A Study on Data Mining Techniques to Improve Student Performance in Primary and Secondary Schools

Zahira Noor Quraishi[1], Dr. Atul Dattarya Newase[2]

[1]Research Scholar, [2]Research Supervisor
[1,2]Dr. A. P. J. Abdul Kalam University, Indore, India
zahira16sep@gmail.com, dr.atulnewase@gmail.com

Presented at International Conference on Trends and Innovations in Management, Engineering, Sciences and Humanities (ICTIMESH-24), London, 24-27 June 2024.

Published in IJIRMPS (E-ISSN: 2349-7300), ICTIMESH-24

Abstract

Educational institutions generate large volumes of data related to student demographics, academic records, attendance, and learning behavior. Effectively analyzing this data can support early identification of learning difficulties and improve educational outcomes. This study explores the role of data mining techniques in enhancing student performance at primary and secondary school levels, where timely intervention is most critical. Based on an extensive review of recent Educational Data Mining (EDM) literature, the study examines commonly used methods such as decision trees, random forest, naïve Bayes, artificial neural networks, and feature selection approaches applied to student performance prediction. Existing research highlights that while higher education dominates EDM applications, primary and secondary education remain under-explored despite their long-term impact on learning trajectories. The findings indicate that student behavioral data, academic history, and demographic attributes are strong predictors of performance, and that applying feature selection significantly improves model accuracy. The study emphasizes the importance of data-driven decision-making for educators and policymakers to identify at-risk students and design targeted interventions. By synthesizing current trends, challenges, and gaps, this research provides a foundation for developing effective data mining models aimed at improving learning outcomes and strengthening early-stage education systems.

Keywords: Educational Data Mining, Student Performance Prediction, Primary Education, Secondary Education, Machine Learning

1. Introduction

Education plays a crucial role in shaping human capital and socio-economic development. In recent years, the rapid digitization of educational processes has led to the generation of massive amounts of educational data, including enrollment records, attendance logs, examination scores, and learning management system interactions. Managing and analyzing this data using traditional statistical

approaches has become increasingly challenging. As a result, **Educational Data Mining (EDM)** has emerged as a specialized field that applies data mining and machine learning techniques to discover meaningful patterns from educational datasets [1].

Student performance is one of the most widely studied problems in EDM because of its direct impact on academic success, dropout prevention, and institutional effectiveness. Predicting student performance at an early stage enables educators to identify learners who are at risk of failure and to provide timely academic support. While significant progress has been made in applying EDM techniques to higher education, studies focusing on **primary and secondary education** remain limited. This gap is critical, as early educational stages strongly influence long-term academic achievement and career outcomes [2].

Recent literature reveals that most EDM studies concentrate on performance prediction using traditional machine learning models such as decision trees, naïve Bayes, support vector machines, random forest, and artificial neural networks. These models have demonstrated high accuracy in predicting academic outcomes when trained on structured datasets. Additionally, research highlights the importance of **feature selection techniques**, such as Boruta, Recursive Feature Elimination, and Lasso regression, in improving model performance by removing irrelevant attributes and reducing computational complexity [3].

Another important observation from the literature is the imbalance in research focus across educational levels. A majority of studies analyze tertiary education datasets, whereas secondary education receives moderate attention and primary education remains significantly underrepresented. Furthermore, many existing studies rely on localized or case-specific datasets, limiting the generalizability of their findings. Advanced techniques such as deep learning, learning analytics dashboards, and knowledge tracing models are still rarely applied in school-level education due to data availability constraints and implementation complexity [4].

Despite these limitations, EDM has demonstrated strong potential in improving decision-making within education systems. Data-driven insights can support curriculum planning, resource allocation, personalized learning, and policy formulation. In primary and secondary schools, EDM applications can assist in monitoring attendance patterns, evaluating assessment outcomes, and identifying behavioral factors that influence learning performance. Such insights are particularly valuable in developing regions, where educational challenges such as low retention rates and uneven learning outcomes persist [5].

This study aims to examine and synthesize existing research on data mining techniques used for improving student performance in primary and secondary education. By analyzing trends, commonly used methodologies, and identified research gaps, the study seeks to highlight the need for broader adoption of EDM at early educational levels. The outcomes of this research are expected to guide future studies in designing robust, scalable, and interpretable data mining models that can support educators in improving learning outcomes and strengthening school education systems.

