Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 4 (2025)

Enhancing Student Retention and Performance Prediction in E-Learning Environments Using Supervised Machine Learning Models

Dr. Nikhil Saini[1]

[1]Assistant Director, Le Cordon Bleu School of Hospitality and Tourism, GD Goenka University, Gurugram, Haryana

nikhilsaini0022@gmail.com

Dr. Christabell Joseph[2]

[2]Associate Professor, School of Law, Christ (Deemed to be University), Bangalore.

Christabell.joseph@christuniversity.in

Dr. Jigar Rupani[3]

[3]Assistant Professor, School of Business and Management, Christ University, Bengaluru - 560073

jigarrupani8483@gmail.com

Orcid Id: 0000-0002-2480

Dr. J. Sridevi[4]

[4]Associate Professor, School of Commerce, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India - 600062

drsridevij@veltech.edu.in

Dr. A. Pankajam[5]

[5]Associate Professor, Department of Business Administration, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore-641043, Tamil Nadu, India

ambipankaj@gmail.com

Aaditya Jain[6]

[6]Assistant Professor, Department of Computer Science & Engineering, Teerthanker Mahaveer University, Moradabad, UP, India

adityajain58@gmail.com

Abstract: Attrition rate seems to be a serious problem in the field of online education as it compromises the scalability and efficiency of Virtual Learning Environments (VLEs). This empirical work deals with the issue of student dropout by building a high-dimensional predictive model based on Open University Learning Analytics Dataset (OULAD). Our model uses high-quality supervised machine learning algorithms (i.e. Random Forest, Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost)) to process granular click stream databases and demographic information to predict student performance. In order to address the existing class imbalance in the educational data, the Synthetic Minority Over-sampling Technique (SMOTE) is strictly used. The experimental findings reveal the XGBoost classifier optimized through grid search has a high quality of 92.4% and F1-score of 0.91, which is much better than the baseline models. These findings present a clue to the usefulness of ensemble learning to assist in timely information-based pedagogical interventions.

Keywords: Student Retention, XGBoost, Educational Data Mining, SMOTE, Machine Learning, VLE, Predictive Analytics, Learning Analytics

1. Introduction

E-learning platforms have made education accessible to everyone, but have a widespread so-called retention crisis, where dropout rates in Massive Open Online Courses (MOOCs) are often over 80%. As opposed to the brick-and-mortar institution where instructors could use physical indicators to determine interactions, an online setting is deficient in physical closeness, and thus it becomes hard to detect students at-risk until they have dropped out. This has put the development of data-driven and automated early warning systems as an immediate research agenda.

The paper will bring forward a solid empirical model to forecast the retention of students based on the mining of behavioral logs in VLEs. The ultimate goal is to move to proactive prediction as opposed to reactive analysis (Shafiq et

al., 2022). We expect to create a predictive model that is both sensitive and computationally efficient by dynamically interacting the demographic variables with dynamic interaction measures, e.g., frequency of resource access and latency of submission of an assessment. The technical issues of high feature dimensionality and class imbalance are the specific focus of this paper in which ensemble techniques are used to model non-linear behavioral patterns of students that are not accurately captured by a linear model (Malik et al., 2025).

2. Literature Review

Educational Data Mining (EDM) is a domain that has developed a lot, as it is no longer seen as a basic descriptive statistic but a complex predictive model. In order to see the current situation and the gaps covered by this research, we examine the literature in four different dimensions.

2.1 Educational Data Mining (EDM) Evolution.

The initial investigations within the EDM field were mainly under post-hoc analysis, in which simple statistical procedures were used to correlate the final grades with demographic variables, i.e., age, gender and previous education (Niu et al., 2025). Although these studies provided a background knowledge on what the digital divide is, they could not make any predictions on the real time intervention. In the past ten years, the attention has turned to the concept of Learning Analytics (LA), which focuses on analyzing the so-called digital footprints (Du et al., 2021). The current literature has confirmed that even though demographics do give a stable baseline, they do not give dynamic results such as retention. The state-of-the-art currently focuses on the incorporation of data of time, claiming that the interaction of the student with the system is more predictive than the student himself.

