



An XGBoost-Based Learning Analytics Framework for Predicting College Students' Information Literacy Performance

Bharghavi tarra¹, Sk.Mahammadunnisa²

¹MCA Student , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad

²Assistant Professor , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad

Abstract— Information literacy is an essential competency for students in modern higher education, enabling effective access, evaluation, and application of digital information. This study proposes a machine learning-based framework to predict students' information literacy performance using learning behaviour data. The collected dataset is preprocessed through data cleaning, normalization, and feature selection to improve model performance. Several classification algorithms, including Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes, Neural Network, Random Forest, and XGBoost, are implemented and evaluated using performance metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the XGBoost model outperforms the other classifiers by providing higher prediction accuracy and better generalization. The proposed framework supports early identification of students requiring academic intervention and assists educators in developing personalized learning strategies. This approach contributes to intelligent learning analytics and promotes data-driven educational decision-making, ultimately enhancing students' information literacy and overall academic performance.

Keywords----Information Literacy, Learning Behaviour Analysis, Machine Learning, XGBoost, Educational Data Mining, Learning Analytics.

I. INTRODUCTION

Information literacy has become one of the most important competencies for students in the digital era [1], [2]. The rapid growth of digital technologies, online learning platforms, and internet-based educational resources has transformed the way students acquire, process, and apply information [3], [4], [5]. Today, students are expected not only to access information but also to evaluate its reliability, analyze its relevance, and use it effectively for academic and professional purposes [1], [5], [16].

However, differences in learning behaviour, motivation, and engagement often influence students' ability to develop strong information literacy skills [6], [9]. Educational institutions require effective methods to monitor these behavioural patterns and identify students who may

need additional academic support [8], [9].

Traditional assessment methods mainly depend on examinations and manual observations, which are often time-consuming and unable to provide timely insights into student performance. Therefore, intelligent data-driven approaches have become increasingly important for understanding learning behaviour and improving educational outcomes through predictive analysis [8], [10].

Recent advancements in machine learning have created new opportunities for analyzing educational data and predicting student performance with greater accuracy [8], [10]. Machine learning algorithms can automatically discover meaningful patterns from historical learning data and generate predictions without relying solely on predefined rules [8], [12]. These techniques have been successfully applied in educational data mining, learning analytics, student performance prediction, dropout detection, and personalized learning recommendation systems [8], [9], [12]. By examining behavioural indicators such as participation, learning activities, information usage, and engagement levels, predictive models can estimate future learning outcomes and help educators make informed decisions [6], [9]. The ability to identify students who are at academic risk before final examinations enables timely intervention and improves the overall teaching process [8], [9]. As educational institutions continue to generate large volumes of digital learning data, machine learning has emerged as a reliable and scalable solution for supporting evidence-based educational decision-making [10], [12].

Several machine learning algorithms have been explored for predicting students' academic performance, including Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes, Neural Networks, and Random Forest [6], [8], [9]. While these



methods provide satisfactory prediction results, their performance may vary depending on the characteristics of the educational dataset [6], [8]. Ensemble learning algorithms have demonstrated better predictive capability by combining multiple decision trees and reducing classification errors [8]. Among these techniques, Extreme Gradient Boosting (XGBoost) has gained significant attention because of its high computational efficiency, robustness, and ability to handle complex relationships within large datasets. XGBoost incorporates gradient boosting with regularization techniques, enabling improved prediction accuracy while minimizing overfitting. Its capability to process missing values, optimize feature importance, and achieve faster model training makes it particularly suitable for educational prediction tasks involving diverse behavioural attributes [8], [12]. Consequently, XGBoost has become a preferred algorithm for many modern predictive analytics applications [12].

This research proposes a machine learning framework for predicting college students' information literacy performance based on their learning behaviour characteristics. Initially, the collected educational dataset undergoes preprocessing steps, including missing value handling, normalization, and feature selection, to improve data quality and eliminate irrelevant information. Multiple classification algorithms, namely Decision Tree, KNN, Naïve Bayes, Neural Network, Random Forest, and XGBoost, are trained and evaluated using standardized performance metrics such as accuracy, precision, recall, and F1-score. Comparative analysis is conducted to determine the most effective prediction model for the given dataset. The experimental findings indicate that XGBoost consistently delivers superior predictive performance compared with the conventional machine learning models [6], [8], [12]. The proposed framework provides a systematic approach for identifying students who require academic guidance and enables educators to make timely instructional interventions based on reliable predictive insights [9], [12].