2. Literature review

Kaur et al. (2023), Educational Data Mining (EDM) and Learning Analytics are increasingly used to address critical educational challenges such as student performance prediction and assessment. This systematic literature review analyzes studies published between 2012 and 2022 to examine techniques employed in both domains. EDM focuses on applying data mining methods to educational data, while learning analytics emphasizes understanding learning processes through visualization and quantitative analysis. Reviewing 41 relevant studies, the paper identifies commonly used statistical, machine learn-

ing, and visualization techniques, highlights research gaps, and provides guidance for educators, researchers, and policymakers seeking to enhance learning outcomes through data-driven approaches. **[1]**

Missah et al. (2023), The application of Data Mining (DM) in education supports evidence-based decision-making by educational leaders. This review examines how DM research is distributed across different educational levels, with a particular focus on Basic Education (BE). Analyzing 94 studies published between 2017 and 2022 from nine major academic publishers, the findings reveal a strong imbalance: most EDM research targets tertiary education, while basic and pre-tertiary levels remain underrepresented. Additionally, existing studies focus heavily on student performance factors, overlooking critical aspects such as pedagogical resources, revealing both population and knowledge gaps in EDM research. **[2]**

Abideen et al. (2023), Student enrollment analysis is a critical but underexplored area in educational data mining, particularly in developing regions. This study focuses on predicting school enrollment in Punjab, Pakistan, using five years of data from 100 schools. Machine learning algorithms—Multiple Linear Regression, Random Forest, and Decision Tree—are applied to identify significant features and predict future enrollment trends. The proposed model supports enrollment target classification and provides insights to address low enrollment levels, thereby contributing to improved planning, literacy rates, and more effective education policy implementation. **[3]**

Syed Mustapha et al. (2023), Accurate prediction of academic success increasingly depends on effective feature engineering and selection in machine learning models. This study compares feature selection methods—Boruta and Lasso for regression, and Recursive Feature Elimination (RFE) and Random Forest Importance (RFI) for classification—using the OULA dataset. Results show that Gradient Boost with Boruta achieved the lowest prediction error, while RFI produced the highest classification accuracy. The findings emphasize that appropriate feature selection significantly improves model performance, offering valuable guidance for developing reliable student success prediction systems. **[4]**

Alghamdi et al. (2023), Predicting academic performance in early secondary education can help institutions identify at-risk students and provide timely support. This study uses data from high school graduates in the Al-Baha region of Saudi Arabia to develop predictive models using Naïve Bayes, Random Forest, and J48 algorithms. Data balancing with SMOTE and feature extraction using correlation coefficients improved model reliability. Performance evaluation through cross-validation showed exceptionally high accuracy, with Naïve Bayes achieving 99.34%, demonstrating the effectiveness of EDM techniques in early academic performance prediction. **[5]**

Chan et al. (2023), Despite the rapid growth of Educational Data Mining (EDM), its application in secondary school contexts remains limited. This literature review analyzes 18 studies published between 2008 and 2021 focusing on secondary school data. Most studies address academic success classification, influence factor analysis, and dropout risk prediction using traditional machine learning methods. Advanced techniques such as deep learning and knowledge tracing are rarely applied. The review identifies a clear research gap and emphasizes the need to expand EDM research into secondary education to better support learning and policy decisions. **[6]**

Huerta et al. (2023), This work applies the Knowledge Discovery in Databases (KDD) methodology using RapidMiner to filter and organize information for more efficient decision-making, with a focus on historical investment per student in the education sector. By mining patterns from stored data, the study aims to reduce waste caused by poor information management and support more accurate prediction. Findings indicate that expenditure per student generally increases over time, though allocation differs by province, while still showing an upward trend overall. The study concludes that KDD-based analysis can visualize spending variation across education grades and provide relevant insights for future research and planning. **[7]**

Alsulami et al. (2023), Focusing on EDM research between 2020 and 2022, this review identifies key factors influencing student performance and the most commonly used EDM methods. The analysis concludes that student behaviors are the strongest contributors to academic performance compared with other factors. It also finds that the most frequently used classifiers for predicting student performance include decision trees, multilayer perceptron models, and support vector machines. By summarizing recent trends, the paper offers a snapshot of dominant features and modeling approaches and supports researchers in selecting methods aligned with current practice in EDM-based performance prediction. **[8]**

