Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 18 May 2026
Revised 20 June 2026
Accepted 25 July 2026

Natural Resources for Human Health 2026, Vol. 6 (8s): 221–238

KEYWORDS:
Student performance prediction, Higher education, Machine learning, Academic achievement, Learning analytics, Bibliometric analysis, Systematic review.

A Bibliometric and Systematic Review of Machine Learning Applications for Predicting Student Performance in Higher Education: Insights via the TCCM Framework

Pawan Painuli[1], Dr. Amar Jeet Rawat[2]

[1]PhD Scholar, Uttaranchal School of Computing Sciences, Uttaranchal University, Dehradun
Email: pawanpainuli@gmail.com
[2] Associate Professor, Uttaranchal School of Computing Sciences, Uttaranchal University, Dehradun
Email: aj.amar.rawat@gmail.com

ABSTRACT:

Background: Higher education institutions collect large amounts of student data through learning management systems, student information systems, online assessments and academic records. These data sources have increased the use of machine learning for predicting student performance. Although this area has grown in recent years, the literature remains scattered across education, computer science and learning analytics.

Purpose: This study presents a bibliometric and systematic review of machine learning applications for predicting student performance in higher education. The review also uses the TCCM framework to examine the literature through four dimensions: Theory, Context, Characteristics and Methodology.

Methods: The study followed the PRISMA framework. Data were collected from the Scopus database. The search covered peer-reviewed English-language publications from 2017 to 2025. The search produced 1,193 records. After duplicate removal and screening, 1,156 records were included in the final review. Bibliometric analysis was conducted using R software, Bibliometrix and Biblioshiny.

Results: The results of this study have shown an upward trend in machine learning based on the number of articles related to predicting students' performance. The 47.69% Annual Growth Rate (AGR) for machine learning shows there has been an increased interest in using Machine Learning methods to predict how well a student will do in school. Most of these studies are found in specific areas around the world; and at the institution level. Student data that can be used as predictors includes: attendance, assignments completed by student; log files from Learning Management Systems (LMS); test scores, quizzes etc.; prior academic history/achievement. Many researchers are now utilizing classification algorithms and ensemble models. A large number of studies rely upon data collected from one single institution. Limited studies utilized external validations, theoretical bases to select variables, explanations for their model's predictions; and actual use within the classroom to evaluate if their model could make valid predictions about students' potential success.

Conclusion: The study indicates that machine learning is a major methodology in using data from students to predict student success at the post-secondary level. While there are many studies in this area; most of these focus on technical aspects of model building rather than providing insight into the meaning and relevance of their findings. Therefore, future

studies will need to develop models that are based on well-established theories of education as well as clearly define constructs used to build those models. Additionally, researchers will need to validate the generalizability of models by testing them across different contexts and provide clear descriptions of how they validated their results.

1. INTRODUCTION

Higher education has been changed by the use of digital learning platforms. Universities have access to a vast amount of student data today using a variety of tools including; student information systems, attendance records, academic databases, online assessments, and learning management systems. These types of data are being used to inform the use of Machine Learning to predict student performance and to develop plans for providing academic support to students [1]. Student performance is predicted based upon the use of machine learning techniques related to the ability of students to complete courses successfully, to remain engaged throughout the semester, to achieve specific levels of academic achievement, and to drop out of school. When the use of predictive modelling is applied, institutions will be able to identify those students that could benefit from early intervention. In addition, it will enable educators to develop interventions tailored specifically to each student's needs to increase the chances that each student will remain in school [2].

There is growing interest in research relating to this topic; however, most research studies related to machine learning in education are located within three separate disciplines (education, computer science and data science) and as a result, much of the existing body of knowledge is fragmented. Furthermore, many studies primarily report on the accuracy of the models used, rather than reporting on the educational significance of the results. Finally, many studies rely solely on data collected at one institution. Therefore, the application of such findings in other educational contexts would be limited [5].

A second primary concern relates to the connection between educational theory and machine learning models. Much of the work completed in this field relies on readily accessible data points such as attendance, grade level scores, assignments submitted, and Learning Management System (LMS) activities. However, these data points do not necessarily reflect educational constructs [3]. Therefore, when conducting research in this field, terms such as student performance, academic success, or student participation are frequently defined differently across studies. Thus, comparing the results from different research papers will be problematic [4].

Reviews have been completed that examine the use of algorithms in machine learning in an educational context. Reviews have also been completed to assess the data sets used in machine learning in an educational context. Some reviews have documented broader patterns and trends regarding the utilization of educational technologies. However, there is no previous review that has comprehensively reviewed both theory/methodologies, along with contextual elements and features of the data sets. Thus, previous reviews have provided evidence not only of what type of models were employed, but whether or not those models informed decisions made by educators [6].

Therefore, the current study uses a systematic process called the TCCM framework to address this gap. TCCM includes four dimensions. The first dimension represents Theory – examining the theoretical foundations underlying prior studies utilizing machine learning for predicting student performance [8]. The second dimension represents Context – identifying countries, institutions, and educational environments represented in prior studies. The third dimension represents Characteristics – identifying variables and outcomes measured in prior studies utilizing machine learning for predicting student performance. The fourth dimension represents Methodology – reviewing methodologies employed by prior researchers including machine learning techniques, data sources utilized by researchers, validation procedures followed by researchers and evaluation criteria employed by researchers [11].

The purpose of this article is to provide a bibliometric and systematic review of machine learning applications for predicting student performance in higher education [7]. The review is based on Scopus-indexed studies published from 2017 to 2025. Bibliometric analysis is used to examine publication growth, major sources, authorship patterns, country contribution and keyword trends. The TCCM framework is used to assess the structure and gaps of the existing literature [10]

The objectives of this study are as follows:

1. To examine the growth, distribution and publication trends of research on machine learning-based student performance prediction in higher education from 2017 to 2025.

2. To analyse the literature through the TCCM framework with reference to theory, context, characteristics and methodology.
3. To identify limitations in current research related to theoretical basis, construct clarity, contextual coverage, model validation and educational use.
4. To suggest future research directions for developing reliable and interpretable predictive models for higher education.

2. METHODOLOGY

Study uses a systematic review and bibliometric analysis followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to ensure both the transparency and consistency of their study selection and reporting." The researchers' methodology was also designed for replication and "provided a clear account of how the literature was identified, screened and analysed."

Data Source

One single database (Scopus) was used as the sole "data source for this review." Scopus provides an "extensive coverage of peer reviewed journals and conference proceedings", and Scopus also "covers interdisciplinary fields including education, computer science and others." The use of one database allowed for "consistent indexing standards, citation data and metadata structure" which are important considerations when conducting bibliometric analysis [12].

Search Strategy

A structured search strategy was developed to find studies that utilize machine learning to predict student performance in higher education. The search string that was used in Scopus was ("student performance" OR "academic achievement" OR "learning outcome") AND "machine learning" AND ("higher education" OR "university"). The search results were limited to peer-reviewed publications published in English between 2017 and 2025. By limiting their search to peer-reviewed publications, the authors ensured that the studies selected represented current methodologies and practices in higher education [13].

Identification and Screening Process

In the initial search, 1,193 records were identified. All records found during the search were exported in CSV format and processed using R software. First, at the identification phase, duplicate records were removed, eliminating 7 duplicate records. Second, an automatic filtering tool was used to eliminate 20 records that failed to meet the minimum inclusion criteria. Third, an additional 10 records were eliminated based upon specific reasons, such as, but not limited to, missing author information, one page conference papers with insufficient content and studies that were not focused on predicting student performance in higher education. After the elimination of these records, 1,156 records remained available for screening purposes. Screening of the remaining records occurred by reviewing the title, abstract and keywords associated with each record to determine the extent to which each record fulfilled the purpose of the study.

Eligibility Assessment

All studies that utilized machine learning techniques to predict student performance or other related academic outcomes in higher education were considered eligible. Since all of the remaining screened records satisfied the eligibility criteria, each of those records were successfully retrieved for full assessment [15]. Therefore, no studies were eliminated during the eligibility assessment. Therefore, the authors determined that a total of 1,156 studies were selected for inclusion in the final review. A summary of the authors' process of identifying, screening and selecting studies is illustrated in the PRISMA flow diagram included in the study.