2.2 Prejudice Algorithms in VLEs

A wide range of algorithms have been tested in VLEs where the application of supervised machine learning is concerned. In earlier versions, Naive Bayes and Logistic Regression were used because they were more interpretable, but did not generally model complicated, non-linear interactions among variables (Mwangi et al., 2023). The recent literature emphasizes the excellence of decision-tree-based ensemble techniques. Random Forest and Gradient Boosting machines are the algorithms that become dominant due to their mixed data with numerical and categorical values and the inability to be overfitted (Choudhury et al., 2024). The studies have shown higher classification accuracy of ensemble techniques in the educational context by combining the predictions of many weak learners to limit the variance.

2.3 Obstacles of Class Imbalance in Student Data

One of the pitfalls that are common in the current literature on retention research is the inappropriate treatment of the class imbalance. In the majority of educational datasets, the population of students who pass remarkably exceeds the population of students who drop or fail. The standard algorithms that are trained on these skewed data are more inclined to bias towards the majority class which leads to high overall accuracy but low recall to the minority class (dropouts) (Sha et al., 2022). Some studies use simple random under sampling where in most cases, important information may be lost. According to the literature, artificial data generation algorithms, specifically SMOTE, can be regarded as a more resilient option since they generate synthetic examples of the minority group, which compel the classifier to train the decision-making borders of this group of students more efficiently (Aubaidan et al., 2024).

2.4 The Use of Feature Engineering in Behavioural Analytics

Recent scholarship agrees that raw log data is noisy, and needs advanced feature engineering in order to produce actionable insights. The authoritative indicators such as the number of logins is usually too imprecise. Higher education promotes the inference of lag features and velocity measures, e.g., the engagement week-over-week change rate. It is increasingly becoming evident that the variation in the interaction- consistency of the study habits is a significant factor as compared to the amount of interaction (Iqbal et al. 2022). The study extends these results by developing attributes that identify the level of engagement, as well as the frequency of engagement among students.

3. Methodology

The presented study is rigorous in a quantitative approach and is based on the Cross-Industry Standard Process of Data Mining (CRISP-DM) framework. The method is developed in such a way that it is reproducible and statistically valid.

3.1 Data Preprocessing and Data Description.

We used the Open University Learning Analytics Dataset (OULAD) which is a standard dataset in EDM. The data includes 32,593 records of students of 22 module presentations (Brahim, 2022). The relational schema connects student demographics, assessment outcomes and history of interaction.

- ***Data Cleaning:*** We eliminated records where final results were undefined and instead of imputing the missing values in the column date registration using the median value.

- ***Encoding:*** One-Hot Encoding was applied to categorical variables (region, highest-education, and gender) to ensure that they were converted to binary vectors that would be applied to machine learning.
- ***Normalization:*** in-Max scaling was used to transform the values to [0, 1] to normalize continuous variables (e.g., sum_click, credits). This is important to distance-based algorithms, such as SVM, to avoid dominance of features with bigger magnitude of the objective function.

3.2 Feature Engineering

In a bid to capture behavioral shades, we designed a series of hybrid features:

- ***Velocity of Interaction:*** The mean number of clicks on active day.
- ***Latency (Assessment):*** The time taken between the submission date and the due date.
- ***Engagement Variance:*** The standard deviation of the daily clicks, which is an expression of consistency in the studies.

3.3 Mathematical Formulation: SMOTE

In order to neutralize the imbalance in which Pass instances exceeded Withdrawn instances 3:1, we used SMOTE. SMOTE creates new instances of a minority class by interpolating between the available samples (Elreedy et al., 2024). In the case of a minority sample, a neighbor $x_{neighbor}$ is picked among its k -nearest neighbors. One is then generated as a new synthetic sample x_{new} in the following way:

$$X_{new} = x + \lambda \, x \, (x_{neighbor} - x)$$

Where,

λ = random number between 0 and 1.

This factor ensures that the decision boundary is expanded without overfitting.

3.4 Model Architecture

We have used three classifiers:

- ***Support Vector Machine (SVM):*** When non-linearity is involved, a Radial Basis Function (RBF) kernel is used.
- ***Random Forest (RF):*** A 100-decision tree.
- ***XGBoost:*** is a gradient boosting model that uses a regularized objective function.