Martinez-Comesana et al. (2023), This systematic review synthesizes research on how AI tools improve assessment of primary and secondary students. From 2010 to 2023, nine original studies (641 participants) met inclusion criteria. The main contributions of AI in lower-level assessment include predicting performance, automating evaluations to improve objectivity (e.g., with neural networks or natural language processing), using educational robots to analyze learning processes, and identifying factors that make classes more engaging. Overall, the review demonstrates existing, practical applications of AI that can strengthen assessment quality and learning experiences at early educational levels. **[9]**

Batool et al. (2023), Educational data mining supports proactive interventions by predicting student achievement before final exams, helping reduce dropout risk and improve outcomes. This paper reviews about 260 studies over the past 20 years, comparing major influencing factors, prediction and feature selection techniques, and frequently used tools. It reports that ANN and Random Forest are the most commonly used algorithms, and WEKA is a widely used tool. Academic records and demographics are highlighted as strong predictors. The review also shows irrelevant features degrade accuracy and increase processing time, explaining why many studies apply feature selection before modeling and offering guidance for future EDM research and implementation. **[10]**

Ampadu et al. (2023), Educational Data Mining (EDM) emerged as a response to the explosive growth of educational data in the big-data era, where institutions store large volumes of enrollment, attendance, and examination records. Because educational settings have unique goals and constraints, traditional data mining cannot always be applied directly, motivating specialized approaches and algorithms. The reviewed work highlights EDM applications that support stakeholders by identifying at-risk students, prioritizing learning needs across groups, improving graduation rates, monitoring institutional performance, managing campus resources, and guiding curriculum renewal. Overall, it surveys key methodologies used to extract knowledge from higher-education datasets for practical decision-making. **[11]**

Table 1. Systematic literature review

Ref	Author (First author et al.)	Year	Title	Methods	Result	Advantage	Limitation
[1]	Kaur et al.	2023	Role of educational data mining and learning analytics techniques used for predictive modeling	Systematic literature review (2012–2022); compares EDM vs Learning Analytics techniques (statistical, ML, visualization) across 41 studies	Identified commonly used techniques, research gaps, and guidance for stakeholders to enhance learning outcomes	Clear synthesis of EDM vs LA approaches and technique landscape; highlights gaps and provides practical guidance	Limited to 41 studies and 2012–2022 window; findings depend on included study quality/coverage
[2]	Missah et al.	2023	Systematic Review of Data Min-	Systematic review of 94 studies (2017–2022)	Found imbalance: EDM research heavily focused on	Identifies population + knowledge	Scope constrained to selected publish-

			ing in Education on the Levels and Aspects of Education	across 9 academic publishers; analyzes distribution by education level and research aspects	tertiary education; basic/pre-tertiary underrepresented; focus mostly on performance factors	gaps; helps redirect EDM research to neglected levels/ aspects	ers and 2017–2022; review highlights gaps but does not propose/validate models
[3]	Abideen et al.	2023	Analysis of enrollment criteria in secondary schools using machine learning and data mining approach	ML prediction using 5-year dataset from 100 schools (Punjab, Pakistan); Multiple Linear Regression, Random Forest, Decision Tree; feature significance + trend prediction	Model predicts future enrollment and supports enrollment target classification; offers insights to address low enrollment	Practical application in underexplored enrollment analytics; supports planning and policy decisions	Region-specific dataset (Punjab) may limit generalizability; limited algorithm set and depends on data quality/ coverage
[4]	Syed Mustapha et al.	2023	Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods	Comparative feature selection on OULA dataset: Boruta & Lasso (regression), RFE & Random Forest Importance (classification); models include Gradient Boost	Gradient Boost + Boruta gave lowest prediction error; RFI yielded highest classification accuracy; feature selection improves performance	Strong evidence that feature selection boosts accuracy; provides guidance on selecting FS methods for student success prediction	Results tied to OULA dataset and chosen models; may not transfer directly to other contexts/ datasets without validation
[5]	Alghamdi et al.	2023	Data mining approach to predict success of secondary school students: A Saudi Arabian case study	Naïve Bayes, Random Forest, J48; SMOTE for class balancing; correlation-based feature extraction; cross-validation evaluation	Very high predictive accuracy reported; Naïve Bayes achieved **99.34%**	Demonstrates strong performance with balancing + feature extraction; supports early identification of at-risk students	Extremely high accuracy may indicate dataset bias/leakage risk; limited to one region/cohort (Al-Baha), affecting generalizability
[6]	Chan et al.	2023	A literature review on educational data mining with secondary school data	Literature review of 18 studies (2008–2021) on secondary school EDM; analyzes tasks (success classification, factor analysis, dropout prediction)	Secondary-school EDM is limited; mostly traditional ML; deep learning/knowledge tracing rarely used; highlights research gap	Clarifies state-of-the-art and gaps specifically for secondary education; motivates advanced method adoption	Small set of studies (18); review period ends 2021 so newer advances may be missing
[7]	Huerta et al.	2023	Data mining: Application of digital marketing in education	Knowledge Discovery in Databases (KDD) methodology using RapidMiner; pattern mining/visual-	Expenditure per student generally increases over time; allocation varies by province but shows upward trend; visu-	Uses KDD to improve decision-making and reduce information waste; pro-	Focus is descriptive/organizational; limited evidence of predictive validity; con-