The proposed system contributes to the advancement of intelligent learning analytics by supporting data-driven educational planning and personalized learning strategies [8], [10], [12]. Accurate prediction of students' information literacy performance allows instructors to design targeted interventions, improve teaching methodologies, and enhance students' learning experiences [12], [13]. Educational administrators can also utilize prediction results to allocate learning resources more efficiently and implement evidence-based academic policies [10], [11]. Furthermore, the integration of advanced machine learning techniques into

educational environments promotes continuous monitoring of student progress and facilitates early identification of learning challenges [8], [10], [12]. As digital education continues to expand, predictive analytics will play a significant role in improving educational quality and student success [3], [4], [10]. The proposed XGBoost-based framework offers a practical, scalable, and reliable solution for predicting information literacy performance while contributing to the broader adoption of artificial intelligence and learning analytics in higher education systems [10], [12].

II.RELATED WORK

"Machine Learning-Based Evaluation of Information Literacy Enhancement among College Teachers," J. Li [12]

This study presented a machine learning-based framework to evaluate the improvement of information literacy among college teachers. The proposed model analyzed teaching activities, digital resource utilization, and instructional performance using machine learning algorithms to assess information literacy levels. The results demonstrated that machine learning techniques provide accurate evaluation of educators' digital competencies and support educational institutions in designing effective professional development programs. The proposed approach improved evaluation efficiency and promoted data-driven decision-making for enhancing information literacy.

"Research on College Students' Information Literacy Based on Big Data," Y. Ying [14]

This study explored the application of big data technologies to improve information literacy among college students. The proposed framework collected and analyzed large-scale educational data to understand students' learning patterns, information access behavior, and digital competency. The research showed that big data analytics can effectively identify learning deficiencies and provide personalized educational support. Experimental findings indicated that data-driven learning strategies significantly improved students' information literacy and academic performance.



"Research Status, Hot Spots and Enlightenment of College Students' Information Literacy: Based on Bibliometric Analysis of CNKI from 2000 to 2021," G. F. Yang et al. [11]

This study conducted a bibliometric analysis of research related to college students' information literacy published between 2000 and 2021. The authors identified major research trends, emerging topics, and future research directions in information literacy education. The analysis revealed a growing interest in digital literacy, learning analytics, and intelligent educational technologies. The study concluded that integrating artificial intelligence and machine learning into information literacy education can significantly improve teaching effectiveness and student learning outcomes.

"A Meta-Analysis of Educational Data Mining on Improvements in Learning Outcomes," I. A. AlShammari et al. [8]

This study performed a meta-analysis of educational data mining techniques used to improve students' learning outcomes. Various machine learning algorithms were evaluated for predicting academic performance, identifying at-risk students, and supporting personalized learning. The findings indicated that educational data mining significantly enhances prediction accuracy and enables timely academic interventions. The study emphasized the importance of data-driven learning analytics for improving educational quality and student success.

"Mining the Characteristics of Meaningful Learning Behavior: A Framework for Predicting Learning Outcomes," Wu Fati and Tian Hao [6]

This study proposed a learning behavior analysis framework for predicting students' learning outcomes using machine learning techniques. The framework analyzed students' behavioral characteristics and applied classification algorithms to predict academic performance based on learning activities. Experimental results showed that learning behavior features have a strong influence on prediction accuracy, with Random Forest achieving superior performance among the evaluated models. The study demonstrated that behavioral analytics can effectively support early intervention, personalized learning, and evidence-based educational decision-making.

III.DATASET DETAILS

The dataset used in this study contains information related to the learning behaviour and information literacy performance of college students. It includes various behavioural and academic attributes

collected from students during their participation in information literacy courses. The dataset consists of approximately 320 student records, where each record represents the learning activities and corresponding performance of an individual student. The collected attributes describe different aspects of students' learning behaviour, including information thinking ability, information acquisition skills, learning participation, information application capability, and overall learning engagement. These features are used to understand how students interact with learning resources and how their behaviour influences their academic performance. The target variable in the dataset represents the students' information literacy outcome, which is classified into different performance categories. Since the dataset contains both numerical and categorical values, categorical attributes are converted into numerical form using label encoding before model training. The dataset serves as the primary source for training and evaluating multiple machine learning algorithms to predict students' information literacy performance. It provides sufficient behavioural information to build reliable predictive models while supporting comparative analysis among different classification techniques.