Data Analysis Tools

Bibliometric analysis was performed utilizing R software and the Bibliometrix package, along with Biblioshiny, a graphical user interface to Bibliometrix. The authors utilized Bibliometrix and Biblioshiny to analyse publication trends and frequency, the productivity of sources, authorship patterns, institutional affiliations, geographic locations contributing to research, citation counts and reference structures [17]. The authors utilized Biblioshiny to create visualizations of the data collected during the study, including but not limited to, publication trend graphs, keyword tree maps, topic evolution diagrams and three field plots illustrating the relationships between keywords, sources and geographic locations [16].

Analytical Framework

The authors utilized the TCCM (Theory, Context, Characteristics and Methodology) framework to organize and assess the systematic review [18]. The authors utilized the TCCM framework to categorize the literature into four categories: Theory - the theoretical bases of studies; Context - the educational and geographic contexts in which the research was conducted; Characteristics - the types of data collected and the variables measured; and Methodology - the machine learning techniques and evaluation approaches employed in studies [20].

Reliability and Transparency

The researchers recorded each step of the evaluation process in sufficient detail for reviewers to evaluate the credibility of the entire research process. This included how databases were selected; what was searched for during the search strategy; which studies met the inclusion criteria; and how the results of those studies were analysed. The use of PRISMA guidelines to guide the study, a specific search string that was used throughout the study, standard bibliometric tools, and an organized analytical structure enhanced the reproducibility of the study [17].

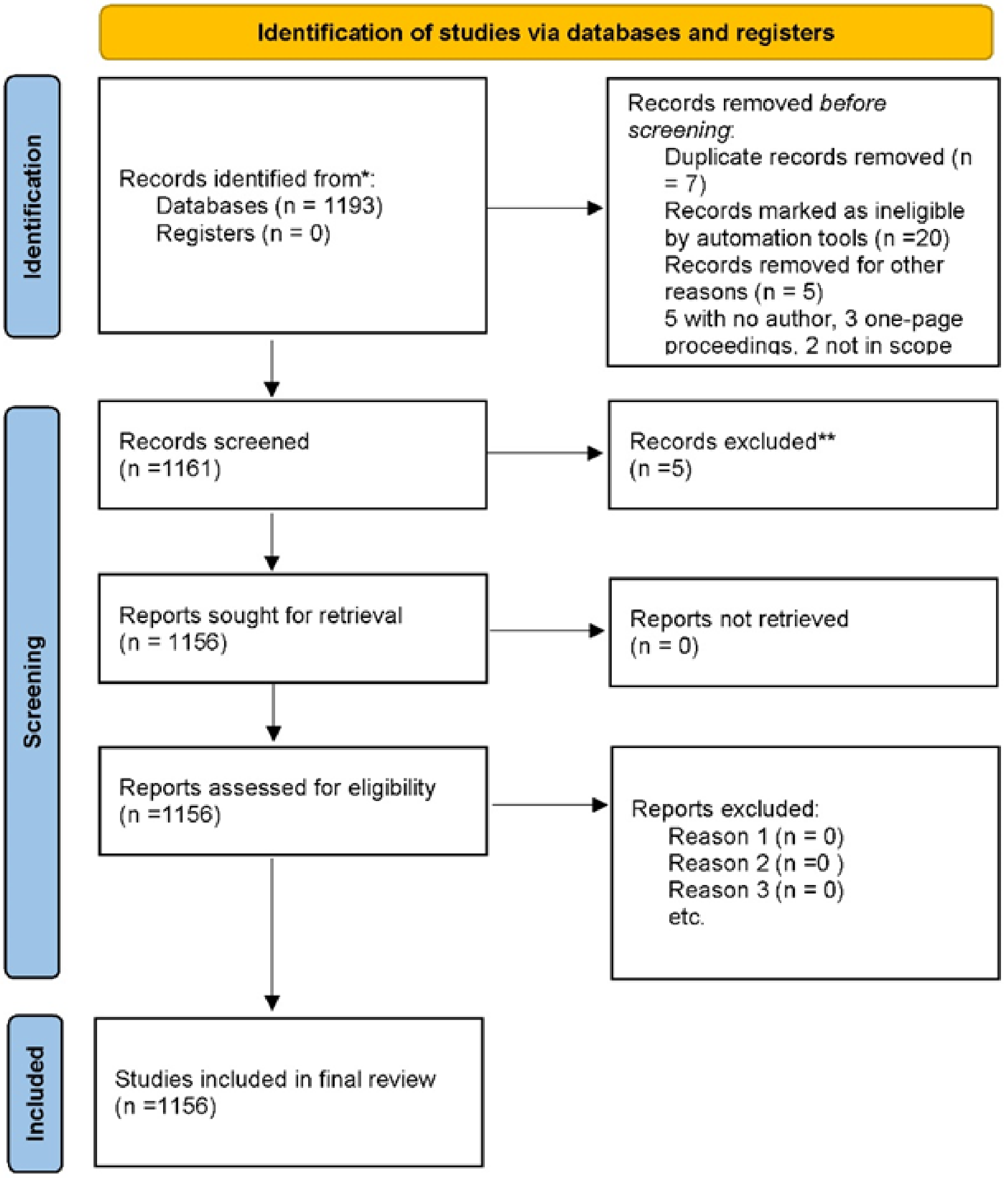

Figure 1. Prisma framework (Source: Author)

3. RESULTS AND DISCUSSION

Table 1. Basic information

Time period	**2017–2025**
Number of documents	1156
Number of sources	323
Annual growth rate	47.69%
Authors	3539
International co-authorship	19.03%
References	8635
Average citations per document	10.34

The dataset spans from 2017 to 2025 and consists of 1156 documents across 323 different information sources. Thus, it represents an increasing number of researchers interested in applying machine learning methods to forecast student performance in higher education. The annual rate of growth in this area of study was 47.69%. Hence, there has been a rapid expansion in the body of literature related to this topic. A fast-growing body of literature would indicate both the importance of the subject matter as well as the urgency felt by institutions to apply data-driven strategies to address educational problems. The scholarship represented by the research contributions involves 3539 authors; therefore, a very large number of scholars were involved in the scholarship on this topic. In addition, international co-authorship accounts for 19.03%, which represents a fairly substantial amount of global collaboration. The limited nature of this global collaboration could be a result of regional differences in prioritizing topics, data privacy concerns, or structural differences in educational systems across countries which inhibit further global collaboration on cross-boundary projects. There are 8635 references associated with the dataset, indicating that the referenced works were largely supported by previous studies. The average number of citations per document (10.34) suggests a moderate amount of scholarly influence. However, the average citations per document must be viewed cautiously because it is comprised of relatively recent documents that have not had sufficient time to obtain citations [19].

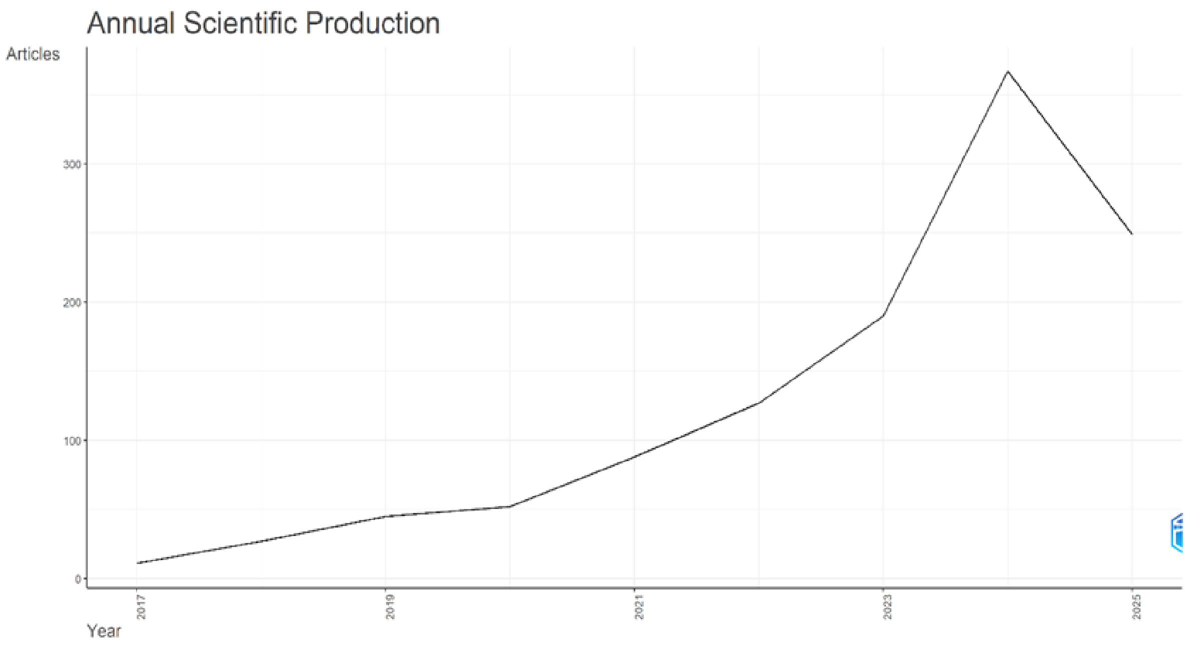

Figure 2. Annual scientific production (Source: Biblioshiny)

