Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Identifying at-Risk Students with Data Analytics and Machine Learning: Insights from a Systematic Review

Patrick Ngulube[1]*, Mthokozisi Masumbika Ncube[2]

[1]*Department of Interdisciplinary Research and Postgraduate Studies, University of South Africa, Pretoria, South Africa, Email: ngulup@unisa.ac.za, Email: http://orcid.org/0000-0002-7676-3931
[2]Department of Interdisciplinary Research and Postgraduate Studies, University of South Africa, Pretoria, South Africa, Email: ncubemm@unisa.ac.za, https://orcid.org/0000-0003-4835-6594
***Corresponding Author:** Patrick Ngulube
*Department of Interdisciplinary Research and Postgraduate Studies, University of South Africa, Pretoria, South Africa, Email: ngulup@unisa.ac.za, Email: http://orcid.org/0000-0002-7676-3931

ARTICLE INFO

Received: 12 Dec 2024
Revised: 18 Feb 2025
Accepted: 26 Feb 2025

ABSTRACT

The purpose of this systematic review was to evaluate how data analytics and machine learning methods are used in Higher Education Institutions (HEIs) to forecast student achievement and identify students who may face academic difficulties. Although these methods have been studied before, there has been inadequate analysis of their strengths and limitations. Furthermore, considering the worldwide commitment to accomplishing the Sustainable Development Goals (SDGs), especially SDG 4, which prioritises inclusive and equitable quality education, tackling problems like student attrition and unequal support is essential. To fill this research gap, a systematic literature review including four major databases was carried out in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) criteria. Studies that used data analytics and machine learning methods to detect at-risk students or identify student achievement in higher education settings were included. To guarantee the rigour of the included studies, a quality assessment utilising the Critical Appraisal Skills Programme (CASP) Checklist was used. The review revealed a wide variety of strategies, including more sophisticated approaches like ensemble methods and neural networks, as well as more conventional ones like logistic regression and support vector machines. These methods were used on a variety of data sources, such as survey data, administrative data, and data from learning management systems. The results demonstrated how data analytics and machine learning may transform higher education by making it easier to identify at-risk students early on, customising support services to meet their requirements, and allocating resources as efficiently as possible. HEIs can therefore use these technologies to make data-driven decisions that will improve teaching and learning methods and increase student achievement in sustainable ways.

Keywords: data analytics, identifying at-risk students, machine learning approaches, predicting academic success, Higher Education Institutions, systematic review

1 Introduction

Achieving the Sustainable Development Goals (SDGs) of the United Nations (UN) by 2030 is a critical concern for the world community (Ncube & Ngulube, 2024). The SDGs cover a wide range of global concerns, such as climatic instability, environmental degradation, and socioeconomic inequality (Le Blanc, 2014; United Nations, 2015). A comprehensive, multidisciplinary approach is necessary to end poverty, slow down climate change, and advance high-quality education (United Nations, 2015). To achieve SDG 4, higher education institutions (HEIs) are essential for providing accessible, equitable, high-quality education and opportunities for lifelong learning (United Nations, 2015).

However, there are a lot of challenges facing HEIs, like high rates of student attrition, delayed graduation, and differences in academic performance amongst different student populations (Aina et al., 2019; Liaison, 2024; Maher & Macallister, 2013; Tamrat, 2022). Conventional approaches to determining whether students need assistance mostly depend on subjective evaluations, small datasets, and retroactive analysis (Ncube & Ngulube, 2024). This can exacerbate already-existing disadvantages by leading to unequal distribution of support resources, delayed interventions, and inaccurate identification of students who are at risk (Bhutoria, 2022; Hung et al., 2015; Osborne & Lang, 2023). These problems affect people and organisations, as well as the achievement of the SDGs (United Nations, 2015).

Global data highlight the urgent issue of student attrition and dropout rates, which impact both developed and developing countries (Tamrat, 2022; UNESCO, 2020). In developing nations, 20% of students drop out of their higher education programmes. Certain student populations are disproportionately affected by the disparities that still exist in industrialised nations (Ressa & Andrews, 2022). These patterns highlight how urgently creative approaches are needed to improve student achievement and fair access to higher education.

Using statistical and computational techniques to analyse large datasets methodically, data analytics and machine learning present promising answers (Abdul-Jabbar & Farhan, 2022; Caspari-Sadeghi, 2023; Sivarajah, 2017; TechTarget, n.d.). Harnessing data-driven insights, higher education institutions (HEIs) can:

- Enhance student progress monitoring,
- Identify at-risk students promptly,
- Allocate resources efficiently (Bozkurt & Sharma, 2022; Nguyen, et al., 2020).

This strategy supports sustainable patterns of production and consumption, which is in line with Sustainable Development Goal 12 (SDG 12) (United Nations, 2015). Within HEIs, data analytics and machine learning maximise resource use and reduce waste (Ncube & Ngulube, 2024).

A comprehensive understanding of student behaviours and academic outcomes is made possible by the application of various data analytics methodologies, including descriptive, predictive, prescriptive, text, and social network analyses, as well as machine learning algorithms and data visualisation (Ncube & Ngulube, 2024). Predictive analytics and machine learning, which are frequently used to predict academic success and identify at-risk learners, are the focus of this study. Educators can identify students who are at risk of academic underperformance or dropout by using predictive analytics, which uses previous data to estimate future academic outcomes (Kumar & Garg, 2018; Brown et al., 2015; Diamant, 2024; Smithers, 2023). Early detection facilitates targeted interventions, significantly enhancing student success rates (Waheed et al., 2020). To identify students at risk, predictive models use past data points, including course choices, engagement metrics, and academic performance. Large student databases are analysed by machine learning algorithms, which reveal patterns and trends that point to academic success or failure (Jagwani, 2019; Pinto, 2023; Yağcı, 2022). This makes it possible for educators to offer individualised assistance, which improves student achievement (Yağcı, 2022).

Higher education settings have used a variety of data analytics and machine learning techniques to identify susceptible students and forecast academic achievement (Järvinen et al., 2021; Rane et al., 2024). Research topics, dataset properties, and intended results must all be carefully considered when choosing an approach. Traditional logistic regression and support vector machines continue to be widely used despite the variety of these methods because of their interpretability, effectiveness, and adaptability (Bhumireddy & Anala, 2022; Johnson et al., 2024). In recent years, ensemble approaches-which combine several models-have become more popular (Kumari et al., 2018). These methods increase projected accuracy by reducing overfitting and improving generalisation by leveraging the distinct characteristics of each model (Zhang et al., 2021). In higher education, predictive analytics and machine learning make use of a range of data sources, such as survey results, gradebook data, administrative records, and data from educational technologies. Student demographics, academic standing, and disciplinary histories are all included in administrative data (Kok et al., 2024). Important insights can be gained from educational technology data from Massive Open Online Courses (MOOCs), learning management systems (LMS), and e-books (Asrowi et al., 2019; Reinhold et al., 2020). Multiple sources are integrated via institutional data marts, which analyse student conduct and forecast academic results (Kurniawan & Halim, 2013). Survey results help forecast student happiness and engagement by providing insight into their attitudes, beliefs, and perceptions (Chili & Madzimure, 2022; Kandiko-Howson & Matos, 2021).

Ensuring reliability and validity in predictive analytics and machine learning approaches in higher education necessitates careful metric selection (Kroese et al., 2024). While accuracy is commonly employed, its limitations, particularly with imbalanced datasets, underscore the importance of complementary metrics (Hastie et al., 2019; Koehrsen, 2024). A comprehensive evaluation framework incorporates classification metrics such as F1-score, precision, recall and AUC-ROC, which assess model performance, especially with imbalanced datasets (Gonçalves et al., 2014; Luque et al., 2019; Richardson et al., 2024). Additionally, regression metrics like Root Mean Squared Error (RMSE) quantify prediction errors (Sharma et al., 2022). Educational performance metrics, including completion rate, engagement rate and withdrawal rate, provide further insights into model effectiveness and student behaviour (Vărzaru et al., 2021).

Although the use of these approaches in HEIs has been studied in the past, there hasn't been a comprehensive study that synthesises the body of research and gives an overview of how they are applied. By reviewing pertinent material, highlighting important findings, and pointing out areas for further investigation, this study filled this knowledge gap. Therefore, this study aimed to synthesise existing research on data analytics and machine learning approaches to support students by looking at how integrated approaches can get a more thorough picture of student academic issues and identify at-risk students. This synthesis offers insightful suggestions for academics, educators, and policymakers on how to best use data analytics and machine learning techniques to identify at-risk students and forecast academic performance.

This research sets itself apart by going beyond theoretical discussion to provide useful suggestions and actionable strategies for applying data analytics and machine learning techniques in a higher education setting. In this context, the following objectives framed this systematic review:

1. Identify the various data analytics and machine learning techniques used in higher education to detect at-risk students and forecast academic performance.
2. Examine the different data sources used in machine learning and predictive analytics techniques to identify at-risk students and predict academic success in higher education institutions.
3. Assess how data analytics and machine learning techniques can be used to detect at-risk students and forecast academic achievement in higher education institutions.