				ization of historical investment per student	alization supports planning	duces interpretable spending patterns	clusions depend on completeness of financial data
[8]	Alsulami et al.	2023	Using Data Mining Techniques To Enhance The Student Performance. A semantic review	Semantic review (2020–2022) summarizing key performance factors and common EDM methods/classifiers	Student behaviors strongest predictors; common classifiers: decision trees, MLP, SVM	Quick snapshot of recent trends; helps researchers choose commonly used methods/features	Short time window (2020–2022) may miss broader trends; "semantic review" may be less rigorous than full systematic review
[9]	Martinez-Comesana et al.	2023	Impact of artificial intelligence on assessment methods in primary and secondary education: Systematic literature review	Systematic review (2010–2023); 9 original studies, 641 participants; covers AI tools for assessment (NN, NLP, robots, etc.)	AI supports performance prediction, automated/objective evaluation, learning-process analysis, and engagement factor identification	Demonstrates concrete AI assessment applications in early education; emphasizes improved objectivity and insight	Only 9 studies met criteria → limited evidence base; heterogeneity across tools/studies may limit general conclusions
[10]	Batool et al.	2023	Educational data mining to predict students' academic performance: A survey study	Survey/review of ~260 studies (≈20 years); compares factors, prediction + feature selection techniques; tools noted (e.g., WEKA)	ANN and Random Forest most common; academic records & demographics strong predictors; irrelevant features reduce accuracy/time	Broad, comprehensive synthesis across two decades; highlights practical tooling and importance of feature selection	Breadth may reduce depth on individual methods; results depend on surveyed literature quality and may underrepresent newer deep learning
[11]	Ampadu et al.	2023	Handling big data in education: a review of educational data mining techniques for specific educational problems	Review of EDM techniques for big educational datasets; discusses specialized approaches vs traditional DM; applications in higher education	EDM supports identifying at-risk students, improving graduation rates, resource management, performance monitoring, curriculum renewal	High-level mapping of EDM applications for decision-making in higher education contexts	Broad review with limited methodological detail; may not provide quantitative comparisons or implementation specifics

3. Research Gap

Across the reviewed studies, a clear research gap emerges around underrepresented contexts and methodological depth: multiple reviews report that EDM research is imbalanced toward tertiary education, with basic/pre-tertiary and secondary settings receiving comparatively less attention despite

their policy and intervention importance. Even within secondary education, literature remains relatively small and focused on traditional machine-learning tasks (e.g., successful classification, dropout prediction), while advanced approaches (e.g., deep learning, knowledge tracing) are rarely applied and insufficiently validated. In addition, several works synthesize trends but highlight limited translation into robust, generalizable, real-world deployments, because results are often tied to specific datasets/regions (e.g., localized enrollment or case-study cohorts) and may lack external validation across diverse populations. Finally, assessment-focused AI evidence at primary/secondary levels remains thin (few qualifying studies), indicating a need for more rigorous, larger-scale evaluations of AI-driven assessment tools, fairness, and impact on learning outcomes.