The graph indicates an increase in the amount of scientific output on the use of Machine Learning (ML) to predict student performance from 2017 through 2025. As shown in Figure 2, the output has grown over time, which demonstrates increasing interest in the application of ML in Higher Education by academics. Output levels grew gradually from 2017 to 2020; there was then a notable surge after 2020. This rapid increase in research output from 2021 through 2023 likely reflects an intense period of academic focus in this field. It is possible that the global shift to digital/remote learning during this period led to an increase in the adoption of predictive models in academia, thereby leading to an increase in research in this field. It seems that there was a decline in output in 2025; however, I believe this decline was primarily caused by an absence of complete data for the year, when analysed, as opposed to a decline in the interest to conduct research on predicting student performance [30]. The trend clearly continues to show increasing and expanding scholarly interest in using Machine Learning (ML) to gain insights into education, particularly as institutions continue to seek out systematic approaches to assist students with their academic success [28]

Table 2. Year-wise Publications in Selected Journals

Year	LECTURE NOTES IN NETWORKS AND SYSTEMS	COMMUNICATIONS IN COMPUTER AND INFORMATION SCIENCE	APPLIED SCIENCES (SWITZERLAND)	IEEE ACCESS	INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS	PROCEEDINGS - FRONTIERS IN EDUCATION CONFERENCE, FIE
2017	0	1	0	0	0	0
2018	1	1	0	1	1	0
2019	1	2	0	1	1	1
2020	1	3	3	4	4	1
2021	1	6	5	6	5	2
2022	4	9	11	7	8	4
2023	12	14	15	12	9	5
2024	27	26	17	17	15	5
2025	43	35	23	22	15	15

The data in Table 2 illustrate a rising trend of publication output across selected journals and conferences for machine learning and higher education from 2017 to 2025. Although many journals showed an increased level of publishing in machine learning and higher education after 2020, prior to 2020 the volume of publication was minimal. Between 2017-2019, very few articles appeared in all but one or two journals per year. However, between 2020 and 2025, there has been a notable increase in both the number of articles published as well as a greater diversity of source types. A notable example of this is the "Lecture Notes in Networks and Systems" and "Communications in Computer and Information Science" series. These series have shown dramatic increases; specifically, "Lecture Notes in Networks and Systems" reached 43 publications in 2025, while "Communications in Computer and Information Science" had 35 in the same year. This indicates a reliance on conference-based dissemination for this area of study [26], [29].

Natural Resources for
Human Health
eISSN: 2583-1194

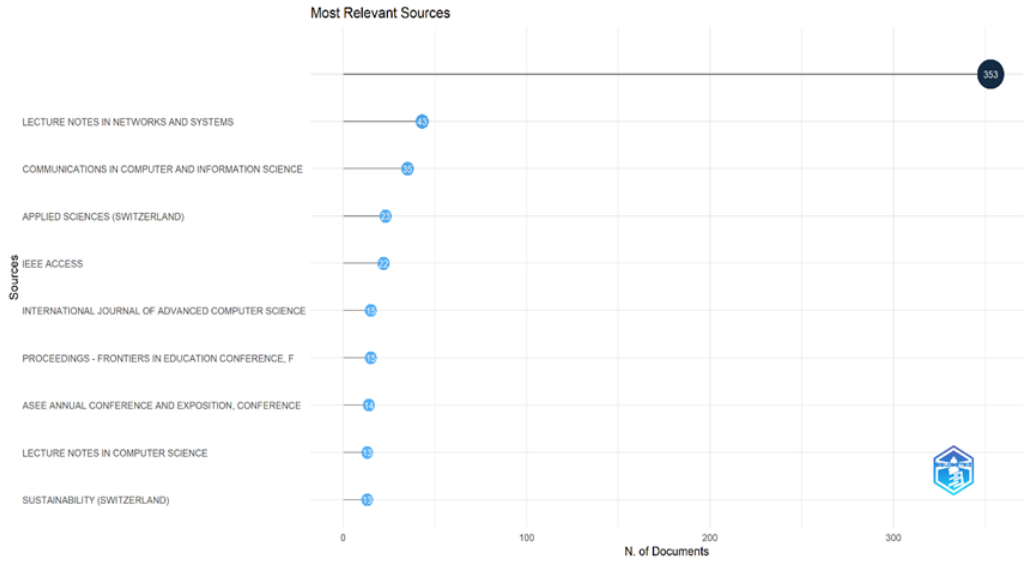

Figure 3. Highly Pertinent Sources (Source: Biblioshiny)

Figure 3 depicts the most important citation sources related to predicting student performance through machine learning. The cited sources are ordered based on the number of citations in the data set. Fig. 3 illustrates the diverse nature of the journals and conference proceedings providing input to the growth of this area of study, as well as the increasing number of inputs over the last few years in each of these areas. A substantial majority of the total number of cited works are represented by "Lecture Notes in Networks and Systems". As a result of the speed at which machine learning techniques develop, it is likely that many of the first presentations of new machine learning methodologies occur through conference-based publications (Bearman et al., 2022).

"CIS", second to "LNNS", reinforces the importance of conference series in this research area. "Applied Sciences (Switzerland)" and "IEEE Access" are also in the top five, illustrating an increase in awareness of this topic from researchers publishing in peer-review, open access journal settings.

Additional sources of note include "IJACSA" and "FIE", indicating that the work being done is in the realm of educational computing and interdisciplinary in nature. The presence of other sources such as "Sustainability (Switzerland)" may indicate that there is an emerging interest in assessing educational performance relative to larger institutional or policy contexts [25].

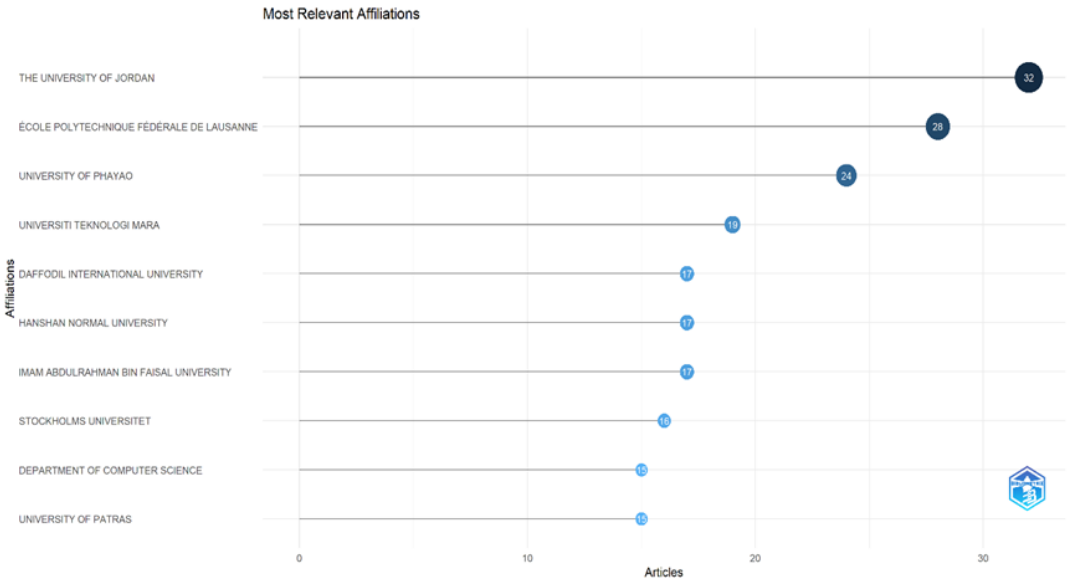

Figure 4. Most relevant Affiliations (Source: Biblioshiny)

As illustrated in Figure 4, the top contributors to the literature on the application of machine learning to predict the performance of students in Higher Education are represented by academic and research institutions that have demonstrated a continued commitment to the use of data-based techniques to enhance educational outcomes. Therefore, in order of the highest number of publications, the top contributors are:

1) University of Jordan.

2) École Polytechnique Fédérale de Lausanne.