This review concentrated on these objectives as they are vital for informing decision-making among researchers, educators, and policymakers regarding the implementation of data analytics and machine learning. Therefore, by determining which strategies work best, institutions can choose and apply strategies that are most likely to produce the best outcomes. Understanding the variety of ways that are accessible is essential for institutions because relying only on one might limit efficacy. Furthermore, identifying the critical elements affecting the precision of machine learning and predictive analytics techniques enables organisations to streamline their data gathering and analysis procedures. Making wise implementation decisions requires an understanding of the ramifications of data analytics and machine learning in higher education. Thus, the results of this review aid in the formulation of policies and procedures controlling the application of machine learning and data analytics techniques in HEIs. Through these focused objectives, this review provides a resource for researchers, educators, and policymakers seeking to enhance student outcomes and institutional effectiveness through data-driven approaches.

2 Methods

To maintain methodological rigour and transparency, this systematic review strictly adhered to the reporting guidelines outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (Page, et al., 2021).

2.1 Literature Search and Selection

A systematic search was conducted across four databases, namely, ACM Digital Library, ERIC, Scopus, and Web of Science, to identify relevant literature on predicting academic success and identifying at-risk students in higher education. The search strategy combined keywords and subject headings related to higher education, academic success, at-risk students, data analytics, and machine learning. The following syntax for each database was formulated: ACM Digital Library: (("higher education" OR "universities" OR "colleges") AND (("academic success" OR "academic achievement" OR "student retention") AND ("at-risk students")) AND (("data analytics" OR "learning analytics" OR "educational data mining") OR ("predictive modeling" OR "machine learning algorithms"))). ERIC: ((higher education OR universities OR colleges) AND (academic success OR academic achievement OR student retention OR at-risk students) AND (data analytics OR learning analytics OR educational data mining OR predictive modeling OR machine learning algorithms)) OR TI=("predicting academic success") OR AB=("data analytics" OR "machine learning" OR "higher education"). Scopus: TITLE-ABS-KEY (((("higher education" OR "universities" OR "colleges") AND ("academic success" OR "academic achievement" OR "student retention" OR "at-risk students")) AND ("data analytics" OR "learning analytics" OR "educational data mining" OR "predictive modeling" OR "machine learning algorithms"))). Web of Science: TS= (("higher education" OR "universities" OR "colleges") AND (("academic success" OR "academic achievement" OR "student retention" OR "at-risk students") AND ("data analytics" OR "learning analytics" OR "educational data mining" OR "predictive modeling" OR "machine learning algorithms. To refine the search, the Population, Intervention, Comparison, Outcome (PICO) framework informed the search strategy and eligibility criteria, yielding peer-reviewed journal articles, conference papers and chapters (Methley et al., 2014). The PICO framework facilitated study selection by guiding the identification of relevant literature. As such, the following eligibility criteria were established:

1. Population (P): Focus was on students enrolled in HIEs, with priority given to graduate students (Master's or Doctoral programmes). While studies on undergraduate students were considered, those exclusively focused on primary, secondary, or vocational education were excluded.
2. Intervention (I): Emphasis was on the application of data analytics and machine learning approaches in higher education. Interventions focused on using data analytics and machine learning approaches for academic success prediction or early warning systems were included. Studies solely focused on the technical development of data analytics without an educational application were excluded.
3. Comparison (C): Attention was on studies comparing the application of data analytics and machine learning interventions to traditional support methods or a control group with no intervention. Baseline measures of academic achievement before intervention are also considered.
4. Outcome (O): Emphasis was on academic success outcomes relevant to higher education students, including academic performance (grades, GPA), student retention and graduation rates, time-to-degree completion, academic progress (credit and course completion), and early warning indicators for at-risk students. Table 1 outlines the additional criteria applied in conjunction with the PICO framework:

Table 1: Additional Eligibility Criteria

Category	Sub-Criteria	Description
Publication	Language	English
	Date	2011 – 2024 (inclusive)
	Type	Peer-reviewed journal articles, conference papers, and chapters
Methodology	Type	Quantitative, Mixed methods
	Requirements	Clear and transparent methodology and data analysis
Context	Education Level	Higher education institutions (universities/colleges)
	Data Analytics and Machine Learning Focus	Explicit discussion of data analytics and/or machine learning
Availability	Full-Text	Accessible full-text versions
Intervention	Details	Structure, data types, and student support activities
Research	Quality	Demonstrates strong data analytics and machine learning expertise, rigorous methodology, clear sampling, and valuable insights into enhancing student success and identifying at-risk students.

The literature search identified 2414 initial studies. To streamline the review process and ensure high-quality selection, a two-stage approach was implemented using the advanced software tools Rayyan and ASReview (ASReview, 2024; Rayyan, 2024). In the first stage, Rayyan was used to remove duplicate and non-English studies, resulting in 1636 remaining articles (Rayyan, 2024. Subsequently, ASReview analysed these titles and abstracts using pre-defined criteria, prioritising the most relevant studies for full-text review (ASReview, 2024). In the second stage, Rayyan was used to thoroughly review the prioritised articles. Applying eligibility criteria resulted in the exclusion of 1559 studies, identifying 77 key studies that directly addressed the context of the study for further quality assessment.

2.2 Quality Assessment Using the CASP Checklist

A quality assessment was conducted using the Critical Appraisal Skills Programme (CASP) checklist (CASP, 2024) to ensure a rigorous and standardised evaluation of the selected studies. This framework assessed four key dimensions:

1. Conceptual Clarity and Comprehensiveness: Reviewers evaluated the authors' understanding of data analytics and machine learning applications in predicting academic success and identifying at-risk students.

2. Methodological Rigour: The checklist guided the assessment of research design, data collection methods, data analysis techniques, and the mitigation of potential biases in predicting academic outcomes.
3. Sampling Clarity: Reviewers assessed the clarity of the target student population, sampling methods, and justification for sample size and representativeness within higher education settings.
4. Valuable Findings: The checklist focused on the clarity of findings, their alignment with research questions or objectives and methodology, and the depth of discussion regarding implications for academic success and at-risk student identification.

The CASP checklist ensured a systematic evaluation of each article's strengths and weaknesses across these four dimensions. This approach led to the selection of 33 high-quality research articles that employed data analytics and machine learning approaches to predict academic success and identify at-risk students. To ensure a consistent and unbiased screening process, inter-rater reliability (IRR) measures were established. A high concordance rate of 85% and a kappa coefficient of 0.80 demonstrated strong agreement between reviewers. Figure 1 presents the PRISMA flow diagram, illustrating the systematic search and study selection methodology.

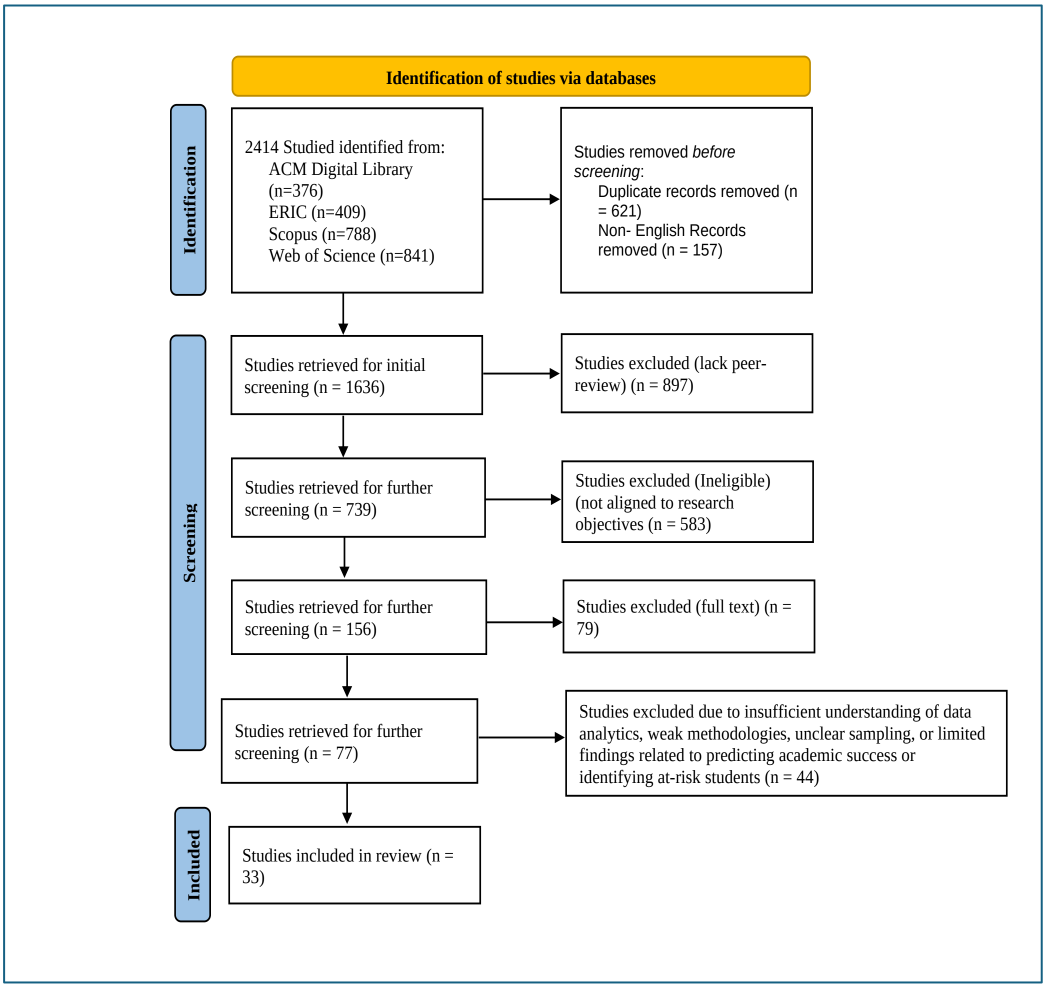
Identification of studies via databases
Identification
2414 Studied identified from:
ACM Digital Library (n=376)
ERIC (n=409)
Scopus (n=788)
Web of Science (n=841)
Studies removed *before screening*:
Duplicate records removed (n = 621)
Non- English Records removed (n = 157)
Screening
Studies retrieved for initial screening (n = 1636)
Studies excluded (lack peer-review) (n = 897)
Studies retrieved for further screening (n = 739)
Studies excluded (Ineligible) (not aligned to research objectives (n = 583)
Studies retrieved for further screening (n = 156)
Studies excluded (full text) (n = 79)
Studies retrieved for further screening (n = 77)
Studies excluded due to insufficient understanding of data analytics, weak methodologies, unclear sampling, or limited findings related to predicting academic success or identifying at-risk students (n = 44)
Included
Studies included in review (n = 33)