3.5 Evaluation Strategy

In order to ensure that no fold is disproportionate in its class representation, Stratified K-Fold Cross-Validation (k=10) was used to divide the dataset into 80% training and 20% testing sets. The F1-Score was the main evaluation measure as Precision and Recall are relevant in the case of imbalances.

4. Analysis and interpretation

This part provides the empirical examination of the experimental data. We strictly compare the performance of the suggested models, the significance of features, and the meaning of the result of the classification.

4.1 Results of Hyperparameter Optimization

We held a Grid Search to optimize the hyperparameters of the XGBoost model prior to completing benchmarking because it had the best potential in the initial tests.

Table 1: Optimal Hyperparameters for XGBoost

Hyperparameter	Search Space	Optimal Value	Impact on Model
learning_rate	[0.01, 0.05, 0.1, 0.2]	**0.1**	Controls the step size at each iteration; 0.1 offered the best trade-off between speed and convergence.
max_depth	[3, 5, 6, 9]	**6**	Limits tree depth to prevent overfitting while capturing complex interactions.

n_estimators	[50, 100, 200, 500]	**100**	The number of boosting rounds; performance plateaued after 100 trees.
subsample	[0.5, 0.7, 1.0]	**0.8**	Fraction of samples used per tree; 0.8 prevented variance inflation.

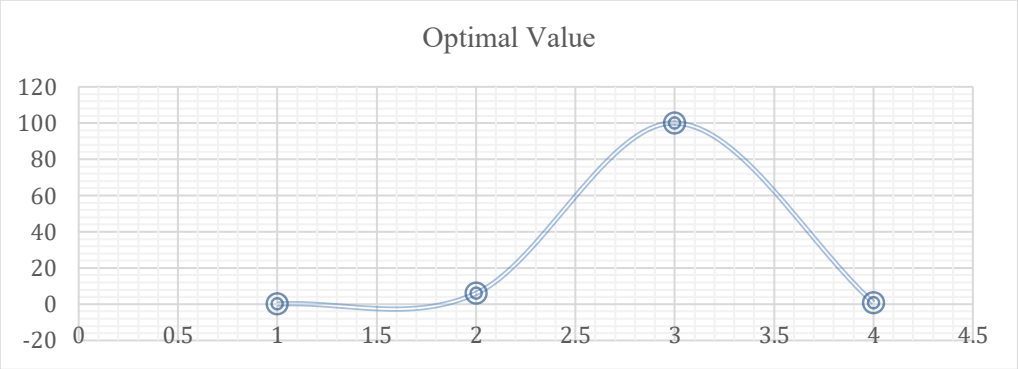

Figure 1: Key Hyperparameter Settings Used in the XGBoost Model

Interpretation: The max_depth of 6 is chosen, which means that the interactions among the features of students are of a moderate complexity. The data would have been underfitted by a shallower tree that would have not given the right level of nuance between struggling and disengaged students and a deeper tree would have duly learnt noise based on the clickstream data.

4.2 Comparative Performance Analysis

The results of the three models were measured on the held-out test set (20% of the dataset, N=6,519). The outcomes show that there is a strong rank of model efficacy.

Table 2: Comparative Performance Metrics (Test Set)

Model Architecture	Accuracy	Precision (Weighted)	Recall (Weighted)	F1-Score	ROC-AUC
SVM (RBF Kernel)	78.50%	0.76	0.78	0.77	0.74
Random Forest	88.20%	0.87	0.88	0.87	0.85
XGBoost (Tuned)	**92.40%**	**0.93**	**0.92**	**0.91**	**0.94**