3) University of Prizren.

These three institutions represent the largest contributor base in Europe, with an indication of established research groups or clusters focused on the application of machine learning for the purpose of enhancing the outcomes of students in Higher Education. Other contributors with notable amounts of publications include:

1) University Teknologi MARA (Malaysia).

2) Daffodil International University (Bangladesh).

3) Hunan Normal University (China), which suggest that there is significant research occurring within Asia.

In addition, these institutions appear to be involved in developing methodology and conducting applied research in utilizing machine learning in educational environments.

As such, the diversity of institutions that are located throughout the world, representing various countries and continents, illustrates the global nature of the topic; thus, predictive analytics in education is an international academic concern that has relevance to a variety of different educational systems and contexts [24].

While some institutions may focus primarily on developing technology-based applications, other institutions may be attempting to integrate educational theory with computational methodologies, as suggested by the interdisciplinary affiliations of the Department of Computer Science, as well as the broader academic profile of universities [23].

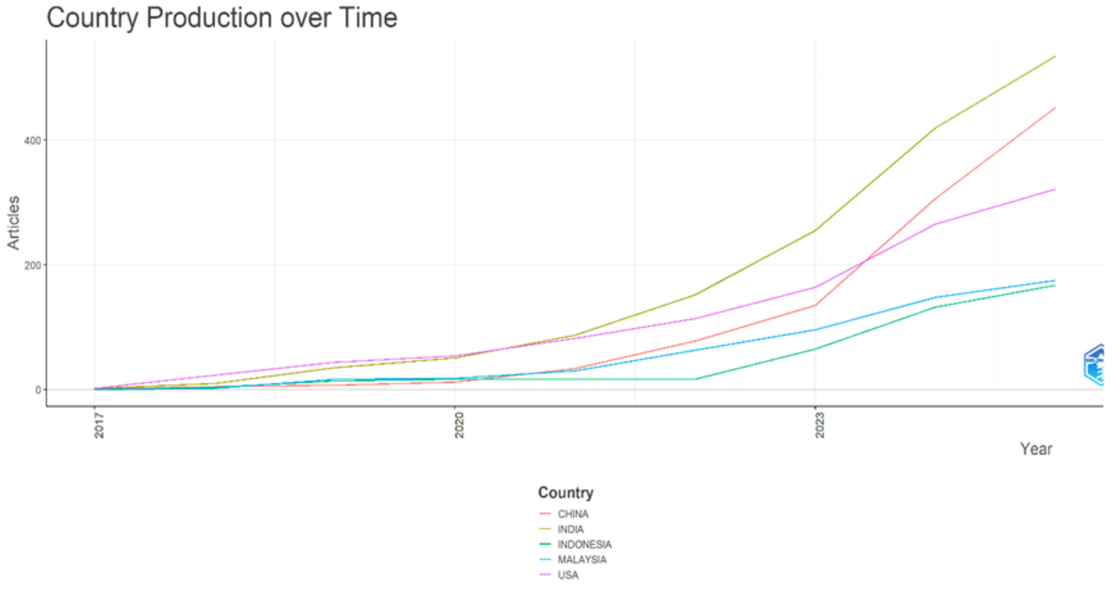

Figure 5. Country production over time (Source: Biblioshiny)

Figure 5 shows how the number of scientific outputs produced over time varies among countries involved in predictive student performance research via machine learning. The data shown represent the top five (China, India, Indonesia, Malaysia, and the U.S.) countries as identified through an analysis of the number of papers published each year. Figure 5 clearly indicates that there has been an increase in international participation in the study of student performance prediction via machine learning; and, further, that this has been most pronounced in the region of Asia. This increased international participation may indicate a world-wide move to make decisions about students using data, and be motivated locally and regionally by academic priorities.

Of the top five countries represented in Figure 5, China represents the largest cumulative volume of research outputs, and has had the greatest rate of increase since approximately 2020. These results suggest that there has been a large commitment to developing data science in education and other applications in the last few years in China. The steady and large increase in research outputs in China indicates continued institutional and governmental support for research in the area of data science in education [21].

India's growth rate was also considerable and, similar to China's, has grown substantially since 2020. However, unlike China, India's growth has been relatively steady and at a somewhat lower volume than China's. Therefore, it appears that both countries have invested heavily in artificial intelligence (AI) research and education technologies; however, China has a larger volume of outputs [22].

Both Indonesia and Malaysia have had moderate, yet stable increases in their research outputs. These countries' increasing outputs reflect the growing role of countries in Southeast Asia in researching machine learning for education purposes, and that universities within these countries are producing empirical studies that utilize data available within their own countries.

Finally, the United States has exhibited a growth trend in the number of research outputs that is more gradual than those of the three countries that are currently exhibiting rapid growth rates. While the overall number of outputs of the United States is significantly less than those of China and India, the relatively steady stream of research outputs over time indicates a stable level of academic interest in this area and the existence of established research infrastructure supporting such studies [28].

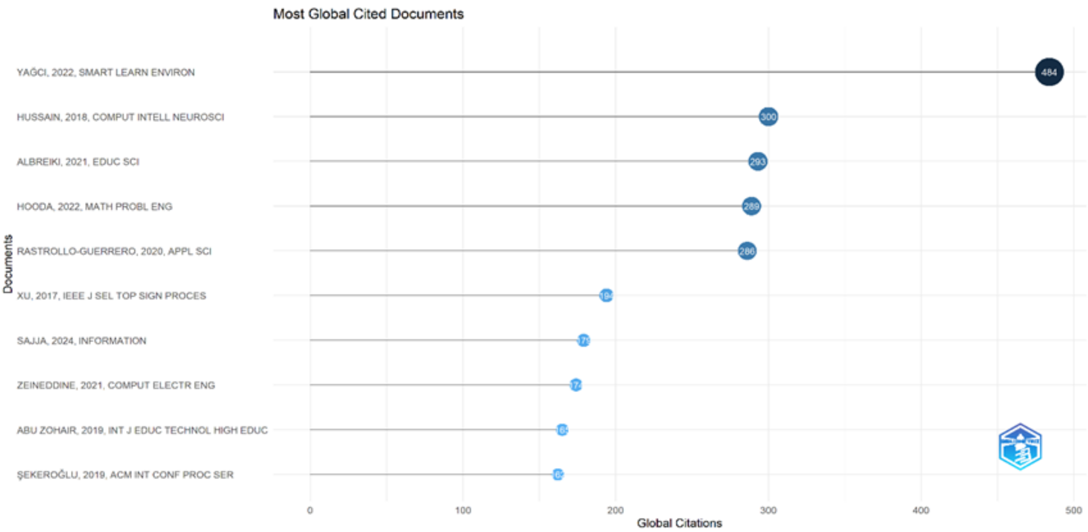

Figure 6. Top 10 global cited documents (Source: Biblioshiny)

Figure 6 presents the top 10 papers in terms of total global citations, which serve as a measure of their academic significance, pertaining to the application of machine learning models to predict performance of students in higher education [30].

The paper with the highest number of citations, thus indicating the greatest potential for impact, was Yago (2022) published in Smart Learn Environment. This suggests that the study has had an impact on either methodology or practice, since it was highly cited, and/or has been widely adopted in educational environments.

Highly cited works included those by Hussain (2018) and Aljerrah (2021) which were published in established journals, such as Comput Intell Neurosci and Educ Sci. The fact that the journals from which these papers originated are primarily in the fields of education and AI reflects the multidisciplinary nature of the subject matter of this literature [35].

A significant number of papers, including Hooda (2022) and Martínez-González (2021), have high numbers of citations; suggesting that recent studies have been referenced extensively, and may be due to an increase in interest in research post-2020 [34].

Collectively, this group of documents forms the foundation of the body of literature. The citation data also indicate that studies combining robust machine learning techniques with accessible educational applications are most likely to attract greater academic attention [36].