Figure 1: Systematic Literature Search and Selection Process (Adapted from PRISMA 2020 Guidelines)

2.3 *Data Extraction and Coding*

Data extraction adhered to PRISMA guidelines (Page et al., 2021). A standardised form developed using CADIMA (2024) ensured consistency and efficiency. The form captured study characteristics, research design, data analytics and machine learning approaches, key findings, and author-identified limitations. Manual coding was then used to extract and document crucial information, including author(s), research design, publisher, and findings. The extracted data is presented in Table 2.

Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2: Data Extraction and Coding

Author(s)	**Design**	**Publisher**	**Findings**
Ahmad et al. (2019)	Quantitative Predictive Analytics Through Case Study and Classification Problem	Springer	Identified significant attributes for student attrition prediction.
Akçapınar et al. (2019)	Quasi-Experimental Design	International Journal of Educational Technology in Higher Education	Accurately predicted unsuccessful students at the end of the term (89%). Early prediction (3 weeks) was possible with 74% accuracy.
Akçapınar et al. (2019)	Quantitative Predictive Analytics Design	Smart Learning Environments	Early prediction system using eBook-based teaching-learning.
Alturki, et al. (2022)	Experiment Using Data Mining Algorithms on a Limited Dataset of Easily Accessible Features from Students	Smart Learning Environments	The most significant features are semester grades, distance from accommodation, and culture.
Anagnostopoulos, et al. (2020)	Quantitative Predictive Analytics	Springer	Developed a prediction model using Learning Management System data to identify at-risk students. Improved performance.
Azizah, et al. (2024)	Quantitative Quasi-Experimental	Artificial Intelligence	Analysed motivational variables to identify at-risk students in blended learning. Time-management variables were key factors.
Baashar, et al. (2022)	Quantitative Predictive Analytics	Alexandria Engineering Journal	Predicted academic performance of postgraduate students using ML algorithms. The Artificial Neural Network (ANN) model achieved the best performance (89% R2).
Bainbridge, et al. (2015)	Quantitative Predictive Analytics	Journal of Public Affairs Education	Used learning analytics to identify at-risk students in an online Master of Public Administration programme. Simple models achieved high sensitivity and low false positive rates.
Bañeres, et al. (2020)	Quantitative Predictive Analytics	Applied Sciences	Developed an early warning system to detect at-risk students in online higher education. Used a predictive model, established a threshold, and implemented an early warning system.
Chui, et al. (2018)	Quantitative Predictive Analytics	Computers in Human Behavior	Developed a probabilistic logistic regression model to identify at-risk students. Incorporated student engagement data, demographic data, and student performance data.
Herodotou, et al. (2020)	Quantitative Experimental Through Randomised Controlled Trial (RCT)	Journal of Learning Analytics	Used a predictive model to identify at-risk students and implemented motivational interventions. Found that interventions were effective in facilitating course completion.
Hu, et al. (2014)	Quantitative Predictive Analytics Through Longitudinal Study	Computers in Human Behavior	Developed an early warning system using data-mining techniques to predict at-risk students in an online undergraduate course.
Jia, et al. (2015)	Quantitative Predictive Analytics Through Longitudinal Study	Higher Education	Developed a predictive model to identify at-risk students using administrative data.

Koutina, et al. (2011)	Quantitative Predictive Analytics Through an Experimental Study	Springer	Investigated the most efficient machine learning approach for predicting the final grades of postgraduate students.
Latif, et al. (2022)	Quantitative Predictive Analytics Through an Experimental Study	International Journal of Computing and Digital Systems	Predicted student grades using Artificial Intelligence algorithms.
Lourens and Bleazard (2016)	Quantitative Predictive Analytics Through Case Study	South African Journal of Higher Education	Identified second-year dropout students using pre-university information and first-semester data. Used Konstanz Information Miner (KNIME) for predictive modeling.
Matz, et al. (2023)	Quantitative Predictive Analytics Design Through Multi-Institutional Study	Scientific Reports	Predicted student dropout using macro-level (socio-demographics, early performance) and meso-level (engagement data) variables.
Nimy, et al. (2023)	Quantitative Learning Analytics Design Through Multi-Stage Study	Applied Sciences	Developed a probabilistic logistic regression model to identify at-risk students at different stages.
Osborne and Lang (2023)	Quantitative Predictive Analytics Through Binary Classification Problem	Journal of Postsecondary Student Success	Used a neural network model to predict student failure based on grade book data.
Paterson and Guerrero (2019)	Quantitative Predictive Analytics Through Binary Classification Problem	Research in Higher Education Journal	Compared logistic regression and discriminant analysis for predicting student retention.
Pecuchova and Drlik (2023)	Quantitative Predictive Analytics Through Classification Problem	Procedia Computer Science	Evaluated ensemble methods for early prediction of at-risk students.
Pek, et al. (2023)	Quantitative Predictive Analytics Through Classification Problem	IEEE Access	Developed a hybrid ensemble model to predict at-risk students.
Psathas, et al. (2023)	Quasi-Experimental Study	Computers	Examined factors contributing to early prediction of Massive Open Online Courses (MOOCs) dropouts.
Queiroga, et al. (2020)	Quantitative Predictive Analytics Through Case Study and Classification Problem	Applied Sciences	Developed a solution using student interactions with the virtual learning environment to predict at-risk students. Applied a genetic algorithm for hyperparameter tuning.
Qushem, et al. (2024)	Quantitative Predictive Analytics Through Regression/Classification Problem	Technology, Knowledge and Learning	Developed an early warning prediction model using computing science degree performance data.
Ramaswami, et al. (2023)	Quasi-Experimental Study	Journal of Learning Analytics	Evaluated a learning analytics dashboard (SensEnablr) for its effectiveness in impacting student learning.
Russell, et al. (2020)	Quantitative Quasi-Experimental Design	Computers & Education	Examined the effects of a learning analytics platform (Elements of Success) on at-risk students.
Seo, et al. (2024)	Quantitative Predictive Analytics Design, Classification Problem	Heliyon	Developed a prediction model using real-world data from an online university. Identified significant dropout indicators.
Smirani, et al. (2022)	Quantitative Predictive Analytics Through Classification Problem	Scientific Programming	Proposed a Stacked Generalization for Failure Prediction (SGFP) model to improve student results.

Syed Mustapha (2023)	Quantitative Comparative Study Through Regression and Classification Problems	Applied System Innovation	Compared feature engineering and selection methods for predicting student success.
Varade and Gupta (2024)	Quantitative Predictive Analytics Through Classification Problem	Journal of Electrical Systems	Developed a deep neural network model to predict student academic performance.
Wan Yaacob, et al. (2020)	Quantitative Predictive Analytics Through Classification Problem and Experiment	Journal of Physics: Conference Series	Used data mining techniques to predict student dropout in Computer Science.
Yukselturk, et al. (2014)	Quantitative Predictive Analytics Through Classification Problem and Experiment	European Journal of Open, Distance and E-Learning	Predicted dropouts using data mining approaches in an online program.

Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

3 Results

This section synthesises the review's findings through a critical analysis, assessing their relevance and contribution to achieving the research objectives. Specifically, it examines data analytics and machine learning approaches predicting academic success and identifying at-risk students in higher education (Objective 1), summarising results in Table 3.

Table 3: Data Analytics and Machine Learning Approaches for Predicting Academic Success and Identifying At-Risk Students

Author(s)	**Approach**	**Evaluation Metric**
Baashar et al. (2022)	Neural Network	Root Mean Square Error
Osborne et al. (2023)	Neural Network	Accuracy
Hu et al. (2014)	Classification and Regression Trees, Adaptive Boosting	Accuracy
Yukselturk et al. (2014)	Data Mining	Accuracy
Varade et al. (2024)	Deep Neural Network	Accuracy
Anagnostopoulos et al. (2020)	Ensemble	F1-score
Matz et al. (2023)	Ensemble	Area Under the Receiver Operating Characteristic Curve
Pecuchova et al. (2023)	Ensemble	Area Under the Receiver Operating Characteristic Curve
Syed Mustapha (2023)	Feature Engineering	Coefficient of Determination
Queiroga et al. (2020)	Genetic Algorithm	Area Under the Receiver Operating Characteristic Curve
Pek et al. (2023)	Hybrid Ensemble	Accuracy
Akçapınar et al. (2019)	k-Nearest Neighbor	Precision, Recall
Herodotou et al. (2020)	Learning Analytics	Completion Rate
Ramaswami et al. (2023)	Learning Analytics	Engagement
Russell et al. (2020)	Learning Analytics	Withdrawal Rate
Seo et al. (2024)	Light Gradient Boosting Machine	Accuracy
Azizah et al. (2024)	Logistic Regression	Coefficient of Determination
Bainbridge et al. (2015)	Logistic Regression	Sensitivity
Jia et al. (2015)	Logistic Regression	Odds Ratio
Paterson et al.	Logistic Regression	Odds Ratio
Qushem et al. (2024)	Logistic Regression	Accuracy
Wan Yaacob et al. (2020)	Logistic Regression	Odds Ratio
Koutina et al. (2011)	Naive Bayes	Accuracy
Lourens and Bleazard (2016)	Predictive Modeling	Sensitivity
Chui et al. (2018)	Probabilistic Logistic Regression	Area Under the Receiver Operating Characteristic Curve
Nimy et al. (2023)	Probabilistic Logistic Regression	Area Under the Receiver Operating Characteristic Curve
Bañeres, et al. (2020)	Quantitative Predictive Analytics	Accuracy
Akçapınar et al. (2019)	Random Forest	Accuracy
Alturki et al. (2022)	Random Forest	Area Under the Receiver Operating Characteristic Curve

Latif et al. (2022)	Sequential Minimal Optimisation, Logistic Regression	Accuracy
Smirani et al. (2022)	Stacked Generalisation	Accuracy
Ahmad et al. (2019)	Support Vector Machines (SVMs)	F1-score
Psathas, et al. (2019)	Support Vector Machines (SVMs)	Recall, Precision

Table 3 presents a comprehensive overview of the diverse data analytics and machine learning techniques that predict student performance. These techniques, from traditional methods like logistic regression to advanced ensemble methods, are applied to various data sources to forecast student outcomes. To evaluate their models, researchers employed a variety of metrics, including classification metrics (F1-score, precision, recall, accuracy, AUC-ROC) and regression metrics (R-squared, RMSE). Additionally, educational performance metrics such as sensitivity, odds ratio, completion rate, engagement, and withdrawal rate were used to assess student outcomes. These metrics provide valuable insights into the accuracy and effectiveness of the models in predicting student performance and identifying at-risk students.

The second research objective of this study was to examine the diverse data sources employed in predictive analytics and machine learning models designed to forecast academic success and identify at-risk students within higher education institutions. To achieve this, the present study analysed the data sources the reviewed studies utilised, summarising the results in Table 4.

Table 4: Diverse Data Sources Employed in Predictive Analytics and Machine Learning Models

Author(s)	**Data Source**
Ahmad et al. (2019), Alturki et al. (2022), Baashar et al. (2022), Chui et al. (2018), Hu et al. (2014), Jia et al. (2015), Koutina et al. (2011), Latif et al. (2022), Lourens and Bleazard (2016), Nimy et al. (2023), Paterson et al. (2019), Pek et al. (2023), Pecuchova et al. (2023), Qushem et al. (2024), Seo et al. (2024), Smirani et al. (2022), Syed Mustapha (2023), Varade et al. (2024), Wan Yaacob et al. (2020)	Administrative Data
Akçapınar et al. (2019)	eBook-based Teaching
Osborne et al. (2023)	Grade Book Data
Bañeres, et al. (2020)	Institutional data mart containing curated data from six semesters
Anagnostopoulos et al. (2020), Bainbridge et al. (2015), Herodotou et al. (2020), Ramaswami et al. (2023), Russell et al. (2020), Akçapınar et al. (2019)	Learning Management System
Psathas, et al. (2023)	MOOC data, including student demographics, interaction data, and self-reported self-regulated learning (SRL) data
Matz et al. (2023)	Multi-Institutional Data
Yukselturk et al. (2014)	Online Programme Data
Azizah et al. (2024)	Survey Data
Queiroga et al. (2020)	Virtual Learning Environment

Table 4 presents a diverse range of data sources, including administrative data, LMS data, survey data, institutional data marts, and specialised data sources like eBook-based teaching data, MOOC data, and online programme data, which were utilised in the reviewed studies to predict student performance and identify at-risk students. The third objective of this study was to explore the implications of data analytics and machine learning approaches for predicting academic success and identifying at-risk students in higher education institutions. The findings from the reviewed studies regarding this objective are summarised in Table 5.

Table 5: Implications of Data Analytics and Machine Learning for Predicting Academic Success and Identifying At-Risk Students

Key Finding	Implications for Higher Education Institutions	Example Studies
Early Identification of At-Risk Students	Proactive interventions to prevent academic failure and improve student success.	Akçapınar et al. (2019), Alturki et al. (2022), Anagnostopoulos et al. (2020)
Accurate Prediction of Academic Performance	Tailored support services and resource allocation based on individual student needs.	Azizah et al. (2024), Baashar et al. (2022), Bainbridge et al. (2015)
Identification of Diverse Factors Influencing Academic Success	A comprehensive approach to student support that addresses multiple factors contributing to academic performance.	Bañeres et al. (2020), Chui et al. (2018), Herodotou et al. (2020)
Customisation of Models to Institutional Contexts	Adaptability of data analytics solutions to the specific needs of different institutions.	Hu et al. (2014), Jia et al. (2015), Koutina et al. (2011)
Cost-Effectiveness of Data Analytics	Efficient use of resources and improved return on investment.	Latif et al. (2022), Ahmad et al. (2019)

Table 5 underscores the transformative potential of data analytics and machine learning in revolutionising higher education by enhancing student support, improving academic performance, and reducing attrition rates. Firstly, these technologies can identify at-risk students early on, allowing for timely interventions. The table shows that data analytics and machine learning approaches can accurately predict academic performance, enabling tailored support services. Thirdly, they can identify various factors influencing academic success, leading to a holistic approach to student support. Additionally, data analytics models can be customised to suit the specific needs of different institutions. Lastly, these technologies can be cost-effective by optimising resource allocation and improving student outcomes.

4 Discussion

The analysis of study results (Table 3) reveals substantial heterogeneity in the application of data analytics and machine learning approaches across the reviewed studies, highlighting the need to select appropriate approaches based on the specific research problem, dataset characteristics, and desired outcomes in predicting academic success and identifying at-risk students in higher education institutions (Järvinen et al., 2021; Rane et al., 2024). Traditional approaches such as logistic regression and support vector machines have emerged as the most used techniques in the investigated studies. This prevalence can be attributed to their interpretability, efficiency, and versatility in modelling binary outcomes and accommodating various predictor variables (Bhumireddy & Anala, 2022). This finding aligns with existing empirical research (Johnson et al., 2024; Bhumireddy & Anala, 2022), underscoring logistic regression's adaptability, ubiquity, and exceptional suitability for data analytics and machine learning applications focused on predicting academic success and identifying at-risk students. The enduring popularity of these traditional approaches can be attributed to their adaptability, interpretability, and ability to accommodate complex datasets.

The analysis of study results further underscores the widespread adoption of ensemble methods across the examined studies, highlighting their efficacy in predictive analytics for student success, thereby informing targeted interventions and supporting data-driven decision-making within higher education institutions. Consistent with these findings, Kumari et al. (2018) and Zhang et al. (2021) underscore the efficacy of ensemble approaches in augmenting predictive performance in higher education, attributable to their ability to harness the collective strengths of multiple models. This synergistic approach mitigates overfitting and enhances generalisation, optimising student outcome predictions (Kumari et al., 2018). Consequently, ensemble methods have gained widespread acceptance among scholars and higher education institutions, who increasingly recognise their potential to inform data-driven decision-making and improve student success.

