interventions to provide necessary support and ultimately helping educational institutions reduce dropout rates and improve academic outcomes.

[14] Created an Artificial Neural Network-Long Short-Term Memory (ANN-LSTM) model for forecasting student performance in Massive Open Online Courses (MOOCs). The model classifies student outcomes into categories like Distinction, Pass, Fail, and Withdrawn. By incorporating both demographic information and clickstream data, the ANN-LSTM model provides day-wise predictions throughout the course. Its unique ability to account for time dependencies, facilitated by the LSTM component, positions it as a more accurate predictor, particularly as the course progresses and more representative student behavior data becomes available beyond the initial days.

[15] demonstrated a prediction framework for student performance adaptable to diverse curriculum guidelines. Key components include feature engineering aligning with students' learning behavior, leveraging domain knowledge to accommodate dataset variations, and employing a two-layer ensemble learning model incorporating algorithms like KNN, SVM, and Random Forest. The framework emphasizes continuous tracking to accommodate the dynamic nature of student performance as they progress through courses. The intended outcome is to aid academic stakeholders in identifying at-risk students, enhancing educational quality, and enabling timely interventions to ensure on-time graduation.

[16] designed machine learning and deep learning models to predict early dropout in online learners by assessing weekly performance. The models, including Random Forest and Deep Neural Network, aimed to provide weekly predictions, enabling timely interventions to prevent dropouts. The Random Forest model demonstrated superior effectiveness with the highest accuracy and fastest processing speed among the tested models. While Deep Neural Network models did not surpass the Random Forest, they showed promise, prompting further investigation into the applicability and performance of various deep learning models, such as Convolutional Neural Networks and Recurrent Neural Networks, for distributed prediction in future studies.

[17] devised predictive models using machine learning to anticipate the risk of dropout among Higher Education students at various study stages. The process involved acquiring and preprocessing data from diverse academic sources, employing algorithms like Gradient Boosting, Random Forest, Support Vector Machine, and an Ensemble model. A unique cascade connection of models allowed

continuous refinement of predictions as more data became available. Evaluation metrics such as precision and recall were used to interpret results. The models, spanning from enrollment to the end of the second year, aimed to identify at-risk students and inform dropout prevention policies. The research also provided guidelines for model reproduction and adaptation across different institutional contexts, presenting a comprehensive system for early intervention to enhance student retention rates.
[6] conducted a study using administrative data from the Anagrafe Nazionale degli Studenti to predict university student dropout behavior in Italy. They applied various prediction algorithms, including least absolute shrinkage and selection operator, random forest, gradient boosting machines, and neural network, to identify at-risk students. The study utilized a comprehensive dataset covering all students in the Italian university system during the 2013–14 academic year, tracking their academic trajectories until March 21, 2018. Exclusions were made for specific programs, international students, and online universities, resulting in a final dataset of 230,336 students. Factors examined included first-year ECTS credits, family income, high-school grade, type of high school, and distance to the nearest university. The research aimed to provide insights into key predictors of student dropout, facilitating potential interventions and policy suggestions to reduce dropout rates in Italian higher education.
[3] Accurately predicting dropout risks empowers students to plan effectively and boosts their dedication to studies. This information aids school management in deciding on repeat course requests, and educators can use it to intervene promptly for enhanced student engagement. The research presented focuses on course-level dropouts, revealing that selected indicators, not reliant on system logs, are beneficial even with small datasets. The methodology is reliable for predicting course completion, allowing timely educator interventions. Classification models demonstrated trustworthy results, achieving 77%-93% accuracy in predicting completion or early dropout. Recognizing at-risk students is the initial step, followed by understanding individual needs and implementing tailored prevention strategies. Instructors should concentrate on students' unique needs for effective dropout prevention. Critical factors identified by proposed models to predict dropout risks could be based on past achievements, forming a basis for course recommendations. Recommendations should consider talents, skills, preferences, and free-time activities, making proposed methods generalizable to other courses.
[18] The application of uplift modeling, an advanced analytical technique, aimed to evaluate the individual treatment effect of actions such as tutorials or retention strategies, specifically tailored for at-risk students. The researchers employed uplift modeling in a targeted manner, focusing on four key aspects. Firstly, they applied this analytical approach to predict individual treatment effects, accurately forecasting the potential impact of tutorials on each student's likelihood of not dropping out. Secondly, the research included a comprehensive model comparison and selection process, with MOAxgboost demonstrating superior performance based on higher Qini values. Additionally, the researchers explored feature importance and student profiling, analyzing characteristics to differentiate treatment responders from non-responders, using pre-treatment data like PSU mathematics scores, family size, and high school type. Finally, the study leveraged uplift modeling for customized intervention design, utilizing insights to refine retention strategies by prioritizing students predicted to be more responsive. This approach aimed to ensure a more efficient allocation of resources, ultimately enhancing the overall effectiveness of intervention efforts.
[19] designed a computational approach employing various machine learning algorithms to predict university students' risk of dropping out. In the context of Brazilian public universities' students with a higher tendency to drop out and to determine the most significant features contributing to this prediction, the Random Forest algorithm yielded the best result, achieving a success rate of approximately 80% in predicting student dropout. The model developed by the researchers indicated that the most determinant characteristics for dropout prediction included age, participation in extracurricular activities, and the total hours of the course.
[20] developed a predictive tool with the aim of early identification of the risk of university student dropouts, either during the application phase or the first year of university. Utilizing well-known