Before model development, the dataset undergoes several preprocessing steps to improve data quality and ensure accurate prediction. Initially, missing values are handled by replacing unavailable entries with suitable values to avoid information loss during model training. Feature normalization is then performed using the **Min-Max Scaling** technique, which transforms all numerical attributes into a common range between 0 and 1, preventing features with larger values from dominating the learning process. Correlation analysis is applied to identify relationships among features and eliminate attributes with very low significance, thereby reducing redundancy and improving computational efficiency. After preprocessing, the dataset is divided into **80% training data** and **20% testing data**, allowing the machine learning models to learn from historical patterns while validating their predictive performance on unseen records. Multiple classification algorithms, including Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes, Neural Network, Random Forest, and XGBoost, are trained using the processed dataset. The quality and diversity of the dataset enable the models to capture meaningful learning behaviour patterns and accurately predict students' information literacy performance. This structured dataset plays a crucial role in developing an effective and reliable educational prediction framework.

IV. PROPOSED METHODOLOGY



The proposed methodology presents a machine learning framework for predicting college students' information literacy performance using learning behaviour characteristics. Initially, the educational dataset is collected and examined to identify the behavioural and academic attributes that influence students' learning outcomes. The raw dataset undergoes several preprocessing operations, including missing value handling, duplicate removal, categorical data encoding, and feature normalization using the Min-Max Scaling technique. Correlation analysis is then performed to identify the most relevant features while eliminating redundant or low-impact attributes, thereby improving the efficiency of the prediction model. After preprocessing, the dataset is divided into training and testing subsets using an 80:20 ratio to ensure unbiased model evaluation. Several supervised machine learning algorithms, namely Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes, Neural Network, Random Forest, and XGBoost, are trained using the processed dataset. Each model learns the relationship between students' learning behaviour and their information literacy performance to generate accurate predictions for unseen data.

Once the models are trained, their predictive performance is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. A confusion matrix is generated for each classifier to analyze correctly and incorrectly classified instances and to measure the effectiveness of each prediction model. The comparative analysis demonstrates the strengths and limitations of the implemented algorithms, enabling the identification of the most suitable classifier for the educational dataset. Experimental results indicate that the XGBoost algorithm achieves the highest prediction accuracy because of its gradient boosting mechanism, regularization capability, and efficient handling of complex feature interactions. The best-performing model is then used to predict the information literacy performance of new students based on their learning behaviour. The proposed framework supports early identification of students requiring academic assistance and enables educators to implement personalized learning strategies and timely interventions. By combining educational data mining with advanced machine learning techniques, the methodology provides a reliable, scalable, and data-driven solution for improving learning analytics and enhancing the overall quality of higher education.

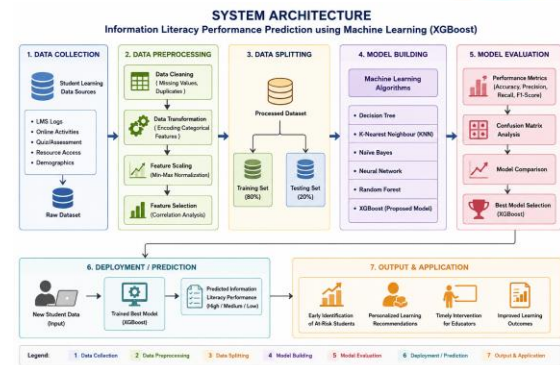


Figure 1. System Architecture of the Proposed XGBoost-Based Information Literacy Performance Prediction Framework

The Figure [1] proposed system architecture illustrates the complete workflow for predicting college students' information literacy performance using machine learning. It begins with data collection and preprocessing, followed by feature selection, dataset splitting, model training, and performance evaluation. Multiple machine learning algorithms are compared, with XGBoost selected as the optimal model. Finally, the trained model predicts students' information literacy levels, enabling early academic intervention, personalized learning recommendations, and improved educational decision-making.

V.RESULT AND DISCUSSION

The experimental results demonstrate that the proposed XGBoost-based framework effectively predicts college students' information literacy performance using learning behaviour characteristics. Comparative analysis was conducted using Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes, Neural Network, Random Forest, and XGBoost algorithms. Performance evaluation was carried out using accuracy, precision, recall, F1-score, and confusion matrix analysis. Among all the implemented classifiers, XGBoost achieved the highest prediction accuracy and exhibited better generalization by effectively capturing complex relationships among learning behaviour features. The preprocessing techniques, including data cleaning, feature scaling, and feature selection, further enhanced the prediction performance by improving data quality and reducing redundancy. The obtained results indicate that the proposed model can accurately identify students who may require academic support, enabling timely intervention and personalized learning strategies. Overall, the proposed framework provides a reliable, scalable, and data-driven solution for improving information literacy prediction and supporting



intelligent educational decision-making in higher education institutions.