Several of the reviewed studies explicitly focus on predictive modelling (Table 3), highlighting its prevalence as a common application of machine learning and data mining approaches, given their ability to identify patterns, extract insights, and make accurate predictions (Cui et al., 2019; Tete et al., 2022). Consistent with the present study's findings, Kuhn and Johnson (2013) and Tete et al. (2022) elucidate that predictive modelling leverages machine learning algorithms to systematically analyse large datasets, thereby uncovering intricate relationships between variables and facilitating evidence-based interventions for at-risk students and data-driven decision-making within higher education institutions. A key finding from the literature review is the effective application of learning analytics in predicting student success and identifying at-risk students by leveraging data about learners and their contextual factors within higher education. Joshi et al. (2020) and Ogata et al. (2022) contend that by harnessing learning analytics, scholars and educators can systematically analyse large datasets to identify early indicators of student struggle, predict academic performance, and inform targeted interventions.

A notable observation from Table 3 is the emergence of neural networks, particularly deep learning architectures, as a prominent approach employed in several studies to address the complex patterns and voluminous datasets inherent in predicting academic success and identifying at-risk students. Characterised by their multi-layered structure, deep learning models have exhibited exceptional proficiency in handling intricate relationships and high-dimensional data (Kovač et al., 2023; Lakkaraju, 2017; LeCun et al., 2015). This proficiency can be attributed to their capacity to learn hierarchical representations, extract salient features, and capture non-linear interactions (Na et al., 2017; LeCun et al., 2015). The application of feature engineering through system innovation in one study is noteworthy, transforming raw data into informative features to enhance predictive models' accuracy in identifying at-risk students and informing targeted support strategies. This approach extracts relevant information from large datasets (Kuhn & Johnson, 2013), reduces dimensionality, and improves model interpretability (Guan et al., 2020). In educational data mining, feature engineering identifies early warning indicators of student struggle (Abdulwahed et al., 2017), predicts student dropout and retention (Ifenthaler & Yau, 2020), and informs personalised learning interventions (Naseer et al., 2024; Rahiman & Kodikal, 2023). Hence, effective feature engineering informs student support by identifying key predictors of academic success, detecting changes in student behaviour, and facilitating timely interventions.

The efficacy of data analytics and machine learning models was contingent upon the judicious selection of evaluation metrics, which quantitatively assessed their performance (Kroese et al., 2024). The reviewed studies (Table 3) reveal a diverse array of evaluation metrics employed across various studies, each with distinct strengths and limitations. Accuracy, the most widely used metric (Table 3), measured the proportion of correctly classified instances. However, as noted by Hastie et al. (2019) and Koehrsen (2024), it may not always provide a comprehensive picture of model performance, particularly in imbalanced datasets. As such, other metrics such as F1-score, precision, and recall were often employed (Table 3). The F1-score, as validated by Powers (2011), offers a balanced measure of precision and recall, providing a more nuanced understanding of model performance. Hastie et al. (2009) further emphasised the complementary nature of precision and recall, with precision measuring the proportion of true positives among all positive predictions and recall measuring the proportion of true positives among all actual positive instances.

For binary classification models, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used to evaluate model performance (Table 3). Gonçalves et al. (2014) and Luque et al. (2019) noted that AUC-ROC plotted true positive rates against false positive rates, providing a thorough evaluation of model performance. Richardson et al. (2024) highlighted the usefulness of this metric in evaluating models on imbalanced datasets. Root Mean Squared Error (RMSE) was used for regression models to measure the difference between predicted and actual values, providing an absolute measure of model error (Sharma et al., 2022). In addition to these core metrics, Table 3 showed that other metrics, such as completion rate, engagement, and withdrawal rate, were employed to gain insights into user behaviour and model effectiveness (Vărzaru et al., 2021).

The study revealed diverse data sources employed in predictive analytics and machine learning models within the education sector. As Aldoseri et al. (2023) noted, these data sources are crucial for training and testing models, ultimately leading to more accurate predictions and informed decision-making. Administrative data emerged as the most frequently used data source (Table 4). This aligns with the findings of Kok et al. (2024), who highlighted the importance of administrative data, including student demographics, enrolment records, academic performance data, and disciplinary records, in predicting student success, dropout rates, and academic performance. eBook-based teaching data was another source employed in the reviewed studies (Table 4). This closely relates to the study of Asrowi et al. (2019) and Reinhold et al. (2020), which underscore the widespread use of eBooks in education and the potential of data derived from student interactions with eBooks to predict engagement and learning outcomes. Grade book data, providing insights into student performance on assignments and quizzes, was also utilised by several scholars (Table 4). Latif and Miles (2020) also explicate the extensive use of grade book data to identify at-risk students and provide timely interventions.

Institutional data marts, integrating data from multiple sources, were employed by various studies (Table 4). In that regard, Kurniawan and Halim (2013) elucidate the widespread use of institutional data marts to analyse student behaviour and predict academic outcomes. Learning Management System (LMS) data was also prevalent, providing information about student interactions with online courses (Table 4). Duin and Tham (2020) also underline the extensive use of LMS data to develop predictive student retention and engagement models. Massive Open Online Course (MOOC) data was another data source used to predict academic success and identify at-risk students. The study of He et al. (2015) also concurs with these findings as it notes the wide usage and need for MOOC data, including student demographics, interaction data, and self-reported learning data, to identify factors influencing student success and dropout rates. Online programme data, including information about student participation in online programmes, was also employed by several studies (Table 4). Shou et al. (2024) highlight the role of online programme data in predicting student performance and identifying at-risk students. Survey data, providing insights into student perceptions, attitudes, and beliefs, was another data source utilised (Table 4). Chili and Madzimure (2022) and Kandiko-Howson and Matos (2021) concur with these findings and underscore the need for and usage of survey data to predict student satisfaction and engagement.

The study revealed significant implications of data analytics and machine learning for predicting academic success and identifying at-risk students in higher education institutions (Table 5). The studies demonstrated the efficacy of early identification in enabling institutions to proactively implement targeted interventions, such as tutoring and academic advising. This proactive approach contributes to improved student success and retention rates, as supported by Shou et al. (2024) and Chili and Madzimure (2022). The study highlighted the importance of accurate prediction models in tailoring support services to individual student needs. By identifying students most likely to benefit from specific interventions, institutions can optimise resource allocation and ensure that support is provided where it is most needed, as noted by He et al. (2015) and Shou et al. (2024). The study emphasised the need for a holistic approach to student support that addresses multiple factors influencing academic performance. In line with this, Duin and Tham (2020), Kok et al. (2024), and Kurniawan and Halim (2013) highlight the importance of considering demographic information, academic engagement, and personal attributes in developing effective interventions.

Additionally, the study demonstrated the adaptability of data analytics and machine learning solutions to various institutional contexts. In synch with these findings, the studies of Aldoseri et al. (2023) and Duin and Tham (2020) emphasise the importance of customising models to account for specific institutional needs and student populations. The study highlighted the cost-effectiveness of data analytics and machine learning in improving student outcomes (Table 5). Thus, by identifying at-risk students early and providing targeted support, institutions can avoid the costs associated with student attrition and remediation efforts. As such, studies by Latif et al. (2022) and Ahmad et al. (2019) provide evidence of the financial benefits of implementing these technologies.

5 Conclusion

The use of data analytics and machine learning techniques to identify at-risk students and predict academic success in higher education was examined in this systematic review. Traditional techniques like logistic regression and support vector machines as well as more sophisticated techniques like ensemble methods and neural networks were among the many different approaches that were found in the review. These methods were used on a variety of data sources, including survey data, administrative data, and data from learning management systems, to produce insightful findings about the performance and behaviour of students. The report highlights how data analytics and machine learning have the power to completely transform higher education. Timely interventions that improve student retention and achievement rates are made possible by early identification of at-risk pupils. By accurately predicting academic success, educational institutions may optimise resource allocation by customising support services to meet the needs of each individual student. Better results can be achieved by addressing the various elements influencing academic performance with a holistic approach to student support that is guided by data analytics. Additionally, data analytics models can be tailored to meet the unique requirements of various organisations, guaranteeing their efficacy in a range of educational settings. To improve teaching and learning methods, maximise resource allocation, and increase student performance, schools can use these tools to make data-driven decisions.

6 Limitations and Suggestions for Further Study

This review prioritised the quality of evidence by focusing on peer-reviewed studies. While this approach ensured robustness, it inherently limited the scope of the review. Potentially valuable unpublished research (grey literature) on predicting academic success and identifying at-risk students in higher education might have been excluded. However, the review mitigated this limitation to some extent by employing a comprehensive search strategy across multiple electronic databases to maximise the capture of relevant peer-reviewed studies. Furthermore, the review methods explicitly justified the rationale behind the inclusion/exclusion criteria, particularly the decision to exclude grey literature. The review was inherently limited by the specific inclusion/exclusion criteria used for selecting studies. This may have restricted the generalisability of findings to all education settings. However, by outlining the inclusion/exclusion criteria, the review allowed readers to assess the applicability of the findings to their specific context.