4. Systematic result analysis

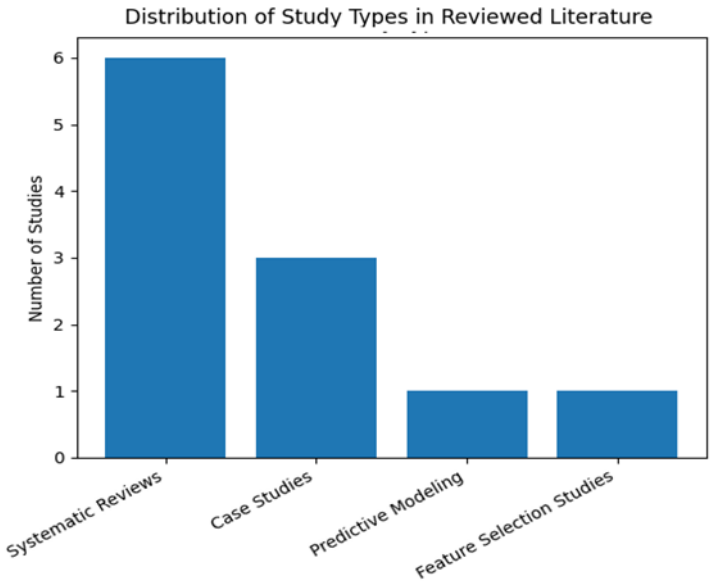

Figure 1. Distribution of Study Types in Reviewed Literature

Figure 1 illustrates the distribution of study types across the reviewed EDM literature. The majority of the studies are **systematic and survey-based reviews**, indicating a strong emphasis on synthesizing existing knowledge rather than proposing novel models. In contrast, **predictive modeling and feature selection studies** are comparatively limited, highlighting a need for more empirical and implementation-focused research.

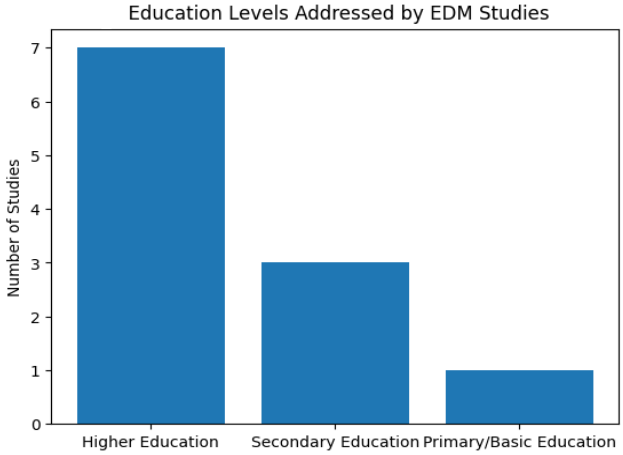

Figure 2: Education Levels Addressed by EDM Studies

Figure 2 presents the distribution of EDM studies across different education levels. Most studies focus on **higher education**, while **secondary education** receives moderate attention and **primary/basic education** remains significantly underrepresented. This imbalance suggests a critical research gap in applying EDM techniques to early educational stages, where early intervention could have substantial long-term impact.

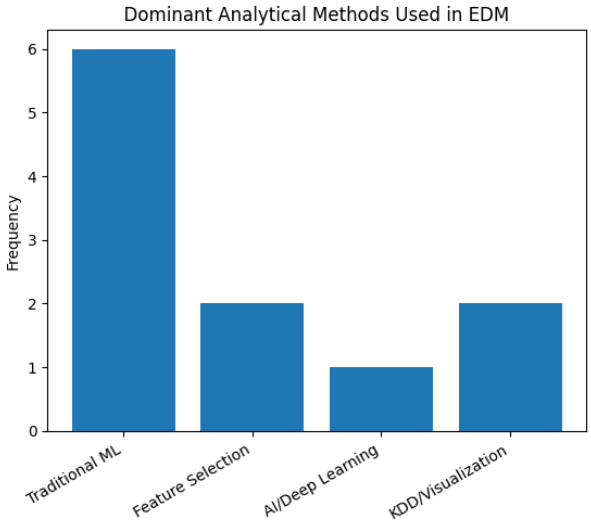

Figure 3: Dominant Analytical Methods Used in EDM

Figure 3 shows the frequency of analytical methods employed in the reviewed studies. **Traditional machine learning algorithms** dominate EDM research, whereas **advanced AI and deep learning techniques** appear infrequently. Feature selection and KDD/visualization methods are used to a limited extent, indicating opportunities to integrate more sophisticated and hybrid analytical approaches in future EDM research.