machine learning classification algorithms, including Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and Random Forest, the study evaluated their performance using different feature sets. For the basic feature set (personal information and high school records), LDA and SVM achieved approximately 62% accuracy, while Random Forest had a slightly lower accuracy of around 56%. Introducing the Academic Learning Rate (ALR) in the feature set improved results, with LDA reaching 75%, SVM reaching 81%, and Random Forest achieving 63%. Further inclusion of Course Credits (CC) enhanced accuracy across all models, with LDA, SVM, and RF achieving 85%, 87%, and 87%, respectively. The research concluded that integrating information on course credits significantly improves predictive performance.

[21] developed an intelligent decision support system to predict academic failure using student data from the Industrial University of Santander. Researchers integrated Multiple-Criteria Decision Making feature extraction guided by TOPSIS-based features and applied the ADASYN technique to address imbalanced datasets. XGBoost was compared with other tree-based machine learning algorithms, including Gradient Boosting Machine, Random Forests, and Decision Trees. These algorithms were selected for their explainability, effectiveness on small datasets, and suitability for ensemble methods, which are valuable in predicting academic failure and dropout. The researchers performed hyperparameter tuning to optimize the predictive model, evaluating its success through various metrics such as precision, recall, F1-score, accuracy, Cohen's kappa score, and the area under the receiver operating characteristic curve.

[22] established associations between each feature and G3 performance. Noteworthy correlations include a strong negative correlation with Past Failures (-0.360415), indicating a significant link between a history of failing classes and lower G3 grades. On the positive side, Mother Education (0.217147) showed that higher maternal educational attainment, Higher Education (0.182465), Father Education (0.152457) correlate with better G3 grades and better student performance. Additionally, Age (-0.161579) displayed a negative correlation, indicating that older students tend to have lower G3 grades compared to younger counterparts. The researchers utilized these features to train machine learning models, aiming to predict student grades based on the identified factors, highlighting the significance of features with higher correlations in influencing student performance.

[23] researchers designed a dropout prediction method, utilizing outlier detection via clustering with unsupervised learning. This method was meant to identify students who may have severe difficulties with the lesson material early on, so that interventions could be made to potentially prevent these students from dropping out of the course.

[24] researchers designed a system, aims to identify students at risk of dropping out so that interventions can be implemented to help retain them. Bayesian classifiers like BayesNet with K2 and a maximum of 5 parents were used and demonstrated to be effective in detecting potential dropouts, although the success rate is not exceptionally high. The study highlights the importance of using a combination of factors within the classifiers to improve the prediction of students who may drop out. By applying such methods, educational institutions can proactively address the issue of student dropout by implementing targeted support measures for those identified as at risk, ultimately enhancing student retention and success.