A strong basis for comprehending the use of data analytics and machine learning in higher education is provided by this comprehensive review. Future studies should address ethical issues like data privacy and algorithmic bias, investigate cutting-edge techniques like deep learning and natural language processing, and promote interdisciplinary cooperation between computer scientists, educators, and social scientists to further develop this field. Effective data governance frameworks can guarantee data confidentiality, integrity, and quality, and longitudinal studies can evaluate the long-term effects of data-driven interventions. Future research can maximise the use of data analytics to improve teaching and learning, increase student achievement, and advance inclusive and equitable higher education by giving priority to these areas.

Acknowledgements The authors acknowledge funding from the National Foundation of South Africa (Grant No. SRUG2205025721), the postdoctoral fellow who assisted with the coding and the anonymous reviewers.

References

[1] Abdul-Jabbar, S.S.; Farhan, A.K. (2022). Data analytics and techniques: A review. *ARO-The Scientific Journal of Koya University*, 10(2). https://doi.org/10.14500/aro.10975.

[2] Abdulwahed, M. (2017). Technology innovation and engineering' education and entrepreneurship (TIEE) in engineering schools: A novel model for elevating national knowledge-based economy and socio-economic sustainable development. *Sustainability, 9*(2), 171. https://www.mdpi.com/2071-1050/9/2/171.

[3] Ahmad Tarmizi, S. S., Mutalib, S., Abdul Hamid, N. H., Abdul-Rahman, S., & Md Ab Malik, A. (2019). A case study on student attrition prediction in higher education using data mining techniques. In M. Berry, B. Yap, A. Mohamed, & M. Köppen (Eds.), *Soft computing in data science* (pp. 181-192). Springer, Singapore. https://link.springer.com/chapter/10.1007/978-981-15-0399-3_15#citeas.

[4] Aina, C., Baici, E., Casalone, G., & Pastore, F. (2019). *Delayed graduation and university dropout: A review of theoretical approaches (GLO Discussion Paper No. 399).* Global Labor Organization. https://www.econstor.eu/bitstream/10419/203475/1/GLO-DP-0399.pdf.

[5] Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop an early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education, 16*(1), 40. https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-019-0172-z#citeas.

[6] Akçapınar, G., Hasnine, M. N., Majumdar, R., Flanagan, B., & Ogata, H. (2019). Developing an early-warning system for spotting at-risk students by using eBook interaction logs. *Smart Learning Environments, 6*(1), 4. https://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/242865/1/s40561-019-0083-4.pdf.

[7] Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-thinking data strategy and integration for artificial intelligence: Concepts, Opportunities, and challenges. *Applied Sciences, 13*(12), 7082. https://www.mdpi.com/2076-3417/13/12/7082.

[8] Alturki, S., Cohausz, L., & Stuckenschmidt, H. (2022). Predicting master's students' academic performance: An empirical study in Germany. *Smart Learning Environments*, 9, 38. https://slejournal.springeropen.com/articles/10.1186/s40561-022-00220-y.

[9] Anagnostopoulos, T., Kytagias, C., Xanthopoulos, T., Georgakopoulos, I., Salmon, I., & Psaromiligkos, Y. (2020). Intelligent predictive analytics for identifying students at risk of failure in Moodle courses. In *Intelligent Tutoring Systems: 16th International Conference, ITS 2020*, Athens, Greece, June 8-12, 2020, Proceedings (pp. 152-162). Springer-Verlag. https://dl.acm.org/doi/10.1007/978-3-030-49663-0_19.

[10] ASReview. (2024). *Join the movement towards fast, open, and transparent systematic reviews*. https://asreview.nl/.

[11] Asrowi, Hadaya, A., & Hanif, M. (2019). The impact of using the interactive e-book on students' learning outcomes. *International Journal of Instruction, 12*, 709–722. https://www.e-iji.net/dosyalar/iji_2019_2_45.pdf.

[12] Azizah, Z., Ohyama, T., Zhao, X., Ohkawa, Y., & Mitsuishi, T. (2024). Predicting at-risk students in the early stage of a blended learning course via machine learning using limited data. *Computers and Education: Artificial Intelligence, 7,* 100261. https://doi.org/10.1016/j.caeai.2024.100261.

[13] Baashar, Y., Hamed, Y., Alkawsi, G., Capretz, L. F., Alhussian, H., Alwadain, A., & Al-amri, R. (2022). Evaluation of postgraduate academic performance using artificial intelligence models. *Alexandria Engineering Journal, 61*(12), 9867–9878. https://doi.org/10.1016/j.aej.2022.03.021.

[14] Bainbridge, J., Melitski, J., Zahradnik, A., Lauría, E. J. M., Jayaprakash, S., & Baron, J. (2015). Using learning analytics to predict at-risk students in online graduate public affairs and administration education. *Journal of Public Affairs Education, 21*(2), 247–262. https://www.tandfonline.com/doi/abs/10.1080/15236803.2015.12001831.

[15] Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A. E., & Karadeniz, A. (2020). An early warning system to detect at-risk students in online higher education. *Applied Sciences, 10*(13), 4427. https://doi.org/10.3390/app10134427.

[16] Bhumireddy, G., & Anala, V. A. S. M. (2022.). *Comparison of machine learning algorithms on detecting the confusion of students while watching MOOCs*. Master's thesis, Blekinge Institute of Technology, Karlskrona, Sweden.

[17] Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, 3, 100068. https://doi.org/10.1016/j.caeai.2022.100068.

[18] Bozkurt, A., & Sharma, R. C. (2022). Exploring the learning analytics equation: What about the "Carpe Diem" of teaching and learning? *Asian Journal of Distance Education, 17*(2), i–xiv. https://files.eric.ed.gov/fulltext/EJ1373860.pdf.

[19] Bozkurt, A. & Sharma, R. C. (2022). Exploring the learning analytics equation: What about the "Carpe Diem" of teaching and learning? *Asian Journal of Distance Education*, 17(2), i-xiv. https://files.eric.ed.gov/fulltext/EJ1373860.pdf.

[20] Brown, D., Abbasi, A., & Lau, R. (2015). Predictive analytics introduction. *IEEE Intelligent Systems, 30*(2), 6–8. https://ahmedabbasi.com/wp-content/uploads/J/Brown_PredAnal_IEEEIntSys_2015.pdf.

[21] CADIMA. (2024). *Evidence synthesis tool and database*. https://www.cadima.info/.

[22] Caspari-Sadeghi, S. (2023). Learning assessment in the age of big data: Learning analytics in higher education. *Cogent Education*, 10(1). https://doi.org/10.1080/2331186X.2022.2162697.

[23] Chili, M., & Madzimure, J. (2022). Using surveys of student engagement to understand and support first-time entering students at a university of technology. *ScienceRise: Pedagogical Education, 6*(51), 4–12. https://doi.org/10.15587/2519-4984.2022.267206.

[24] Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2018). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior, 107*, 105584. https://www.mdpi.com/2076-3417/13/6/3869.

[25] Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral science*. Routledge. https://asu.elsevierpure.com/en/publications/applied-multiple-regressioncorrelation-analysis-for-the-behaviora.

[26] Critical Appraisal Skills Programme (CASP). (2024). *CASP checklists*. https://casp-uk.net/casp-tools-checklists/.

[27] Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences, 120*(3/4), 208–227. https://doi.org/10.1108/ILS-10-2018-0104.

[28] Diamant, A. (2024). Introducing prescriptive and predictive analytics to MBA students with Microsoft Excel. *INFORMS Transactions on Education, 24*(2), 152–174. https://doi.org/10.1287/ited.2023.0286.

[29] Duin, A. H., & Tham, J. (2020). The current state of analytics: Implications for learning management system (LMS) use in writing pedagogy. *Computers and Composition, 55,* 102544. https://doi.org/10.1016/j.compcom.2020.102544.

[30] Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). ROC curve estimation: An overview. *REVSTAT – Statistical Journal, 12*(1), 1–20. https://www.ine.pt/revstat/pdf/rs140101.pdf.

[31] Guan, C., Mou, J., & Jiang, Z. (2020). Artificial intelligence innovation in education: A twenty-year data-driven historical analysis. *International Journal of Innovation Studies, 4*(4), 134–147. https://www.sciencedirect.com/science/article/pii/S2096248720300369.

[32] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer. https://link.springer.com/book/10.1007/978-0-387-84858-7.

[33] He, J., Bailey, J., Rubinstein, B., & Zhang, R. (2015). Identifying at-risk students in massive open online courses. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1). https://doi.org/10.1609/aaai.v29i1.9471.