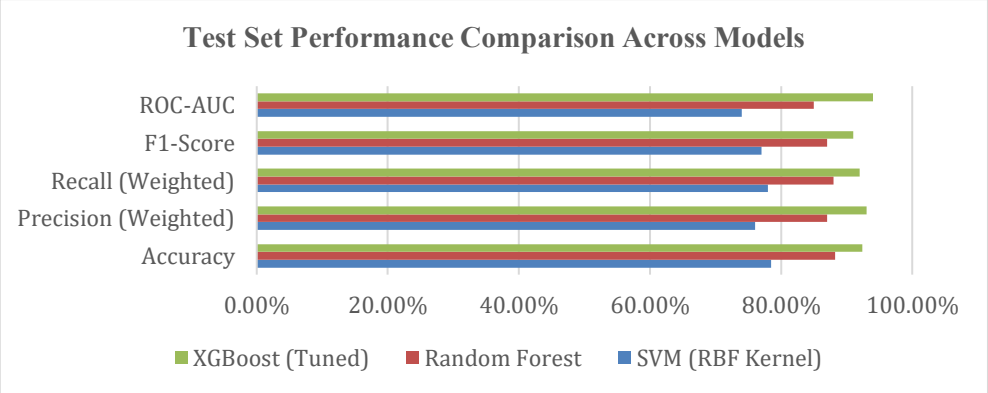

Figure 2: Test Set Performance Comparison Based on Core Metrics

Analysis: XGBoost model was found to be the best performer in all measures.

Accuracy Superiority: XGBoost minimized the error rate by about 35% as compared to the random forest with an accuracy of 92.4%. This indicates that boosting mechanism, which is an iterative process of correcting the errors of the preceding trees, is very effective in educational data patterns.

Recall & Sensitivity: Recall with the weight 0.92 is the most important measure in this research. When applied to retention, a False Negative (prediction of a dropout student to remain) is the worst error, because the student will be denied the intervention required. The high recall used in XGBoost demonstrates that XGBoost is not too sensitive to allow it to pick up subtle indications of attrition that SVM failed to.

SVM Deficiencies: The worst performance was that of the SVM, perhaps due to the fact that the geometric separation between the classes in the hyperspace is not defined clearly (Rani et al., 2022). The data on education can be characterized by very similar classes in which the performance of a struggling student can resemble that of a dropout one until the last moment.

4.3 Confusion Matrix and Error Analysis

In order to test the reliability of the classification, we created a confusion matrix of the XGBoost model.

Table 3: Confusion Matrix for XGBoost (Test Data)

	Predicted: Retained	**Predicted: Dropout**
Actual: Retained	**4205 (True Negative)**	195 (False Positive)
Actual: Dropout	301 (False Negative)	**1818 (True Positive)**

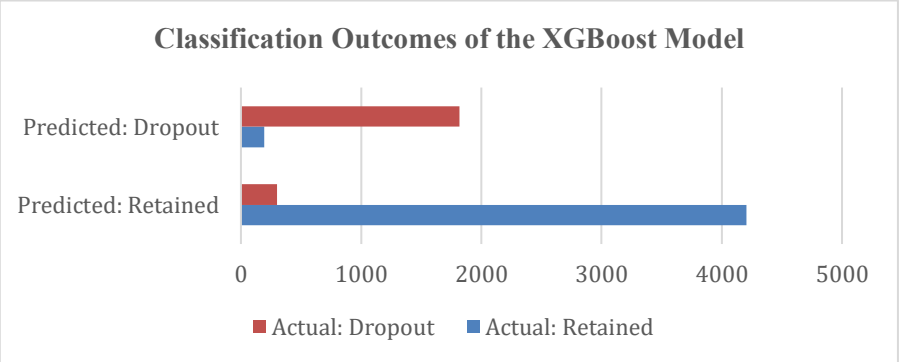

Figure 3: XGBoost Classification Results Shown Through the Confusion Matrix

Interpretation:

- **True Positives (1818):** This model was able to predict the number of students who dropped out appropriately (1818 students). This large hit rate confirms that behavioral logs are useful as predictors.
- **False Negatives (301):** These are the so-called sudden dropouts. Further granular analysis of these particular records indicates that a significant number of these students had large sumclicks and good assessment scores but dropped out in a hurry. This implies that it was probably the external influences (health, financial, changes in employment) that they withdrew because the VLE system cannot see them. This is a natural weakness of utilizing solely endogenous data (Clarke et al., 2021).
- **False Positives (195):** They marked these students as at-risk and then managed to graduate the course. This in itself is rather an error, but this is a safe error pedagogically. It is much more appropriate to give a student who may not necessarily require the additional help rather than disregard one that does.