Natural Resources for Human Health
eISSN: 2583-1194

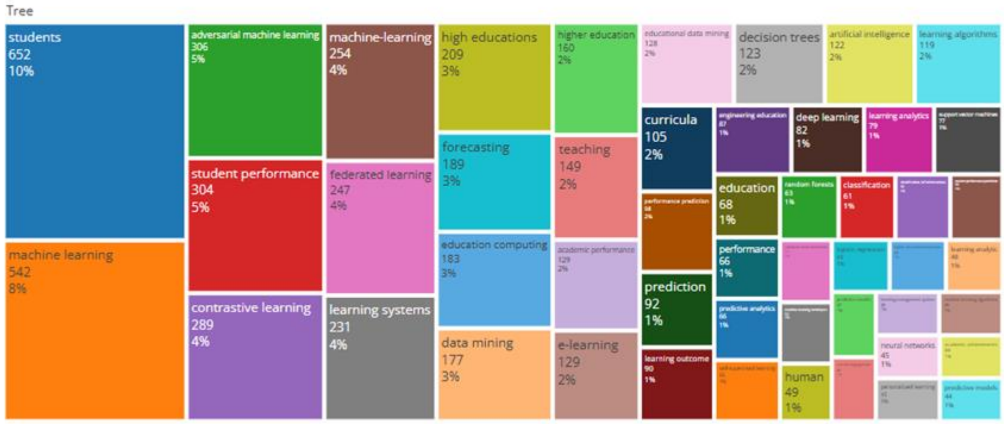

Figure 7. Tree map (keywords) (Source: Biblioshiny)

A tree map of the most common author keywords for the articles in Figure 7 about predicting student success in post-secondary education using machine learning can be seen in figure 7. Keywords are represented by rectangles where their sizes are representative of how many times each was used in the study [33].

The largest keyword is "students" (used 626 times) followed by "machine learning" (used 542 times) and then "student performance" (used 304 times) which illustrates the literature's concentration on predictive modelling to improve student success.

Some of the keywords related to machine learning techniques are also very noticeable including "deep learning", "federated learning", "contrastive learning" and "decision trees." Therefore, the research area is not just about generic models but also about developing specialized models for educational data environments [32].

Also, some educational context keywords are noticed including "higher education," "teaching," and "curricula" that illustrate the educational context of the research and also illustrate the interdisciplinary nature of the subject matter that includes computational methods and educational theory [30], [31].

Terms such as “data mining,” “learning algorithms” and “predictions” which are technical terms associated with the use of data-driven methodologies; and "personalized learning," "adaptive systems," and "academic performance" which indicate an applied focus on improving the learning experience and academic achievement using predictive techniques, reveal that this study will be using a data-driven approach to understand the research area of applying computer science to education [36].

Overall, the tree map illustrated a wide range of research topics that integrate computer science and education. It shows that even though the core research areas continue to focus on predictive modeling and performance, the research area continues to develop towards incorporating both advanced methodologies for developing algorithms and educational theories [37].

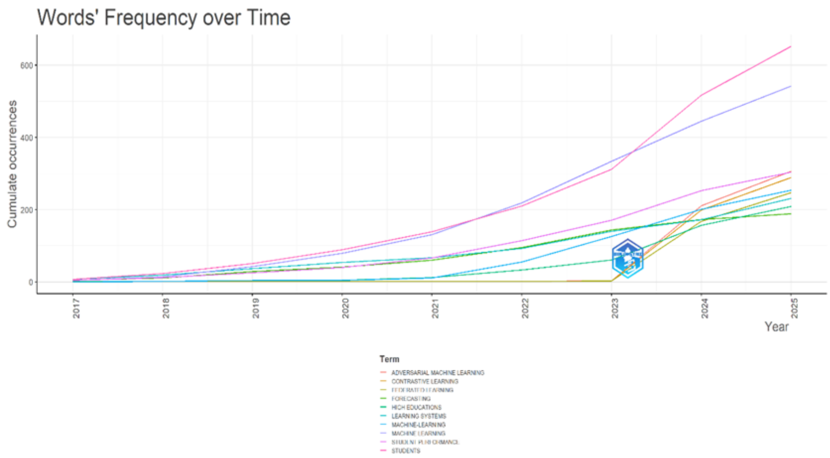

Fig 8. Words’ frequency over time

Natural Resources for
Human Health

eISSN: 2583-1194

As indicated in Figure 8, this is an illustration of the cumulative frequency of several of the key terms that were used in the literature from 2017-2025. As illustrated by Figure 8, the trends in the frequency of the selected key words can be seen to demonstrate an increase in the application of certain key words over time, which correlates with the progression of focus on student performance prediction through the use of machine learning in the literature.

Key words that demonstrated increased frequency of use over time include "students", "machine learning" and "student performance". Following 2020, the frequency of use of these three key words was dramatically increased. This pattern of increase in the use of these key words parallels the broader acceptance of machine learning in education and the demand for predictive models within the educational sector [38].

Further, the rising frequency of key word usage such as "learning analytics", "education", and "prediction", indicate a sustained interest in how computational methods can be used to improve learning outcome and provide institutional decision makers with information about students and their learning [40].

The frequency of the use of the key words "classification", "academic performance", and "models" indicates that there has been a growing focus on developing and testing sophisticated models of prediction designed for education-specific data sets [39].

Overall, the increased frequency of usage of the above key words collectively represents an increase in research focusing on the integration of learning theory, data science and algorithmic modelling. Additionally, this trend suggests a transition from exploratory research to more formally designed and outcome focused research.

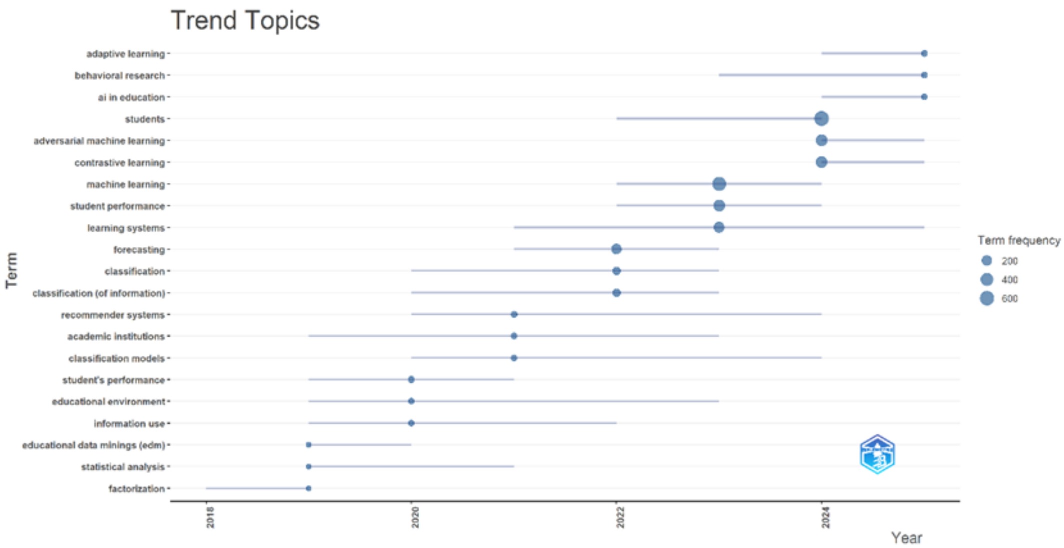

Figure 9. Trend of topics

As Figure 9 illustrates, there was a progression of the primary research subject matter over time relative to the use of machine learning for student performance prediction. The primary subject matter in early research (2018-2020), represented by the general terms "Student Performance", "Prediction" and "Academic Performance", illustrated the early stages of the field where most researchers were employing simple machine learning models to predict student outcomes using student data.

By 2021, we see a shift towards more specialized and sophisticated subjects; such as "Federated Learning," "Deep Learning," "Adversarial Learning" and "Contrastive Learning." As noted above, this trend represents an increasing emphasis on Algorithmic Innovation and the Use of More Complex Models to Address Educational Problems.

Further, the entry of "Personalized Education," "Intelligent Systems" and "Adaptive Learning" illustrate the increasing trend of combining learning analytics with instructional design. Additionally, these trends suggest that contemporary research is becoming increasingly concerned with both predicting student performance accurately and with practical implementation in teaching and learning environments.

The latter years also show an increase in "Educational Data Mining" and "Student Dropout Prediction," illustrating the expansion of the Field into Risk Analysis and Student Retention Strategies - both are critical Issues in Higher Education.