[34] Herodotou, C., Naydenova, G., Boroowa, A., Gilmour, A., & Rienties, B. (2020). How can predictive learning analytics and motivational interventions increase student retention and enhance administrative support in distance education? *Journal of Learning Analytics, 7*(2), 72–83. https://files.eric.ed.gov/fulltext/EJ1273869.pdf.

[35] Hossain, B. M. (2016). Dropout at tertiary education in Bangladesh: Configurations and determinants. *Feni University Journal, 1*(1). https://www.researchgate.net/publication/354157150_Dropout_at_Tertiary_Education_in_Bangladesh_Configurations_and_Determinants.

[36] Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*. http://dx.doi.org/10.1016/j.chb.2014.04.002.

[37] Hung, J. L., Wang, M., Wang, S., Abdelrasoul, M., Li, Y. & He, W. (2015). Identifying at-risk students for early interventions? A time-series clustering approach. *IEEE Transactions on Emerging Topics in Computing*, *5*(1), 1-1. https://ieeexplore.ieee.org/document/7339455.

[38] Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilizing learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development, 68*(4), 1961–1990. https://link.springer.com/article/10.1007/s11423-020-09788-z#citeas.

[39] Jagwani, A. (2019). Review of machine learning in education. *Journal of Emerging Technologies and Innovative Research, 6*(5), 384–386. https://www.scribd.com/document/775098724/A-REVIEW-OF-MACHINE-LEARNING-IN-EDUCATION.

[40] Järvinen, P., Siltanen, P., Kirschenbaum, A. (2021). Data analytics and machine learning. In: Södergård, C., Mildorf, T., Habyarimana, E., Berre, A.J., Fernandes, J.A., Zinke-Wehlmann, C. (eds) *Big data in bioeconomy*. Springer, Cham. https://doi.org/10.1007/978-3-030-71069-9_10.

[41] Jia, P., & Maloney, T. (2015). Using predictive modelling to identify students at risk of poor university outcomes. *Higher Education, 70*(1), 127–149. https://link.springer.com/article/10.1007/s10734-014-9829-7#citeas.

[42] Johnson E. A., Inyangetoh J. A., Rahmon H. A., Jimoh T. G., Dan E. E. & Esang M. O. (2024) An intelligent analytic framework for predicting students academic performance using multiple linear regression and random forest. *European Journal of Computer Science and Information Technology, 12* (3),56-70. https://tudr.org/id/eprint/3111/1/An%20Intelligent%20Analytic%20Framework.pdf.

[43] Joshi, A., Desai, P., & Tewari, P. (2020). Learning analytics framework for measuring students' performance and teachers' involvement through problem-based learning in engineering education. *Procedia Computer Science*, 172, 954–959. https://www.sciencedirect.com/science/article/pii/S1877050920314678.

[44] Kandiko-Howson, C., & Matos, F. (2021). Student surveys: Measuring the relationship between satisfaction and engagement. *Education Sciences, 11*(6), 297. https://doi.org/10.3390/educsci11060297.

[45] Koehrsen, W. (2024). *Precision and recall in machine learning*. https://builtin.com/data-science/precision-and-recall.

[46] Kok, C. L., Ho, C. K., Chen, L., Koh, Y. Y., & Tian, B. (2024). A novel predictive modeling for student attrition utilizing machine learning and sustainable big data analytics. *Applied Sciences, 14*(21), 9633. https://www.meta.ai/c/cd5d2a15-1dc1-4d2a-ba5f-ba257753ade1.

[47] Koutina, M., & Kermanidis, K. L. (2011). Predicting postgraduate students' performance using machine learning techniques. In L. Iliadis, I. Maglogiannis, & H. Papadopoulos (Eds.), *Artificial intelligence applications and innovations* (pp. 143-150). Springer. https://link.springer.com/chapter/10.1007/978-3-642-23960-1_20#citeas.

[48] Kovač, V. B., Nome, D. Ø., Jensen, A. R., & Skreland, L. Lj. (2023). The why, what and how of deep learning: critical analysis and additional concerns. *Education Inquiry*, 1–17. https://doi.org/10.1080/20004508.2023.2194502.

[49] Kroese, D. P., Botev, Z. I., Taimre, T., & Vaisman, R. (2024). *Data science and machine learning: mathematical and statistical methods*. https://people.smp.uq.edu.au/DirkKroese/DSML/DSML.pdf.

[50] Kuhn, M., & Johnson, K. (2013). *Applied predictive modelling*. Springer. https://link.springer.com/book/10.1007/978-1-4614-6849-3.

[51] Kumar, V., & Garg, L. M. (2018). Predictive analytics: A review of trends and techniques. *International Journal of Computer Applications, 182*(1), 31–37. https://doi.org/10.5120/ijca2018917434.

[52] Kumari, P., Jain, P. K., & Pamula, R. (2018). An efficient use of ensemble methods to predict students' academic performance. In *Proceedings of the 4th International Conference on Recent Advances in Information Technology (RAIT)* (pp. 1-6). https://ieeexplore.ieee.org/document/8389056.

[53] Kurniawan, Y., & Halim, E. (2013). Use data warehouse and data mining to predict student academic performance in schools: A case study (perspective application and benefits). In *Proceedings of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering* (TALE 2013) (pp. 98-103). *IEEE.* https://ieeexplore.ieee.org/document/6654408.

[54] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)* (pp. 1909-1918). https://dl.acm.org/doi/10.1145/2783258.2788620.

[55] Latif, E., & Miles, S. (2020). The impact of assignments and quizzes on exam grades: A difference-in-difference approach. *Journal of Statistics Education, 28*(3), 289–294. https://doi.org/10.1080/10691898.2020.1807429.

[56] Latif, G., Alghazo, R., Pilotti, M. A. E., & Brahim, G. B. (2022). Identifying "at-risk" students: An AI-based prediction approach. *International Journal of Computing and Digital Systems, 11*(1). https://dx.doi.org/10.12785/ijcds/110184.

[57] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature,* 521, 436–444. https://www.nature.com/articles/nature14539#citeas.

[58] Liaison. (2024). Using predictive analytics for student success and retention at community colleges. https://www.liaisonedu.com/using-predictive-analytics-for-student-success-and-retention-at-community-colleges/

[59] Lourens, A., & D Bleazard. (2016). Applying predictive analytics in identifying students at risk: A case study. *South African Journal of Higher Education 30* (2), 129-42. https://www.journals.ac.za/sajhe/article/view/583.

[60] Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91, 216-231. https://www.sciencedirect.com/science/article/pii/S0031320319300950.

[61] Maher, M., & Macallister, H. (2013). Retention and attrition of students in higher education: Challenges in modern times to what works. *Higher Education Studies, 3*(2), 62. https://www.ccsenet.org/journal/index.php/hes/article/view/25268.

[62] Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. Scientific *Reports, 13*(1), 5705. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10082180/.

[63] Messaoud, S., Bradai, A., Bukhari, S. H. R., Quang, P. T. A., Ahmed, O. B., & Atri, M. (2020). A survey on machine learning in internet of things: Algorithms, strategies, and applications. *Internet of Things, 12*, 100314. https://www.sciencedirect.com/science/article/abs/pii/S2542660520301451.

[64] Methley, A., O'Malley, K. & Barber, J. (2014). Applying PICO: Using the PICO framework to build clinical questions. *BMJ, 348,* g1898. https://bestpractice.bmj.com/info/us/toolkit/learn-ebm/how-to-clarify-a-clinical-question/.

[65] Na, K. S., & Tasir, Z. (2017). Identifying at-risk students in online learning by analysing learning behaviour: A systematic review. In *2017 IEEE Conference on Big Data and Analytics (ICBDA)* (pp. 118-123). IEEE. https://ieeexplore.ieee.org/document/8284117.

[66] Naseer, F., Khan, M. N., Tahir, M., Addas, A., & Aejaz, S. M. H. (2024). Integrating deep learning techniques for personalized learning pathways in higher education. *Heliyon, 10*(11), e32628. https://www.sciencedirect.com/science/article/pii/S2405844024086596.

[67] Ncube, M. M. & Ngulube, P. A. (2024). Systematic review of postgraduate programmes concerning ethical imperatives of data privacy in sustainable educational data analytics. *Sustainability, 16*, 6377. https://doi.org/10.3390/su16156377.

[68] Nguyen, A., Gardner, L., & Sheridan, D. (2020). Data analytics in higher education: An integrated view. *Journal of Information Systems Education, 31*(1), 61–71. http://jise.org/Volume31/n1/JISEv31n1p61.html.

[69] Nguyen, A., Gardner, L. & Sheridan, D. (2020). Data analytics in higher education: An integrated view. *Journal of Information Systems Education*, 31(1), 61-71. http://jise.org/Volume31/n1/JISEv31n1p61.html.

[70] Nimy, E., Mosia, M., & Chibaya, C. (2023). Identifying at-risk students for early intervention—A probabilistic machine learning approach. *Applied Sciences, 13*, 3869. https://www.mdpi.com/2076-3417/13/6/3869.