Table 1 Review Summary

Authors	Domain	Data Source	Algorithms
[12]	A predictive model designed to pinpoint students at risk of school dropout in Chile.	Chilean Ministry of Education (MINEDUC)	SVM, Gradient Boosting Decision trees
[7]	A predictive model to identify the predictive features impacting	Data from a higher education institution in Malaysia	C4.5 Decision tree, Naïve Bayes

	dropout in Higher Education		
[8]	A model designed to accurately anticipate students who may encounter challenges or fail on an e-learning platform.	Data from a Hungarian e-learning platform	Random Forest, Support Vector Machines, Bayesian Networks
[9]	Created a set of machine learning models to accurately predict early dropout in MOOC students	Data from Massive Open Online Courses	Logistic Regression(LR), Decision Trees, Random Forest, Gradient Boosting
[10]	Develop a machine learning framework for the early prediction of medical students' performance in high-stakes examinations like the Comprehensive Medical Basic Sciences Examination	Data from medical students in Iran	Support Vector Machines K-Nearest Neighbors, Voting Ensemble, Bagging, Random Forest, Adaptive Boosting, Extreme Gradient Boosting, Stacking Ensemble
[11]	designed a modified ensemble machine learning model to predict the adaptability level of students in online entrepreneurship education	Data from online entrepreneurship education programs	Modified ensemble machine learning model (combination of decision tree, logistic regression, and neural network)
[13]	Designed a stacked generalization ensemble machine learning model to support educational institutions in identifying and assisting at-risk students proactively.	Data from university classes	Ensemble machine learning approach (combination of logistic regression, random forest, and neural network
[14]	Designed an Artificial Neural Network-Long Short-Term Memory model to predict the performance of students in Massive Open Online Courses	Data from Massive Open Online Courses	ANN-LSTM (combination of Artificial Neural Networks and Long Short-Term Memory)
[15]	Designed a prediction framework to help academic stakeholders	Data from educational platforms	Two-layer ensemble machine learning approach (combination

	identify at-risk students, improve education quality, and intervene appropriately to ensure students can graduate on time		of decision tree and neural network)
[16]	Developed a model for predicting early dropouts by checking students' performance on a weekly basis in an online learning context	Data from online university learning	Random Forest, Support Vector Machine, Decision Tree, Deep Neural Network
[17]	Designed a set of predictive models to anticipate the dropout risk of Higher Education students at different stages of their studies	engineering school within a Spanish public university	Supervised learning algorithms including Logistic Regression(LR), Decision Trees, Random Forest, and Neural Networks
[6]	Designed an extensive predictive analysis to identify key predictors of student dropout in Italian higher education, allowing for potential interventions and policy suggestions to reduce dropout rates	Data from Italian higher education institutions	Random Forest, Support Vector Machine
[25]	Create predictive model to accurately identify students who are at risk of academic failure, such as dropping out or falling behind in their courses	Data from a specific course	LR, Random Forest, Decision Trees, K-Nearest Neighbors, Support Vector Machines, AdaBoost, XGBoost
[18]	Designed uplift models to identify students who need academic interventions	Data from higher education institutions	Modified Outcome Approach with Random Forest, Modified Outcome Approach with XGBoost, Modified Covariate Approach with Random Forest, Modified Covariate Approach with XGBoost, Separate Model Approach using Random Forest, Separate Model Approach using

			XGBoost, X-Learner, R-Learner
[26]	predict students who are at risk of dropping out of school in Chile	Data from primary and secondary schools	NA
[27]	Predict the pinpoint area where interventions necessary	Data from a university level course	LR, Decision Tree, Support Vector Machine, Random Forest
[19]	Designed a computational approach employing to predict university students' risk of dropping out from Brazilian public universities	Data from engineering courses in Brazil	Naive Bayes, K-Nearest Neighbors, Decision Trees, Random Forest, Neural Networks
[20]	Designed a predictive tool intended to help to identify the risk of university student dropouts at an early stage	Data from multiple educational contexts	Random Forest, Support Vector Machine, Neural Network
[28]	design a framework for developing a predictive information system to help the educational institution in preventing student dropout	ICFES in Colombia, Colombian Ministry of National Education and the Colombian Directorate of Social Development	Random Forest, Decision Tree, Grad-Boosting