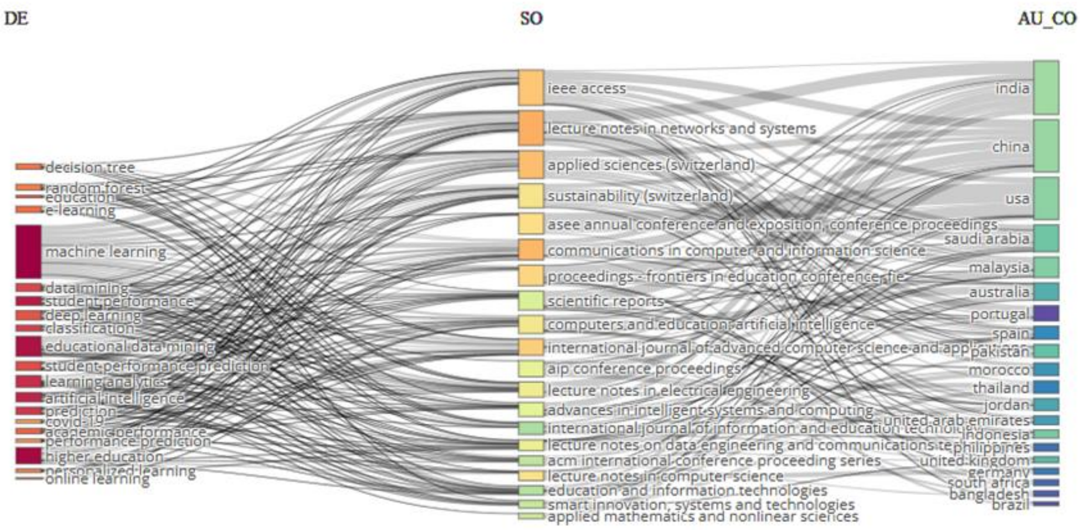

Figure 10. Journal-keyword-country three field plot (Source: Biblioshiny)

In Figure 10 we see a three-field layout representing three core components of our dataset: Keywords (left), Source Titles (middle), Countries (right). The three-field is a useful tool for examining the distribution of Research Themes (Keywords) across Journals and which countries are contributing to those themes.

The keywords most often cited in the leftmost section of the graph include: machine learning, student performance, prediction and academic performance. All of these appear to relate to several journals in the central section of the graph, showing that these subjects have been widely studied via papers in multiple publication outlets.

In addition, the central section of the graph includes some prominent Journal titles (IEEE Access, Lecture Notes in Networks and Systems, Applied Sciences (Switzerland)) which act as leading publication outlet for papers about Applying Predictive Models and Educational Data Mining Techniques.

On the right-hand section of the graph shows the top five countries from which contributions have come. The three largest contributors are India, China and the U.S., and all three contribute to a large number of keywords and source items, signifying a great deal of diversity and scope in their participation in this area. Several other countries (Saudi Arabia, Malaysia, Australia and Portugal) are also well-represented and therefore indicate an increasing geographic distribution of interest in this subject.

This graphic demonstrates the interrelatedness of research areas, communication channels and geographic origins. It provides a picture of an interdisciplinary and collaborative area of study, where the same or very similar topics are being researched in multiple countries and communicated through a wide array of scholarly publication vehicles.

4. TCCM FRAMEWORK ANALYSIS

Future Research Agenda Based on the TCCM Framework

The TCCM (Theory, Context, Characteristics, Method) framework is to be used to structure the analysis of the current state of literature regarding the use of Machine Learning to predict student performance in Higher Education. The TCCM framework provides a conceptual framework that includes four main dimensions Theory, Context, Characteristics, and Methodology; these provide a conceptual framework for examining the nature, scope and limitations of the research in this field.

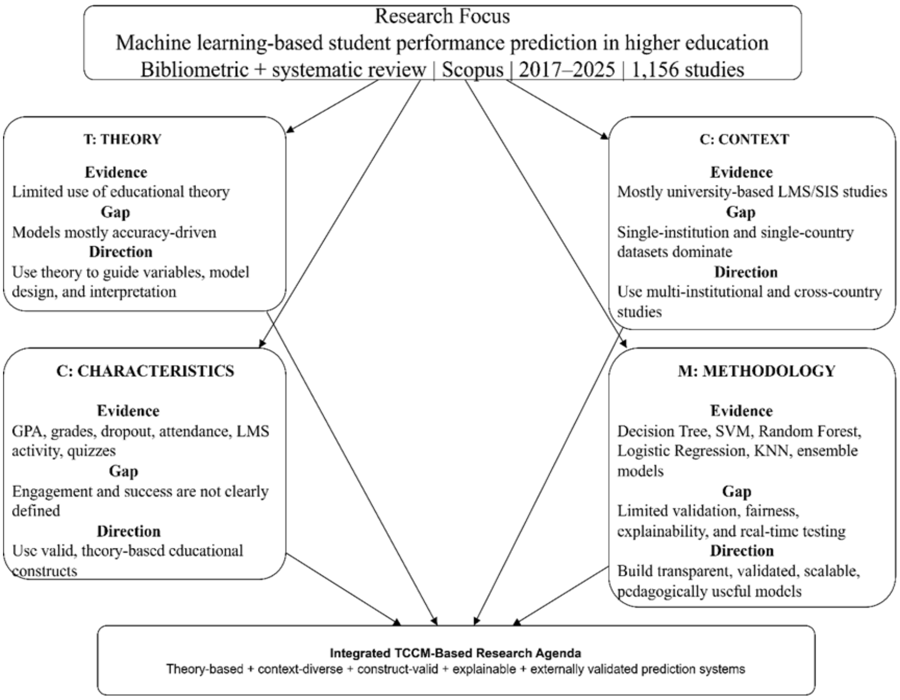

Figure:11. TCCM framework overview

Theory

Theoretical Foundations used to Guide Research on Educational Outcomes (Theory) – Only a few studies provided references to existing theoretical perspectives (e.g., constructivist views of learning, cognitive learning theory, behaviourist theory) [2]. The majority of the studies referenced these theoretical perspectives in a cursory manner to explain their results, rather than using them as the basis for how they developed their methodology and algorithms. For instance, although several studies referenced self-regulated learning theory or Bloom's taxonomy while describing student performance metrics, none of the studies utilized these theoretical frameworks to guide the selection of variables or choice of algorithms or interpretations of their predictive outputs. Overall, there appears to be little evidence that models are being driven by theory. As such, most models view machine learning as a solely technical solution with no relationship to educational processes [26]. The lack of an explicit connection between model development/interpretation and well-established educational theories limits the potential for these models to provide actionable recommendations for instruction and/or policy. It is imperative that future research utilizes existing educational theories as a foundation for developing and interpreting models of student success [36].

Context

The "Context" factor is about the environment in which each study took place, i.e., geography and institutional settings. Most of these studies were based in universities and concerned with students at either undergraduate or graduate levels. Much of the data used by researchers was collected through various digital learning tools, e.g., learning management systems (LMS), student information systems, etc., or as part of an assessment process carried out on-line. The research has largely been focused on a number of countries that include, but are not limited to, India, China, the U.S.A., Malaysia, and Indonesia [17]. Together, these countries represent a significant amount of the literature and suggest clusters of research activity and institutional interest in predictive analytics within these regions; however, many of the studies were based upon data from a single institution, or a single national context and very few involved collaborative research across institutions, or comparative analyses [20]. Therefore, relatively little predictive modelling work has been done in community colleges, vocational schools, and other non-traditional, informal learning environments. As a result, the findings of this body of research are subject to a number of limitations, and the applicability of those findings to different types of educational environments is reduced [10].

Characteristics

The characteristics dimension assesses the constructs, attributes, and dependent variables examined across all of the studies that were reviewed. Most of the studies evaluated academic performance metrics including GPA, course grades, test scores, the likelihood of dropping out, and retention rates. Many of the predictor variables most frequently utilized included attendance data, assignment completion records, frequency of logins to an institution's Learning Management System (LMS), student participation in online discussion forums, and students' performance on quizzes. Unfortunately, many of the predictor variables listed above were selected primarily because they were available via institutional databases and had no correlation to previously researched educational frameworks [23]. This selection process created a significant number of differences regarding the conceptualization and measurement of important terms/concepts (e.g., "student engagement" and "academic success") across the various studies. Only a few of the researchers in this body of research made an attempt to correlate their predictor variables to established theoretical constructs and/or utilize a well-established scale for assessing student motivation, self-efficacy, or cognitive load. The lack of definition and clarity regarding the constructs utilized across the various studies limits the ability to compare results across studies, and the lack of consistency in the use of terminology and methods limits the potential for developing cumulative knowledge about this topic [32]. In order to better advance our understanding of how to evaluate and predict student outcomes, future studies should have clearer definitions of the constructs being studied and utilize measurement tools that are valid and grounded in educational theory [6]

Methodology

While the Technical Dimension indicates that the application of Machine Learning in Higher Education has experienced substantial technical progress; the remaining three dimensions indicate several shortcomings of the literature. In particular, the Theoretical Dimension appears to be somewhat underdeveloped in terms of its relationship to Educational Theory. In fact, the majority of studies using Machine Learning did not clearly relate their ML models to the applicable Educational Theory (Kizilcec, 2023). Additionally, the Contextual Dimension of the literature is heavily concentrated upon a few geographic locations and/or institutional types. Thus, the generalizability of the findings is limited. Furthermore, the Methodological Dimension of the literature has experienced significant technical improvements but generally lacks the evaluation of the educational implications and/or the feasibility of practical implementation of the predictive models that were developed in the studies. Consequently, it is essential to address these short comings in order for the continued development and usefulness of predictive models in higher education [33].