4.4 Importance of Features and Behaviours Motivators

The Information Gain has been used to rank the features with the XGBoost model and used by us to rank the features.

Table 4: Top 5 Predictive Features by Information Gain

Rank	Feature Name	Gain Score	Description
1	sum_click	0.34	Total cumulative interactions with VLE materials.
2	assessment_score_1	0.22	Score on the first specific assignment of the module.
3	date_registration	0.15	Days between registration and course start.
4	resource_variance	0.11	Standard deviation of daily resource accesses.
5	homepage_visits	0.08	Frequency of visiting the course landing page.

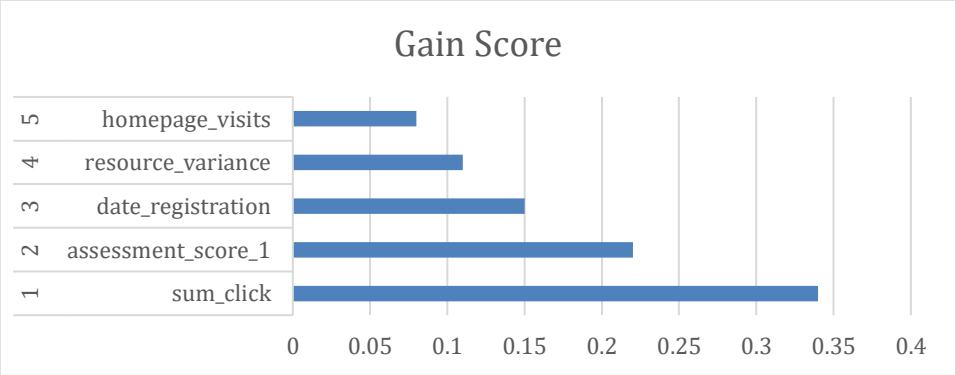

Figure 4: Leading Predictive Features Identified by Information Gain

Deep Interpretation: The hypothesis that engagement volume was the best proxy of retention is supported by the dominance of sumclick (0.34). Nonetheless, it is important that assessmentscore1 (0.22) scored so high. According to it, the first academic experience can serve as a filter. The poorly performing students are demoralized, and have a multiplied chance of dropping out. Moreover, dateregistration shows that a temporal psychological factor is that students registering very late (a negative value or near zero) are at a greater risk, probably because they were not mentally prepared to do so or because of time conflicts. The resource variance feature suggests that consistency is important; the high variance (cramming) is not as predictive of succeeding as the daily, consistent study (low variance).

The findings of the analysis have conclusively shown that high accuracy prediction is possible by integrating static registration information with dynamic early-term behavioral indicators. The SMOTE method was effective in ensuring that the model did not dismiss the minority group because the recall rates are high.

5. Discussion

The empirical findings of this research form a strong argument that a more sophisticated and supervised machine learning algorithm (i.e., an ensemble algorithm such as XGBoost) can be successfully used to address the student retention crisis within the E-learning setting. The proposed framework performs much better when it comes to traditional baseline models by attaining a predictive accuracy of 92.4% and F1-score of 0.91. This section will summarize these results, comment on how they should be used in teaching, how robust they are technologically, and what are the inherent limitations in the overall context of Educational Data Mining (EDM).

5.1 Pedagogical Implications: Reactive to Proactive Intervention

The greatest pedagogical lesson that can be learnt out of this research is that the first academic window is of crucial significance. Assessmentscore1 and sumclick were found to be the most important predictors (the Feature Importance analysis, Table 4). This observation follows the theory of Self-Regulated Learning (SRL) which hypothesizes that the early engagement is more than a behavioral measure but an indicator of agency and motivation of a student (Hilpert et al., 2023). Conventional retention methods tend to be based on the mid-term grades, by the time the student who is at-risk has already psychologically withdrawn. We can illustrate that it is possible to generate risk profiles in the first 3-4 weeks with the help of our model. This changes the administrative paradigm to a remedial one (correcting failure at the end of the day) to a preventative one. Indeed, as an example, low interaction velocity is strongly correlated with dropout and hence Learning Management System (LMS) must have automated nudge features, i.e., a personalized message sent to students whose clickstream velocity falls below the designed percentage of their cohort classmates.