Table 2 : Positive and Negative insights of existing studies

Authors	Positive Insights	Negative Insights
[12]	discusses the significant negative consequences of school dropout on individuals and society, such as lower qualification jobs, higher levels of poverty, and lower life expectancy. incorporates a vast range of factors at individual, family, school, and extra-school levels that could influence a student's likelihood of dropping out.	generalizability of findings, reliance on administrative data impact of external factors, such as the pandemic in 2019
[7]	High accuracy in predicting student drop-out, identification of key factors like grades in certain courses influencing drop-out, potential for early intervention	Limited generalizability to other institutions, reliance on specific dataset from a single institution

[8]	Evaluated the effectiveness of diverse indicators, encompassing both static elements like demographic information and dynamic components such as student performance and engagement data. Examined factors including course complexity, instructional format (full-time vs. correspondence), and cohort size.	Focused on a specific set of courses at the University of Dunaújváros Although different indicators were examined, the research might not comprehensively address all factors leading to student dropout, like personal life circumstances, mental health issues, financial constraints, and other extracurricular activities.
[9]	Achieved a prediction accuracy exceeding 96% for early student dropout in a MOOC, based on first-week data. Engagement data within the MOOC, proved to be a more predictive indicator of dropout compared to relying solely on demographic information.	Dropout patterns might develop or change over time, so analysis throughout the course duration could potentially provide additional insights. Massive Open Online Course (MOOC) designed for professionals in smart city domains only
[10]	framework enables the early detection of students who may fail high-stakes exams, thus allowing educators and administrators to intervene early and provide additional support to those students. suggests that including additional data sources, like university entrance exam scores, and testing on larger datasets could improve model performance.	more diverse and larger population across different educational settings would help to generalize the findings. passing thresholds for the CMBSE changes annually, which introduces complexity in predictive modeling. Addressing this variability might refine prediction quality.
[11]	model enables educators to identify students requiring additional support, formulate precise instructional strategies implement targeted interventions aimed at improving students' adaptability	data set could be more diverse Incorporating qualitative data, such as student feedback and instructor evaluations may provide a deeper understanding of adaptability factors.
[13]	provided insights into how educational models can be implemented in actual university settings, serving as a reference for other higher education institutions looking to deploy predictive analytics for student support. novel stacked ensemble model, the research provides a more accurate tool for predicting student dropout, which is a critical factor in student retention strategies.	Incorporating qualitative data like student feedback, psychological assessments, and socio-economic background could add depth to the model's predictive ability. understanding the causality behind dropout could lead to more effective intervention strategies.

[14]	Educational content providers and instructors can leverage the ANN-LSTM model to monitor student success and tailor educational interventions, support, and resources. It could be integrated into the backend of MOOC platforms to provide real-time analytics on student performance.	The model may require substantial computational resources due to the massive data generated by students in MOOCs. inclusion of more varied data, such as assessment results, and might look into different model architectures or class weight adjustments to further improve accuracy.
[15]	Educational institutions can leverage from this to better allocate resources, such as tutors or additional support services, to students and courses where they are needed most. can be useful for academic advisors to provide targeted support to students who are at risk of graduating late.	use of sensitive student data, raising concerns about data privacy and security. Ensuring ethical use of data and compliance with privacy regulations is critical. in other regions may differ, requiring adjustments to the algorithm.
[16]	advances the application of predictive analytics within the educational sector, particularly focusing on improving retention rates by identifying students who are likely to drop out early in their courses. aids in optimizing online learning platforms and management systems by identifying key factors tied to dropout, which can lead to design improvements in online courses and LMS features.	Very specific to the institution from which the data was gathered. Different institutions might have different contributing factors to dropout risks Technological enhancements based on this research might not address non-technological factors contributing to dropouts, like financial or personal issues.
[17]	predictive capability can enable institutions to intervene early and provide targeted support to at-risk students demonstrates the ability to predict dropout at different stages of a student's academic journey with a high degree of accuracy.	focuses on a specific group of students from engineering degrees in a particular institution, which may limit the generalizability focuses on predicting student dropout, but does not establish causality between the identified risk factors and dropout
[6]	enables the identification of potent predictors of dropout behaviour, such as the number of ECTS earned in the first year, family income, high school grades, and high school type focus on prediction aligns with the consensus	considers data from the one academic year, which may not be representative of more recent trends in dropout behaviour relies on administrative data, which may be subject to selection bias or measurement error