Additionally, The Methodology dimension includes examination of the machine learning approaches, data management strategies, and assessment techniques used in the research. The vast majority of studies relied on Supervised Learning Methods (e.g., Decision Trees, Support Vector Machines, Random Forests, Logistic Regression, K-Nearest Neighbors, etc.). More recently, some researchers have included newer Machine Learning Architectures such as Federated Learning and Ensemble Models. Also, while studies evaluated the performance of their models using Accuracy, Precision, Recall, and F1-Score metrics, there was a general lack of discussion concerning how these metrics relate to real educational outcomes and how they can result in actionable interventions. One common limitation across all studies, however, was their reliance on single-institution datasets which were used extensively without external validation and replication [40]. Additionally, only a few studies were based upon longitudinal designs and/or involved real-time implementations, and only two studies mentioned fairness, explainability and scalability issues associated with the use of different models in varying institutional settings. Hence, this represents both technological maturity in developing models, as well as a lack of emphasis on deployment, educational relevance and interpretability [35].

In summary, therefore, the application of the TCCM framework identifies a number of important patterns in the literature. Although there has been a considerable increase in the amount of research activity, the theoretical dimension remains somewhat underdeveloped in terms of its relationship to educational theory. The contextual dimension of the literature is highly concentrated in a few regions and institutional types, and thus limits generalizability. Additionally, variable selection often lacks construct validity. Finally, although the methodological dimension shows a great deal of technical progress, it does not consistently evaluate the educational significance or practical implementation of the developed predictive models. Therefore, addressing these gaps will be necessary to continue the development and usefulness of predictive models in higher education [20].

5. CONTRIBUTION

Machine Learning to forecast student performance in higher education has several contributions to offer to the literature on forecasting student performance in higher education. First, it conducts an extensive and bibliometric review of 1,156 studies, published in the period 2017–2025, using the Scopus database. It does so by utilizing the TCCM (Theory, Context, Characteristics, Methodology) framework to provide an increased and more systematic understanding of the research field compared to previous reviews which generally focus on a single institution's data, or a single type of algorithm.

Second, the review highlights the lack of theoretical frameworks in most of the existing research and illustrates that while machine learning methods are widely applied, they are often formulated in the absence of relevant educational theories to enhance their explanatory power and relevance. This research offers the opportunity for researchers to extend its findings with the incorporation of pedagogical theories into model development and validation.

Third, the review shows how the vast majority of the research in this area was conducted in a small number of geographically situated and institutionally located contexts. As such, the findings illustrate the need for future research to be conducted in more geographically and institutionally diverse and inclusive contexts in order to increase the generalizability of predictive models to other educational systems.

Finally, the review critically examines the variables and outcomes employed in the predictive models under study to demonstrate the inconsistency in the definitions of the constructs and the selection criteria in the data collection processes. The findings also contribute to the continued efforts to develop and evaluate reliable and comparable predictive models of student performance.

Additionally, the review contributes to the methodological clarity of the area by documenting the most common algorithms, methods of evaluating performance, and methods of validating predictive models in the area.

6. LIMITATIONS OF THE STUDY

There are several limitations to the findings of this study. The limitation relates to the fact that the bibliometric data for this study was limited to the Scopus database; although Scopus is one of the most well-recognized and largest databases available today, it does not include every publication index found in other databases (e.g., Web of Science, IEEE Xplore, Google Scholar). As a result, the study may have missed some important research studies related to this topic.

7. CONCLUSION

A large-scale literature analysis was performed to provide an extensive overview of machine learning application in predicting student performance in higher education. Using the TCCM framework (Theory, Context, Characteristics, Methodology), 1,156 peer-reviewed articles published between 2017-2025 were reviewed using the Scopus database.

This study found that although there has been a substantial increase in the number of studies utilizing machine learning in educational settings, there is a considerable lack of theoretically based works. A majority of studies focused on developing the technical aspects of a model, but did not relate these efforts back to established educational theories. In terms of context, research is heavily concentrated within a small group of nations, and many studies are based upon a single institution's dataset limiting the generalizability of results. Furthermore, when selecting variables to utilize in models, the primary concern is typically data availability, rather than the use of consistently valid and well-defined constructs. Additionally, despite the variety of algorithms and evaluation metrics being used to evaluate models, model transparency, external validation, and real-time implementation are often ignored.

In addition to highlighting areas of concern with regard to applying machine learning in higher education, this study also points out a number of significant gaps in current literature that will need to be researched and addressed in the future. This study argues that it is time for researchers to move from developing purely algorithmic solutions and build upon their work to connect their solutions more directly to educational theory; to apply their solutions across different contexts; and to identify the pedagogical implications of predictive analytics. In order to further develop the field, researchers are required to clearly define the constructs they study, address ethics, and successfully integrate predictive models into the environments in which students learn. Ultimately, this study serves as a foundation for the next generation of studies to develop more contextualized, theoretically-based, and pedagogically meaningful applications of machine learning in higher education.

AUTHOR CONTRIBUTIONS

Pawan Painuli: Conceptualization, Investigation, Writing & Editing.

Dr. Amar Jeet Rawat: Supervision, Analysis

CONFLICT OF INTEREST

There is no conflict of interest for the publication of this research article among the authors. The financial, personal, professional or any other relationship has not affected the results or interpretation of this research.

ETHICS STATEMENT

The research has been carried out in accordance with the existing ethical norms and principles of responsible research. Participation in the research was on a voluntary basis and all the respondents were aware of the nature of the research and its extent. Informed consent has been taken from all the participants before conducting the research. The privacy and anonymity of the respondents have been ensured and the data collected have been used only for academic and research purposes. The personal identity of the participants has not been revealed at any stage. The research has followed the ethical norms relevant to the research.

DATA AVAILABILITY STATEMENT

Availability of data and materials

The datasets supporting the conclusions of this article are available on request from the corresponding author. The datasets are not publicly available because of privacy concerns of the participating respondents and institutions. Scientists who are interested in using the data can contact the corresponding author with a justified request.

REFERENCES

[1] Abuhassna, H. (2024). The Information Age for Education via Artificial Intelligence and Machine Learning: A Bibliometric and Systematic Literature Analysis. International Journal of Information and Education Technology, 14(5), 700–711. https://doi.org/10.18178/ijiet.2024.14.5.2095

[2] AI-Driven Student Performance Prediction Models in EdTech. (2025). International Journal for Research in Education, 14(10). https://doi.org/10.63345/ijre.v14.i10.3

[3] Almalawi, A., Soh, B., Li, A., & Samra, H. (2024). Predictive Models for Educational Purposes: A Systematic Review. Big Data and Cognitive Computing, 8(12), 187–187. https://doi.org/10.3390/bdcc8120187

[4] Bates, T., Cobo, C., Mariño, O., & Wheeler, S. (2020). Can artificial intelligence transform higher education? International Journal of Educational Technology in Higher Education, 17(1). https://doi.org/10.1186/s41239-020-00218-x

[5] Bond, M., Khosravi, H., Laat, M. de, Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. International Journal of Educational Technology in Higher Education, 21(1). https://doi.org/10.1186/s41239-023-00436-z

[6] Boujmiraz, S., Darhmaoui, H., & Maliani, A. D. el. (2026). Predicting student performance: A comprehensive review of machine learning, deep learning, and explainable AI approaches. Computers and Education Artificial Intelligence, 10, 100548–100548. https://doi.org/10.1016/j.caeai.2026.100548

[7] Duro, B., Gomes, A., Correia, F. B., Borges, A. R., & Bernardino, J. (2026). Machine Learning and Deep Learning for Dropout Prediction in Higher Education: A Review. Computers, 15(3), 164–164. https://doi.org/10.3390/computers15030164

[8] Eden, C. A., Chisom, O. N., & Adeniyi, I. S. (2024). Integrating AI in education: Opportunities, challenges, and ethical considerations. Magna Scientia Advanced Research and Reviews, 10(2), 6–13. https://doi.org/10.30574/msarr.2024.10.2.0039