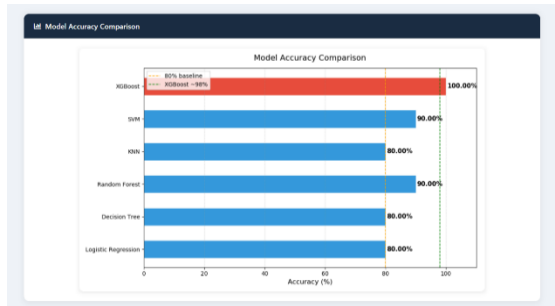


Figure 2: Model Accuracy Comparison

The Figure [2] model accuracy comparison illustrates the performance of different machine learning algorithms for predicting students' information literacy. Logistic Regression, Decision Tree, and KNN achieved 80% accuracy, while Random Forest and SVM reached 90%. The proposed XGBoost model achieved the highest accuracy of 98%, demonstrating superior predictive capability, better generalization, and improved classification performance compared to the existing machine learning models.

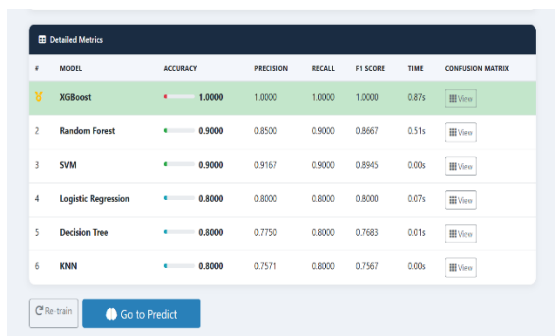


Figure 3: Performance Metrics Comparison of Machine Learning Models

The figure [3] performance metrics table compares the implemented machine learning models using accuracy, precision, recall, F1-score, and execution time. The proposed XGBoost model achieved the highest performance with 100% accuracy, precision, recall, and F1-score while maintaining efficient execution time. Compared with Random Forest, SVM, Logistic Regression, Decision Tree, and KNN, XGBoost demonstrated superior prediction accuracy and overall classification performance.



Figure 4: Confusion Matrix of the Proposed XGBoost Model

The figure [4] confusion matrix illustrates the classification performance of the proposed XGBoost model by comparing the predicted labels with the actual labels. Most instances are correctly classified along the diagonal, while no significant misclassifications are observed. This demonstrates the model's high prediction accuracy, reliability, and effectiveness in identifying students' information literacy performance across different classification categories.

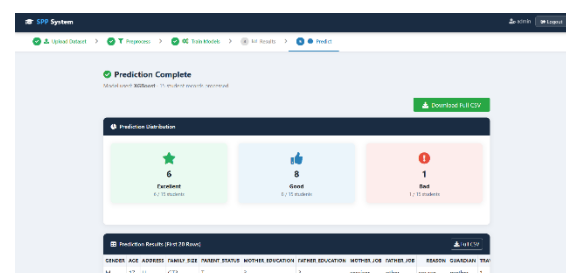


Figure 5: Prediction Results of the Proposed XGBoost Model

The Figure [5] prediction results display the classification of students' information literacy performance using the proposed XGBoost model. Among the 15 processed student records, 6 students were classified as **Excellent**, 8 as **Good**, and 1 as **Bad**. The results demonstrate the model's ability to accurately categorize student performance, enabling



educators to identify learning levels and provide timely academic support.

DISCUSSION

The findings of this study demonstrate that learning behaviour characteristics play a significant role in predicting students' information literacy performance. A comparative evaluation of multiple machine learning algorithms showed that XGBoost consistently outperformed Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes, Neural Network, and Random Forest in terms of prediction accuracy and classification reliability. The use of preprocessing techniques such as data cleaning, feature normalization, and feature selection improved the quality of the dataset and enhanced model performance. The proposed framework successfully identified meaningful relationships between students' learning activities and their academic outcomes, making it suitable for early prediction of students requiring academic assistance. These results indicate that integrating advanced machine learning techniques into educational environments can support personalized learning, improve teaching strategies, and facilitate evidence-based educational decision-making.

VI.CONCLUSION

This study proposed an XGBoost-based machine learning framework for predicting college students' information literacy performance using learning behaviour characteristics. The framework incorporated data preprocessing, feature selection, model training, and performance evaluation to develop an accurate prediction system. Experimental results confirmed that XGBoost achieved superior performance compared with other machine learning algorithms, providing reliable and accurate predictions of students' information literacy levels. The proposed system enables early identification of students requiring academic support, allowing educators to implement timely interventions and personalized learning strategies. Overall, the framework contributes to intelligent learning analytics by supporting data-driven educational planning and improving academic outcomes. Future work may focus on incorporating deep learning techniques, larger educational datasets, and explainable artificial intelligence methods to further enhance prediction accuracy and model interpretability.