[3]	educators can use the predictive models to identify students who may need additional support and intervene knowledge obtained from the research can serve as the basis for designing course recommendations, considering the talents, skills, preferences, and free-time activities of students in their schedules	selection of classifiers used in the study is noted as a limitation dataset used was small
[25]	demonstrated superior prediction performance compared to base models, as indicated by higher precision, recall, and F1-score results features characterizing students' performance and their interactions in the course are considered the most significant for identifying students at risk of dropping out	quantity and quality of individual features in the dataset acknowledges that the objective of which characteristics contribute the most to accurate predictions is personal and may not always be universally true
[18]	highlights the importance of pre-matriculation attributes in designing proactive retention efforts implementation of uplift modeling to maximize the effectiveness of retention efforts in higher education institutions	Integration academic data from upcoming semesters could improve model predictions and enhance understanding of the long-term effects of the program Utilization of the uplift modeling framework in diverse institutional settings
[26]	identifies common predictors of student performance evaluates the quality of existing prediction models	NA
[27]	influenced student dropout rates in the leveling course were regime, type of leveling course, application grade, and vulnerability index. highest risk of dropping out were those in vulnerable situations, with low application grades, who were enrolled in the leveling course for technical degrees.	NA

Conclusion

In conclusion, the prediction of student dropout using machine learning approaches has shown promise in various educational contexts. Researchers have utilized algorithms such as decision trees, random forest, support vector machines, and neural networks to develop models for predicting student dropout. These models have the potential to capture a wide range of factors influencing dropout and can be used to identify students at risk of failure or dropout at an early stage. Additionally, the use of machine learning algorithms allows for the prediction of dropout patterns and the implementation of timely interventions to prevent student attrition. Overall, the development of predictive tools for student

dropout using machine learning is an important area of research. It has the potential to support educational institutions in identifying at-risk students and implementing proactive measures to prevent dropout. Furthermore, the implementation of predictive information systems can contribute to more effective dropout prevention strategies in various educational settings. Several factors identified in these studies, such as personal data, academic records, and course credits, can be useful in predicting student dropout. These predictive tools have the potential to significantly reduce the negative consequences of student dropout on both individuals and institutions. Additionally, the use of machine learning approaches has allowed for the creation of predictive models that outperform traditional statistical methods in terms of accuracy and performance. These findings highlight the importance of further research and development in this area, as well as the potential for implementing predictive systems for student dropout prevention using machine learning algorithms. In conclusion, the predictive system for university-level dropout prevention using machine learning approaches holds great potential for implementation in various educational contexts.

Moving forward, it is important for future research to focus on addressing the limitations of current predictive models for student dropout. Specifically, there is a need to explore ways to enhance the generalizability of these models to different educational contexts. This could involve the development of more adaptable algorithms or the integration of diverse datasets from various educational settings. Furthermore, there is a potential for incorporating qualitative data alongside quantitative metrics in predicting student dropout. Qualitative insights such as student behavior, engagement, and socio-economic background could offer a more comprehensive understanding of the factors influencing dropout, thereby improving the accuracy of predictive models.