[71] Ogata, H., Majumdar, R., Yang, S. J. H., & Warriem, J. M. (2022). Learning and Evidence Analytics Framework (LEAF): Research and practice in international collaboration. *Information and Technology in Education and Learning, 2*(1), Inv-p001. https://www.jstage.jst.go.jp/article/itel/2/1/2_2.1.Inv.p001/_article.

[72] Osborne, J. B., & Lang, A. S. (2023). Predictive identification of at-risk students: Using learning management system data. *Journal of Postsecondary Student Success, 2*(4), 108–126. https://journals.flvc.org/jpss/article/view/132082/138293.

[73] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E. et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *J. Clin. Epidemiol. 134*, 178-189. https://doi.org/10.1016/j.jclinepi.2021.03.001.

[74] Paterson, K., & Guerrero, A. (2019). Predictive analytics in education: Considerations in predicting versus explaining college student retention. *Research in Higher Education Journal,* 44, 1-12. https://files.eric.ed.gov/fulltext/EJ1401369.pdf.

[75] Pecuchova, J., & Drlik, M. (2023). Predicting students at risk of early dropping out from course using ensemble classification methods. *Procedia Computer Science*, 225, 3223–3232. https://doi.org/10.1016/j.procs.2023.10.316.

[76] Pek, R. Z., Özyer, S. T., Elhage, T., Özyer, T., & Alhajj, R. (2023). The role of machine learning in identifying students at-risk and minimizing failure. *IEEE Access*, 11, 1224–1243. https://ieeexplore.ieee.org/document/10002336.

[77] Pinto, A. S., Abreu, A., Costa, E., & Paiva, J. (2023). How machine learning (ML) is transforming higher education: A systematic literature review. *Journal of Information Systems Engineering and Management, 8*(2), 21168. https://doi.org/10.55267/iadt.07.13227.

[78] Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63. https://arxiv.org/pdf/2010.16061.

[79] Psathas, G., Chatzidaki, T. K., & Demetriadis, S. N. (2023). Predictive modeling of student dropout in MOOCs and self-regulated learning. *Computers, 12*(10), 194. https://www.mdpi.com/2073-431X/12/10/194.

[80] Queiroga, E. M., Lopes, J. L., Kappel, K., Aguiar, M., Araújo, R. M., Munoz, R., Villarroel, R., & Cechinel, C. (2020). A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course. *Applied Sciences, 10*(11), 3998. https://www.mdpi.com/2076-3417/10/11/3998.

[81] Qushem, U. B., Oyelere, S. S., Akçapınar, G., Kaliisa, R., & Laakso, M.-J. (2024). Unleashing the power of predictive analytics to identify at-risk students in computer science. Technology, *Knowledge and Learning*, 29, 1385–1400. https://link.springer.com/article/10.1007/s10758-023-09674-6#citeas.

[82] Rahiman, H. U., & Kodikal, R. (2023). Revolutionizing education: Artificial intelligence empowered learning in higher education. *Cogent Education, 11*(1). https://doi.org/10.1080/2331186X.2023.2293431.

[83] Ramaswami, G., Susn jak, T., & Mathrani, A. (2023). Effectiveness of a learning analytics dashboard for increasing student engagement levels. *Journal of Learning Analytics, 10*(3), 115–134. https://files.eric.ed.gov/fulltext/EJ1411453.pdf.

[84] Rane, N. L., Paramesha, M., Choudhary, S. P., & Rane, J. (2024). Machine learning and deep learning for big data analytics: A review of methods and applications. *Partners Universal International Innovation Journal (PUIIJ), 2*(3). https://doi.org/10.5281/zenodo.12271005.

[85] Rayyan. (2024) *Faster systematic reviews*. https://www.rayyan.ai/.

[86] Reinhold, F., Strohmaier, A., Hoch, S., Reiss, K., Böheim, R., & Seidel, T. (2020). Process data from electronic textbooks indicate students' classroom engagement. *Learning and Individual Differences*, 83-84, 101934. https://doi.org/10.1016/j.lindif.2020.101934.

[87] Ressa, T., & Andrews, A. (2022). High school dropout dilemma in America and the importance of reformation of education systems to empower all students. *International Journal of Modern Education Studies, 6*(2), 423–447. https://doi.org/10.51383/ijonmes.2022.23.

[88] Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2024). The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns, 5*(6), 100994. https://www.sciencedirect.com/science/article/pii/S2666389924001090.

[89] Russell, J.-E., Smith, A., & Larsen, R. (2020). Elements of success: Supporting at-risk student resilience through learning analytics. *Computers & Education*, 152, 103890. https://doi.org/10.1016/j.compedu.2020.103890.

[90] Seo, E. Y., Yang, J., Lee, J. E., & So, G. (2024). Predictive modelling of student dropout risk: Practical insights from a South Korean distance university. *Heliyon, 10*(11), e30960. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11145199/.

[91] Sharma, D. K., Chatterjee, M., Kaur, G., & Vavilala, S. (2022). Deep learning applications for disease diagnosis. In D. Gupta, U. Kose, A. Khanna, & V. E. Balas (Eds.), *Deep learning for medical applications with unique data* (pp. 31-51). Academic Press. https://www.sciencedirect.com/science/article/abs/pii/B9780128241455000058.

[92] Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting student performance in online learning: A multidimensional time-series data analysis approach. *Applied Sciences, 14*(6), 2522. https://doi.org/10.3390/app14062522.

[93] Sivarajah, U., Kamal, M. M., Irani, Z. & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70(2), 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001.

[94] Smirani, L. K., Yamani, H. A., Menzli, L. J., & Boulahia, J. A. (2022). Using ensemble learning algorithms to predict student failure and enabling customized educational paths. *Scientific Programming, Article ID* 3805235. https://onlinelibrary.wiley.com/doi/10.1155/2022/3805235.

[95] Smithers, L. (2023). Predictive analytics and the creation of the permanent present. *Learning, Media and Technology, 48*(1), 109–121. https://doi.org/10.1080/17439884.2022.2036757.

[96] Stojanov, A. & Daniel, B. K. (2024). A decade of research into the application of big data and analytics in higher education: A systematic review of the literature. *Educ Inf Technol*, 29, 5807–5831. https://doi.org/10.1007/s10639-023-12033-8.

[97] Syed Mustapha, S. M. F. D. (2023). Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods. *Applied System Innovation, 6*(5), 86. https://doi.org/10.3390/asi6050086.

[98] Tamrat, W. (2022). The indelible challenges of student attrition in Ethiopian higher education: Imperatives for a closer scrutiny. *Widening Participation and Lifelong Learning, 24*(1), 86-113. https://www.ingentaconnect.com/content/openu/jwpll/2022/00000024/00000001/art00005;jsessionid=1wm771ldnm6ta.x-ic-live-01.

[99] TechTarget. *Data analytics definition*. https://www.techtarget.com/searchbusinessanalytics/resources/Data-science-and-analytics.

[100] Tete, M. F., Sousa, M. M., Santana, T. S., & Fellipe, S. (2022). Predictive models for higher education dropout: A systematic literature review. *Education Policy Analysis Archives, 30*(149). https://doi.org/10.14507/epaa.30.6845.

[101] United Nations Educational, Scientific and Cultural Organization. (2020). *Global Education Monitoring Report 2020*. https://gem-report-2020.unesco.org/.

[102] Varade, R. V., & Gupta, S. (2024). A data-driven model for predicting the academic performance of students employing ANN-PSO hybrid approach. *Journal of Electrical Systems, 20*(10s). https://journal.esrgroups.org/jes/article/view/6393/4477.

[103] Vărzaru, A. A., Bocean, C. G., Rotea, C. C., & Budică-Iacob, A.-F. (2021). Assessing antecedents of behavioral intention to use mobile technologies in e-commerce. *Electronics, 10*(18), 2231. https://www.mdpi.com/2079-9292/10/18/2231.

[104] Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 104, 106189. https://doi.org/10.1016/j.chb.2019.106189.

[105] Wan Yaacob, W. F., Mohd Sobri, N., Md Nasir, S. A., Norshahidi, N. D., & Wan Husin, W. Z. (2020). Predicting student drop-out in higher institution using data mining techniques. *Journal of Physics: Conference Series*, 1496, 012005. https://iopscience.iop.org/article/10.1088/1742-6596/1496/1/012005/pdf.

[106] Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*, 11. https://doi.org/10.1186/s40561-022-00192-z.

[107] Yıldız, M. B., & Börekçi, C. (2020). Predicting academic achievement with machine learning algorithms. *Journal of Educational Technology & Online Learning, 3*(3), 372–392. https://dergipark.org.tr/en/download/article-file/1214052.

[108] Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting dropout students: An application of data mining methods in an online education program. *European Journal of Open, Distance and E-Learning, 17*(1), 118–133. https://files.eric.ed.gov/fulltext/EJ1017900.pdf.

[109] Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in Psychology, 12*, 698490. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.698490/full.