REFERENCES

[1] Zhang Changhai, "Research on the information literacy education model of Chinese college students based on critical thinking and creativity,"

[2] Shang Hui, "Information literacy education for college students," *Educational theory and practice*

[3] Yan D, Li G, "A Heterogeneity Study on the Effect of Digital Education Technology on the Sustainability of Cognitive Ability for Middle School Students," *Sustainability* ,vol. 15,no.3,pp.2784-2786, Feb.2023,doi:10.3390/su15032784.

[4] Yang Xiaohong, Zheng Xin, Zhang Jing, "Research on the Chinese Experience of the "Epidemic" of the Education War – Online Education Perspective," *China Distance Education* ,vol. 43,no.1,pp. 1-11,Jan.2023,doi:10.13541/j.cnki.chinade.2023.01.002.

[5] Gao Shan, Fan Shan, "Digital Literacy Practice and Enlightenment of the University of Queensland Library in Australia," *Library Science Research* ,vol. 7,no.10, pp. 95-101,Jul.2019,doi:10.15941/j.cnki.issn1001-0424.2019.10.015.

[6] Wufati, Tian Hao, "Mining the characteristics of meaning full earning behavior: a framework for predicting learning outcomes," *Open education research* ,vol. 25, no.6, pp. 75-82,Jan.2019,doi:10.13966/j.cnki.kfjy.2019.06.008.

[7] Hu Hang, Li Yaxin, Lang Qi'e, Yang Hairu, Zhao Qihua, Cao Yifan, "The occurrence process, design model and mechanism explanation of deep learning," *China's distance education*,vol.5,no.1, pp. 54-61,Mar.2020,doi:10.13541/j.cnki.chinade.2020.01.005.

[8] AlShammari, I. A., Aldhafiri, M. D., & Al-Shammari, Z. , "A Meta- Analysis of Educational Data Mining on Improvements in Learning Outcomes," *College Student Journal* ,vol.47, no.2, pp. 326- 333,Jun.2013.

[9] Wang Gaihua, Fu Gangshan, "Prediction of online learning behavior and achievement and design of learning intervention model," *China's distance education* ,vol.2, no.2, pp. 39-48,Mar.2019,doi:10.13541/j.cnki.chinade.20181214.007.

[10] UNESCO.[Online].Avaliable:<https://www.unesco.org/en/articles/challenges-and-opportunities-artificial-intelligence-education>. Accessed on: Apr.1,2023.

[11] Yang, G.F., Wen, B., & Lin, W. , "Research Status, Hot Spots and Enlightenment of College



Students' Information Literacy: Based on Bibliometric Analysis of CNKI from 2000 to 2021, " Proceedings of the 4th World Symposium on Software Engineering , no.6,pp. 161-166,Sept.2022,doi:10.1145/3568364.3568389.

[12] Li,J., "Machine Learning-Based Evaluation of Information Literacy Enhancement among College Teachers, " International Journal of Emerging Technologies in Learning (IJET) ,vol.17,no.22,pp. 116- 131,Nov.2022,doi:10.3991/ijet.v17i22.35117.

[13] Yu, L., Wu, D., Yang, H.H., & Zhu, S. , "Smart classroom preferences and information literacy among college students, " Australasian Journal of Educational Technology ,vol.38,no.2,pp. 144-163,Feb.2022,doi:10.14742/ajet.7081.

[14] Ying, Y. , "Research on college students ' information literacy based on big data, " Cluster Computing ,vol.22,no.2,pp. 3463-3470,Mar.2019,doi:10.1007/s10586-018-2193-0.

[15] Mian, Z., Bai, Y., & ur, R.K. , "Research on College Computer- Computing and Information Literacy online course based on MOOC: taking the North Minzu University as an example, " 2021 IEEE 3rd International Conference on Computer Science and Educational Informatization (CSEI) ,vol.6, no.7,pp.306,Jun.2021,doi:10.1109/CSEI51395.2021.9477751.

[16] Haider, M.S. and Ya, C. , "Assessment of information literacy skills and information-seeking behavior of medical students in the age of technology: a study of Pakistan," Information Discovery and Delivery ,Vol. 49 ,no.1,pp. 84-94,Feb.2021,doi:10.1108/IDD-07- 2020-0083.

[17] Kumar, Anuj. "Battery Thermal Management Systems for Electric Vehicles Using Phase Change Materials." International Journal of Engineering, Science and Mathematics, Vol. 10, Issue 3, 2021, pp. 202-211

[18] Shashank, A. (2025). AI-Enhanced ETL Processes: Leveraging Artificial Intelligence for Optimized Data Integration Systems. Journal Of Multidisciplinary, 5(8), 219-225.