Another area for future investigation is the ethical implications of utilizing predictive systems for student dropout prevention. It is crucial to ensure that the implementation of such systems respects students' privacy, autonomy, and diversity. Research on the ethical considerations of predictive modeling in education could lead to the development of guidelines and best practices for responsible implementation. Lastly, longitudinal studies that track the effectiveness of interventions based on predictive models could provide valuable insights into the impact of early identification and support for at-risk students. This could inform evidence-based strategies for educational institutions to effectively allocate resources and support mechanisms. In summary, while machine learning approaches present an opportunity to revolutionize student dropout prevention, it is essential for future research to address the current limitations and ethical considerations in order to realize the full potential of predictive information systems in diverse educational contexts.

References

[1] S. C. Matz, C. S. Bukow, H. Peters, C. Deacons, A. Dinu, and C. Stachl, “Author Correction: Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics (Scientific Reports, (2023), 13, 1, (5705), 10.1038/s41598-023-32484-w),” *Sci. Rep.*, vol. 13, no. 1, pp. 1–2, 2023, doi: 10.1038/s41598-023-36579-2.

[2] Fisnik Dalipi; Ali Shariq Imran; Zenun Kastrati, “MOOC dropout prediction using machine learning techniques: Review and research challenges,” [Online]. Available: https://ieeexplore.ieee.org/document/8363340.

[3] J. Kabathova and M. Drlik, “Towards predicting student’s dropout in university courses using different machine learning techniques,” *Appl. Sci.*, vol. 11, no. 7, 2021, doi: 10.3390/app11073130.

[4] and J. Z. Huade Huo, Jiashan Cui, Sarah Hein, Zoe Padgett, Mark Ossolinski, Ruth Raim, “Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach,” [Online]. Available: https://journals.sagepub.com/doi/10.1177/1521025120963821.

[5] S. Kim, E. Yoo, and S. Kim, “Why Do Students Drop Out? University Dropout Prediction and Associated Factor Analysis Using Machine Learning Techniques.”

[6] M. Delogu, R. Lagravinese, D. Paolini, and G. Resce, "Predicting dropout from higher education: Evidence from Italy," *Econ. Model.*, vol. 130, no. January 2023, p. 106583, 2024, doi: 10.1016/j.econmod.2023.106583.

[7] W. F. Wan Yaacob, N. Mohd Sobri, S. A. M. Nasir, W. F. Wan Yaacob, N. D. Norshahidi, and W. Z. Wan Husin, "Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques," *J. Phys. Conf. Ser.*, vol. 1496, no. 1, 2020, doi: 10.1088/1742-6596/1496/1/012005.

[8] L. Bognár and T. Fauszt, "Factors and conditions that affect the goodness of machine learning models for predicting the success of learning," *Comput. Educ. Artif. Intell.*, vol. 3, no. January, 2022, doi: 10.1016/j.caeai.2022.100100.

[9] T. Panagiotakopoulos, S. Kotsiantis, G. Kostopoulos, O. Iatrellis, and A. Kameas, "Early dropout prediction in moocs through supervised learning and hyperparameter optimization," *Electron.*, vol. 10, no. 14, 2021, doi: 10.3390/electronics10141701.

[10] H. Mastour, T. Dehghani, E. Moradi, and S. Eslami, "Early prediction of medical students' performance in high-stakes examinations using machine learning approaches," *Heliyon*, vol. 9, no. 7, p. e18248, 2023, doi: 10.1016/j.heliyon.2023.e18248.

[11] A. Malik *et al.*, "Forecasting students' adaptability in online entrepreneurship education using modified ensemble machine learning model," *Array*, vol. 19, no. June, p. 100303, 2023, doi: 10.1016/j.array.2023.100303.

[12] P. Rodríguez, A. Villanueva, L. Dombrovskaia, and J. P. Valenzuela, *A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile*, vol. 28, no. 8. Springer US, 2023.

[13] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Comput. Educ. Artif. Intell.*, vol. 3, no. March, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.

[14] F. A. Al-azazi and M. Ghurab, "ANN-LSTM: A deep learning model for early student performance prediction in MOOC," *Heliyon*, vol. 9, no. 4, p. e15382, 2023, doi: 10.1016/j.heliyon.2023.e15382.