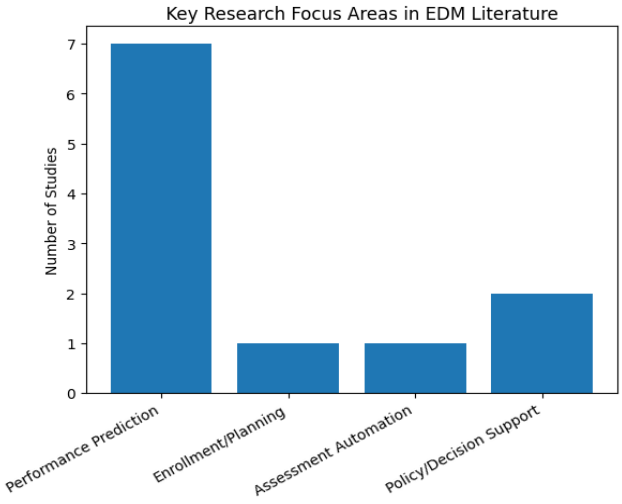

Figure 4: Key Research Focus Areas in EDM Literature

Figure 4 summarizes the main research focus areas within the reviewed literature. Student performance prediction is the most prominent theme, reflecting its importance for academic decision-making. However, areas such as enrollment planning, assessment automation, and policy/decision support receive considerably less attention, revealing underexplored domains where EDM could provide high practical value.

5. Conclusion

This study highlights the growing importance of data mining techniques in improving student performance, particularly within primary and secondary education systems. The review of existing literature reveals that while Educational Data Mining has been extensively applied in higher education, its adoption at early educational stages remains limited. Commonly used techniques such as decision trees, random forest, naïve Bayes, and artificial neural networks have shown strong potential for predicting academic performance, especially when combined with effective feature selection methods. The findings also indicate that behavioral, academic, and demographic factors play a significant role in determining student outcomes. However, challenges such as limited datasets, lack of generalizability, and underuse of advanced analytical models restrict the full potential of EDM in school-level education. Addressing these gaps can enable early identification of at-risk students and support data-driven interventions to enhance learning outcomes. Future work should focus on developing scalable, explainable, and region-specific EDM frameworks using real-time school data to support continuous performance improvement.

References

1. Kaur, Kanksha, and Omdev Dahiya. "Role of educational data mining and learning analytics techniques used for predictive modeling." In 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), pp. 1-6. IEEE, 2023.
2. Missah, Yaw Marfo, Fuseini Inusah, Najim Ussiph, and Twum Frimpong. "Systematic Review of Data Mining in Education on the Levels and Aspects of Education." (2023).
3. Abideen, Z.U., Mazhar, T., Razzaq, A., Haq, I., Ullah, I., Alasmary, H. and Mohamed, H.G., 2023. Analysis of enrollment criteria in secondary schools using machine learning and data mining approach. *Electronics*, *12*(3), p.694.
4. Syed Mustapha, S. M. F. D. "Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods." *Applied System Innovation* 6, no. 5 (2023): 86.
5. Alghamdi, Amnah Saeed, and Atta Rahman. "Data mining approach to predict success of secondary school students: A Saudi Arabian case study." *Education Sciences* 13, no. 3 (2023): 293.
6. Chan, Ka Ian, Philip IS Lei, and Patrick Cheong-Iao Pang. "A literature review on educational data mining with secondary school data." In *Proceedings of the 9th International Conference on Education and Training Technologies*, pp. 1-7. 2023.
7. Huerta, C.M., Atahua, A.S., Guerrero, J.V. and Andrade-Arenas, L., 2023. Data mining: Application of digital marketing in education. *Advances in Mobile Learning Educational Research*, *3*(1), pp.621-629.
8. Alsulami, Abdulkream, Abdullah S. Al-Malaise Al-Ghamdi, and Mahmoud Ragab. "Using Data Mining Techniques To Enhance The Student Performance. A semantic review." In *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, pp. 1-5. IEEE, 2023.
9. Martinez-Comesana, Miguel, Xurxo Rigueira-Díaz, Ana Larranaga-Janeiro, Javier Martínez-Torres, Iago Ocarranza-Prado, and Denis Kreibel. "Impact of artificial intelligence on assessment methods in primary and secondary education: Systematic literature review." *Revista de Psicodidáctica (English ed.)* 28, no. 2 (2023): 93-103.
10. Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.Y. and Hussain, A., 2023. Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), pp.905-971.

11. Ampadu, Yaw Boateng. "Handling big data in education: a review of educational data mining techniques for specific educational problems." *AI, Computer Science and Robotics Technology* 13 (2023).