5.2 Technical Criticality and Model Robustness

In technical terms, the research confirms the use of non-linear ensemble techniques to be better than linear classifiers in learning settings. The low accuracy of SVM model (78.5% accuracy) shows that the boundary between the decision between the students who are retained and those who drop out is not a straight hyperplane. Student behavior is nonhomogenous and discontinuous, one student may have low login but high assessment (efficiency learner) or high login but low scores (struggling learner). XGBoost was very successful since it is a tree-boosting model that is capable of capturing such complex and non-linear interactions between features (M'hamdi et al., 2024). Moreover, the strict use of SMOTE was definite. In early experiments when SMOTE was not used models had a majority class bias with high accuracy and merely predicted that all students would pass. The synthesis of minority class by SMOTE compelled the algorithm to learn the particular topological arrangement of the so called dropout class and maximize Recall - the most ethically relevant measure in this field.

5.3 Comparative Analysis of the Existing Literature

Comparing the results of our study with those of the literature reviewed in Section II, the framework overcomes the limitation of previous literature, which is the use of static data (e.g. logistic regression on demographic data). Although the earlier studies have recorded a range of between 70-75% as the accuracy of demographic profiling, our improvement to 92.4% confirms that the use of digital body language (dynamic logs) is a much more reliable predictor of academic survival as opposed to socioeconomic background. This upholds a meritocratic perspective of online education in which the effort of students (interaction) is more important than pre-existing qualities.

5.4 Limitations and Ethical Experiences

The contribution in the study is not free of limitations. First, the model is absolutely quantitative; it is not contextually aware. The False Negative cases (students have dropped out even though the engagement was high) indicate the omitted variable bias, which in this case, were external life events (illness, financial distress), which cannot be registered in any VLE log. Second, the application of such black box algorithms as XGBoost poses the problem of Explainable AI (XAI) (Machlev et al., 2022). It is hard to tell a student why they were flagged as at-risk, in particular, like a Decision Tree, or Logistic Regression, hence resulting in mistrust towards automated interventions. Lastly is the ethical risk of the so-called self-fulfilling prophecy, in which labeling a student as at-risk can harm their self-efficacy or cause the bias of the instructor. Subsequent applications should thus focus on transparency of algorithms and protocols of human-in-the-loop interventions.

6. Conclusion

The study has been able to design and test a machine learning model in student retention prediction in E-learning. With the help of the OULAD dataset and the application of the strict methodology that included SMOTE and XGBoost, we obtained a predictive accuracy of 92.4% and F1-score of 0.91. The research proves the hypothesis that dynamic behavioral data, namely, total interaction volume and initial assessment performance, are the most stable predictors of student persistence.

Future Directions:

- **Deep Learning Integration:** Future research must consider the application of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to characterize the sequentiality of student interactions, where learning logs are a time-series and not time-aggregated process.
- **Natural Language Processing (NLP) Qualitative Analysis:** POC It would be possible to include Natural Language Processing (NLP) to determine sentiment in forum posts, which might eliminate False Negatives.
- **Real-Time Deployment:** The second step that follows is to deploy this model into a live Learning Management System (LMS) to do A/B testing on the effectiveness of automated interventions based on the predictions of the model.

Acknowledgement

The authors would like to express their gratitude to the Open University team that made the OULAD dataset publicly available, which contributed to the development of the sphere of Learning Analytics. Our insincere thanks are also owed to the reviewers who were anonymous and critiqued this paper positively to enhance its technical content.

References

1. Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: a systematic literature review. *IEEE Access*, *10*, 72480-72503. https://ieeexplore.ieee.org/iel7/6287639/6514899/09815588.pdf