[15] S. A. Priyambada, T. Usagawa, and M. ER, "Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge," *Comput. Educ. Artif. Intell.*, vol. 5, no. January, p. 100149, 2023, doi: 10.1016/j.caeai.2023.100149.

[16] H. S. Park and S. J. Yoo, "Early Dropout Prediction in Online Learning of University using Machine Learning," *Int. J. Informatics Vis.*, vol. 5, no. 4, pp. 347–353, 2021, doi: 10.30630/JOIV.5.4.732.

[17] A. J. Fernandez-Garcia, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, and F. Sanchez-Figueroa, "A real-life machine learning experience for predicting university dropout at different stages using academic data," *IEEE Access*, vol. 9, pp. 133076–133090, 2021, doi: 10.1109/ACCESS.2021.3115851.

[18] D. Olaya, J. Vásquez, S. Maldonado, J. Miranda, and W. Verbeke, "Uplift Modeling for preventing student dropout in higher education," *Decis. Support Syst.*, vol. 134, no. January, p. 113320, 2020, doi: 10.1016/j.dss.2020.113320.

[19] A. M. Mariano, A. B. De Magalhães Lelis Ferreira, M. R. Santos, M. L. Castilho, and A. C. F. L. C. Bastos, "Decision trees for predicting dropout in Engineering Course students in Brazil," *Procedia Comput. Sci.*, vol. 214, no. C, pp. 1113–1120, 2022, doi: 10.1016/j.procs.2022.11.285.

[20] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, *Student dropout prediction*, vol. 12163 LNAI. Springer International Publishing, 2020.

[21] A. López-García, O. Blasco-Blasco, M. Liern-García, and S. E. Parada-Rico, "Early detection of students' failure using Machine Learning techniques," *Oper. Res. Perspect.*, vol. 11, no. September, p. 100292, 2023, doi: 10.1016/j.orp.2023.100292.

[22] N. R. Beckham, L. J. Akeh, G. N. P. Mitaart, and J. V. Moniaga, "Determining factors that affect student performance using various machine learning methods," *Procedia Comput. Sci.*, vol. 216, no. 2022, pp. 597–603, 2022, doi: 10.1016/j.procs.2022.12.174.

[23] S. Oeda and G. Hashimoto, "Log-Data Clustering Analysis for Dropout Prediction in Beginner Programming Classes," *Procedia Comput. Sci.*, vol. 112, pp. 614–621, 2017, doi: 10.1016/j.procs.2017.08.088.

[24] A. Viloria, O. B. P. Lezama, and N. Varela, "Bayesian classifier applied to higher education dropout," *Procedia Comput. Sci.*, vol. 160, pp. 573–577, 2019, doi: 10.1016/j.procs.2019.11.045.

[25] J. Pecuchova and M. Drlik, "Predicting Students at Risk of Early Dropping Out from Course Using Ensemble Classification Methods," *Procedia Comput. Sci.*, vol. 225, pp. 3223–3232, 2023, doi: 10.1016/j.procs.2023.10.316.

[26] L. S. Rodrigues, M. Dos Santos, I. Costa, and M. A. L. Moreira, "Student Performance Prediction on Primary and Secondary Schools-A Systematic Literature Review," *Procedia Comput. Sci.*, vol. 214, no. C, pp. 680–687, 2022, doi: 10.1016/j.procs.2022.11.229.

[27] I. Sandoval-Palis, D. Naranjo, J. Vidal, and R. Gilar-Corbi, "Early dropout prediction model: A case study of university leveling course students," *Sustain.*, vol. 12, no. 22, pp. 1–17, 2020, doi: 10.3390/su12229314.

[28] S. Guzmán-Castillo *et al.*, "Implementation of a Predictive Information System for University Dropout Prevention," *Procedia Comput. Sci.*, vol. 198, no. 2020, pp. 566–571, 2022, doi: 10.1016/j.procs.2021.12.287.