[9] Ersozlu, Z., Taheri, S., & Koch, I. (2024). A review of machine learning methods used for educational data. Education and Information Technologies, 29(16), 22125–22145. https://doi.org/10.1007/s10639-024-12704-0

[10] Fahd, K., Venkatraman, S., Miah, S. J., & Ahmed, K. (2021). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. Education and Information Technologies, 27(3), 3743–3775. https://doi.org/10.1007/s10639-021-10741-7

[11] Hooshyar, D., Šír, G., Yang, Y., Kikas, E., Hämäläinen, R., Kärkkäinen, T., Gašević, D., & Azevedo, R. (2025a). Towards responsible AI for education: Hybrid human-AI to confront the Elephant in the room. In ArXiv.org. https://doi.org/10.48550/arxiv.2504.16148

[12] Hooshyar, D., Šír, G., Yang, Y., Kikas, E., Hämäläinen, R., Kärkkäinen, T., Gašević, D., & Azevedo, R. (2025b). Towards responsible AI for education: Hybrid human-AI to confront the elephant in the room. Computers and Education Artificial Intelligence, 9, 100524–100524. https://doi.org/10.1016/j.caeai.2025.100524

[13] Ifenthaler, D., Majumdar, R., Gorissen, P., Judge, M., Mishra, S., Raffaghelli, J. E., & Shimada, A. (2024). Artificial Intelligence in Education: Implications for Policymakers, Researchers, and Practitioners. Technology Knowledge and Learning, 29(4), 1693–1710. https://doi.org/10.1007/s10758-024-09747-0

[14] Ilić, M., Mikić, V., Kopanja, L., & Vesin, B. (2023). Intelligent techniques in e-learning: a literature review. Artificial Intelligence Review, 56(12), 14907–14953. https://doi.org/10.1007/s10462-023-10508-1

[15] Karim-Abdallah, B., Ayitey, M., Appiahene, P., Harris, E., & Binful, D. K. (2025). Application of Machine Learning Algorithms in Predicting Academic Performance of Students in Higher Education Institutes (HEIs): A Systematic Review and Bibliographic Analysis. African Journal Of Applied Research, 11(1), 536–559. https://doi.org/10.26437/ajar.v11i1.869

[16] Koedinger, K. R. (2023). Navigating Ethical Benefits and Risks as AIED Comes of Age. International Journal of Artificial Intelligence in Education, 34(1), 136–143. https://doi.org/10.1007/s40593-023-00350-5

[17] Lin, C.-C., Huang, A. Y. Q., & Lu, O. H. T. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. Smart Learning Environments, 10(1). https://doi.org/10.1186/s40561-023-00260-y

[18] Luan, H., Géczy, P., Lai, H., Gobert, J. D., Yang, S. J. H., Ogata, H., Baltes, J., Guerra, R. da S., Li, P., & Tsai, C. (2020). Challenges and Future Directions of Big Data and Artificial Intelligence in Education. Frontiers in Psychology, 11, 580820–580820. https://doi.org/10.3389/fpsyg.2020.580820

[19] Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. British Journal of Educational Technology, 50(6), 2824–2838. https://doi.org/10.1111/bjet.12861

[20] Maghsudi, S., Lan, A., Xu, J., & Schaar, M. van der. (2021). Personalized Education in the Artificial Intelligence Era: What to Expect Next. arXiv (Cornell University), 38(3), 37–50. https://doi.org/10.1109/msp.2021.3055032

[21] Nagy, M., & Molontay, R. (2023). Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention. International Journal of Artificial Intelligence in Education, 34(2), 274–300. https://doi.org/10.1007/s40593-023-00331-8

[22] Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. International Journal of Educational Technology in Higher Education, 20(1), 4–4. https://doi.org/10.1186/s41239-022-00372-4

[23] Pacheco, A. J., Figueredo, Ó. R. B., Chiappe, A., & Bedout, L. F. de. (2025). AI-powered learning analytics for metacognitive and socioemotional development: a systematic review. Frontiers in Education, 10. https://doi.org/10.3389/feduc.2025.1672901

[24] Pawar, M., & Dongardive, J. (2025). Leveraging Artificial Intelligence in Learning Management Systems: A Systematic Review of Applications and Impacts in Higher Education. Indian Journal of Science and Technology, 18(38), 3124–3135. https://doi.org/10.17485/ijst/v18i38.1648

[25] 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 & Management, 8(2), 21168–21168. https://doi.org/10.55267/iadt.07.13227

[26] Ramaswami, G., Sušnjak, T., & Mathrani, A. (2023). Effectiveness of a Learning Analytics Dashboard for Increasing Student Engagement Levels. Journal of Learning Analytics, 10(3), 115–134. https://doi.org/10.18608/jla.2023.7935

[27] Ramaswami, G., Sušnjak, T., Mathrani, A., & Umer, R. (2022). Use of Predictive Analytics within Learning Analytics Dashboards: A Review of Case Studies. Technology Knowledge and Learning, 28(3), 959–980. https://doi.org/10.1007/s10758-022-09613-x

[28] Remegio, A. N. (2025). Predictive Analytics in Higher Education: A Systematic Literature Review of Machine Learning Models for Forecasting Academic Performance and Ethical Integration. 213–218. https://doi.org/10.1109/iceit64364.2025.10976097

[29] Rodríguez-Ortiz, M. A., Santana‐Mancilla, P. C., & Anido, L. (2025). Machine Learning and Generative AI in Learning Analytics for Higher Education: A Systematic Review of Models, Trends, and Challenges. Applied Sciences, 15(15), 8679–8679. https://doi.org/10.3390/app15158679

[30] Rosé, C. P., McLaughlin, E. A., Liu, R., & Koedinger, K. R. (2019). Explanatory learner models: Why machine learning (alone) is not the answer. British Journal of Educational Technology, 50(6), 2943–2958. https://doi.org/10.1111/bjet.12858

[31] Roshanaei, M., Olivares, H., & Lopez, R. R. (2023). Harnessing AI to Foster Equity in Education: Opportunities, Challenges, and Emerging Strategies. Journal of Intelligent Learning Systems and Applications, 15(4), 123–143. https://doi.org/10.4236/jilsa.2023.154009

[32] Sajja, R., Sermet, Y., Cwiertny, D. M., & Demir, İ. (2025). Integrating AI and Learning Analytics for Data-Driven Pedagogical Decisions and Personalized Interventions in Education. Technology Knowledge and Learning. https://doi.org/10.1007/s10758-025-09897-9

[33] Saqr, M., & López‐Pernas, S. (2024). Why explainable AI may not be enough: predictions and mispredictions in decision making in education. Smart Learning Environments, 11(1). https://doi.org/10.1186/s40561-024-00343-4

[34] Sušnjak, T. (2023). Beyond Predictive Learning Analytics Modelling and onto Explainable Artificial Intelligence with Prescriptive Analytics and ChatGPT. International Journal of Artificial Intelligence in Education, 34(2), 452–482. https://doi.org/10.1007/s40593-023-00336-3

[35] Turkmenbayev, A., Abdykerimova, E., Nurgozhayev, S., Karabassova, G., & Baigozhanova, D. (2025). The application of machine learning in predicting student performance in university engineering programs: a rapid review. Frontiers in Education, 10. https://doi.org/10.3389/feduc.2025.1562586

[36] Ujkani, B., Ujkani, S., Minkovska, D., & Hinov, N. (2025). A Bibliometric Analysis of AI-Driven Performance Prediction in Higher Education. Information, 16(9), 713–713. https://doi.org/10.3390/info16090713

[37] Viberg, O., Kizilcec, R. F., Wise, A. F., Jivet, I., & Nixon, N. (2024). Advancing equity and inclusion in educational practices with AI ‐powered educational decision support systems (AI ‐ EDSS). British Journal of Educational Technology, 55(5), 1974–1981. https://doi.org/10.1111/bjet.13507

[38] Xu, X., & Thien, L. M. (2026). Artificial Intelligence in Higher Education: Insights From a Bibliometric Analysis and Systematic Review of Highly Cited Articles. European Journal of Education, 61(1). https://doi.org/10.1111/ejed.70520

[39] Yin, Z., Chinta, S. V., Wang, Z., Gonzalez, M., & Zhang, W. (2024). FairAIED: Navigating Fairness, Bias, and Ethics in Educational AI Applications. In arXiv (Cornell University). Cornell University. https://doi.org/10.48550/arxiv.2407.18745

[40] Zawacki‐Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16(1). https://doi.org/10.1186/s41239-019-0171-0