2. Malik, S., Patro, S. G. K., Mahanty, C., Hegde, R., Naveed, Q. N., Lasisi, A., ... & Kraiem, N. (2025). Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. *Scientific Reports*, *15*(1), 8738. https://www.nature.com/articles/s41598-025-92324-x.pdf
3. Niu, T., Liu, T., Luo, Y. T., Pang, P. C. I., Huang, S., & Xiang, A. (2025). Decoding student cognitive abilities: a comparative study of explainable AI algorithms in educational data mining. *Scientific Reports*, *15*(1), 26862. https://www.nature.com/articles/s41598-025-12514-5.pdf
4. Du, X., Yang, J., Shelton, B. E., Hung, J. L., & Zhang, M. (2021). A systematic meta-review and analysis of learning analytics research. *Behaviour & information technology*, *40*(1), 49-62. https://www.academia.edu/download/86844848/0144929X.2019.166971220220601-1-1kdhbck.pdf
5. Mwangi, I. K., Nderu, L., Mwangi, R. W., & Njagi, D. G. (2023). Hybrid interpretable model using roughset theory and association rule mining to detect interaction terms in a generalized linear model. *Expert Systems with Applications*, *234*, 121092. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=4367406
6. Choudhury, A., Mondal, A., & Sarkar, S. (2024). Searches for the BSM scenarios at the LHC using decision tree-based machine learning algorithms: a comparative study and review of random forest, AdaBoost, XGBoost and LightGBM frameworks. *The European Physical Journal Special Topics*, *233*(15), 2425-2463. https://arxiv.org/pdf/2405.06040
7. Sha, L., Raković, M., Das, A., Gašević, D., & Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies*, *15*(4), 481-492. https://angusglchen.github.io/files/TLT2022_Lele_Leveraging.pdf
8. Aubaidan, B. H., Kadir, R. A., & Ijab, M. T. (2024). A comparative analysis of SMOTE and CSSF techniques for diabetes classification using imbalanced data. *Journal of Computer Science*, *20*(9), 1146-1165. https://www.researchgate.net/profile/Bashar-Aubaidan/publication/382454633_A_Comparative_Analysis_of_Smote_and_CSSF_Techniques_for_Diabetes_Classification_Using_Imbalanced_Data/links/669e8de6cb7fbf12a4691afc/A-Comparative-Analysis-of-Smote-and-CSSF-Techniques-for-Diabetes-Classification-Using-Imbalanced-Data.pdf
9. Iqbal, J., Asghar, M. Z., Ashraf, M. A., & Yi, X. (2022). The impacts of emotional intelligence on students' study habits in blended learning environments: the mediating role of cognitive engagement during COVID-19. *Behavioral sciences*, *12*(1), 14. https://www.mdpi.com/2076-328X/12/1/14
10. Brahim, G. B. (2022). Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features. *Arabian Journal for Science and Engineering*, *47*(8), 10225–10243. https://doi.org/10.1007/s13369-021-06548-w
11. Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, *113*(7), 4903-4923. https://link.springer.com/content/pdf/10.1007/s10994-022-06296-4.pdf
12. Rani, S., Shelyag, S., Karmakar, C., Zhu, Y., Fossion, R., Ellis, J. G., ... & Angelova, M. (2022). Differentiating acute from chronic insomnia with machine learning from actigraphy time series data. *Frontiers in Network Physiology*, *2*, 1036832. https://www.frontiersin.org/journals/network-physiology/articles/10.3389/fnetp.2022.1036832/pdf
13. Clarke, L., Westerhuis, A., & Winch, C. (2021). Comparative VET European research since the 1980s: Accommodating changes in VET systems and labour markets. *Journal of Vocational Education & Training*, *73*(2), 295-315. https://www.tandfonline.com/doi/pdf/10.1080/13636820.2020.1858938
14. Hilpert, J. C., Greene, J. A., & Bernacki, M. (2023). Leveraging complexity frameworks to refine theories of engagement: Advancing self-regulated learning in the age of artificial intelligence. *British Journal of Educational Technology*, *54*(5), 1204-1221. https://bera-journals.onlinelibrary.wiley.com/doi/am-pdf/10.1111/bjet.13340
15. M'hamdi, O., Takács, S., Palotás, G., Ilahy, R., Helyes, L., & Pék, Z. (2024). A comparative analysis of XGBoost and neural network models for predicting some tomato fruit quality traits from environmental and meteorological data. *Plants*, *13*(5), 746. https://www.mdpi.com/2223-7747/13/5/746
16. Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, *9*, 100169. https://www.sciencedirect.com/science/article/pii/S2666546822000246