Evaluating The Learning Patterns And Analyzing Efficiency In Students Using Machine Learning Algorithms

[1] MD Altaf, [2]Peddireddy Shiva Reddy,[3] Polasi Praneeth,[4] G Thirupathi

[1,2,3] Graduate Student-CSE, Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad
[4] Asst.Professor, Department of CSE, Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad

***Abstract*—**This project explores the critical role that information literacy plays in the learning outcomes and behaviors of college students. Using a range of supervised classification algorithms, predictive models were created by looking at different learning behaviors and emphasizing information literacy. This study expands the analysis by including new methodologies, building upon the original paper's successful use of Decision Trees, KNN, Naive Bayes, Neural Networks, and Random Forest, which produced an outstanding 92.50% accuracy. The accuracy shot up to 100% when XGBoost and a Voting Classifier were added to the ensemble approach. This improvement represents the possibility for sophisticated techniques to improve the models' predictive power, providing insightful information on customized interventions to maximize information literacy instruction. The results highlight how important it is to comprehend and take use of a variety of learning behaviors in order to develop creative people who can learn new things their entire lives and adjust to changing social demands. This study adds to the growing body of knowledge about information literacy's critical role in postsecondary education and its implications for developing flexible, independent learners.

Index Terms: Machine learning, information literacy, learning behavior characteristics, learning effect, innovative talents.

INTRODUCTION

Overview

The rapid development of information technology has significantly impacted various sectors, making it crucial for college students to acquire competencies such as creativity, critical thinking, and information literacy. Information literacy is essential for fostering creative genius and ensuring the long-term development of future human resources. Educational institutions worldwide prioritize information literacy instruction due to its importance.

In recent years, online and hybrid teaching modalities and advancements in artificial intelligence technologies have led to the rise of specialized information literacy courses in colleges. However, there are still obstacles in college-level training, such as effective learning outcome prediction. Machine learning techniques can be used to optimize the learning process by acquiring insights into learners' progress and customizing interventions accordingly.

Learning prediction uses variables like learning achievement, goals, and abilities to predict learning experiences and results. Techniques such as regression analysis, neural networks, and Bayesian approaches are used to predict students' learning outcomes. The integration of machine learning and educational data mining technologies has emerged as a promising path toward developing data-driven prediction models.

UNESCO's 2019 report on Artificial Intelligence in Education highlights the potential of integrating artificial intelligence and education for advancing quality and equity in educational institutions. Teachers can use data-driven insights to improve learning outcomes and provide individualized learning experiences for students using educational data mining and machine learning. This research investigates the relationship between learning behavior analysis, predictive modeling, and information literacy in higher education contexts, aiming to fill gaps and tackle issues in learning prediction methodologies and information literacy education.

Furthermore, as higher education continues to change in the digital age, information literacy's importance in preparing students for success in a variety of disciplines is becoming more widely acknowledged.

Objective

This research attempts to investigate the relationship between learning behavior analysis, predictive modeling, and information literacy in the context of higher education in light of these advancements. This study aims to fill in the gaps and tackle the issues in the field of learning prediction methodologies and information literacy education by looking at the current state of the art. It also provides insights into how machine learning techniques might be used to improve information literacy instruction and maximize learning outcomes for college students.

LITERATURE SURVEY

Related Work

The construction of an information literacy education model for Chinese college students is the focus of Z. Chinghai's work [1], which integrates creativity and critical thinking. This model emphasizes how crucial it is to develop students' capacity for critical analysis and innovative use of information in order to raise their level of information literacy as a whole. In a similar vein, S. Hui [2] addresses information literacy teaching tactics designed for university students, emphasizing the necessity of a comprehensive strategy that includes both theoretical understanding and practical competence.

Using information from literature indexed in the CNKI database from 2000 to 2021, G. Yang, B. Wen, and W. Lin [3] propose a bibliometric analysis of research trends and hotspots in college students' information literacy. Their research highlights regions that are ready for more examination by identifying major themes, hot subjects, and research trajectories in the discipline.

L. Yu, D. Wu, H. H. Yang, and S. Zhu [4] investigate college students' choices for smart classrooms and information literacy. They investigate the relationship between students' information literacy skills and their preferences for technology-enhanced learning settings through empirical research, providing insightful data for instructional design and pedagogical practice.

Y. Ying [5] uses big data analytics to examine information literacy among college students. The study finds patterns, trends, and connections pertaining to students' information-seeking activities and information processing abilities by examining large-scale datasets. The aforementioned study enhances our comprehension of the complex characteristics of information literacy and its consequences for pedagogical approaches.

The promotion strategies and influencing factors related to information literacy among college students are examined by X. Ouyang, Y. Xiao, and J. Zhong [6]. They identify important factors influencing students' information literacy levels through a qualitative investigation, and they suggest focused interventions to improve information literacy instruction in higher education settings.

The information technology literacy of newly enrolled female college students in Japan is evaluated by T. Nishikawa and G. Izuta [7]. Their study examines potential factors impacting students' technological competencies as well as their competency with a variety of information technologies. The results support initiatives to close the digital divide and increase college students' digital literacy.

Based on multifarious data, Y. Sun, Z. Tan, Z. Li, and S. Long [8] use machine learning approaches to forecast and analyze the performance of college students. Through the utilization of many data sources, such as extracurricular activities, academic records, and demographic data, the research creates predictive models that are able to anticipate the academic outcomes of students. This study highlights how data-driven strategies can improve student support programs and educational decision-making.

The research review concludes by highlighting the multifaceted character of information literacy instruction for college students, which includes predictive modeling of learning outcomes, technological competence, critical thinking, and creativity. This survey provides an extensive summary of current research trends, opportunities, and difficulties in the subject by combining insights from various studies. The future course of information literacy education is expected to be shaped by multidisciplinary collaboration, novel pedagogical approaches, and technology breakthroughs, which will enable college students to prosper in an increasingly complicated and linked world.

METHODOLOGY

A) Proposed Work

Utilizing pre-analyzed data on learning behavior and its relationships to learning outcomes, the proposed study seeks to create predictive models by applying techniques from Decision Tree[9], K-Nearest Neighbor (KNN)[10], Naive Bayes[11], Neural Network (NN), and Random Forest. The goal of this research is to shed light on the complex relationship that exists between academic achievement and learning behavior patterns among students.

Preprocessing the data is part of the methodology to guarantee its accuracy and applicability. The predictive usefulness of characteristics such as levels of engagement, study habits, and involvement in educational activities will be closely examined. To assess each model's performance, the data will then be divided into training and testing sets.

The models' efficacy will be assessed by the application of criteria including F1 score, recall, accuracy, and precision. Furthermore, the models' interpretability will be given top priority in order to find practical insights for focused interventions. In the end, the goal of this project is to provide a systematic framework for using machine learning to forecast learning outcomes based on students' behavior in the classroom. This will improve learning outcomes and advance personalized learning strategies in higher education.

B) System Architecture

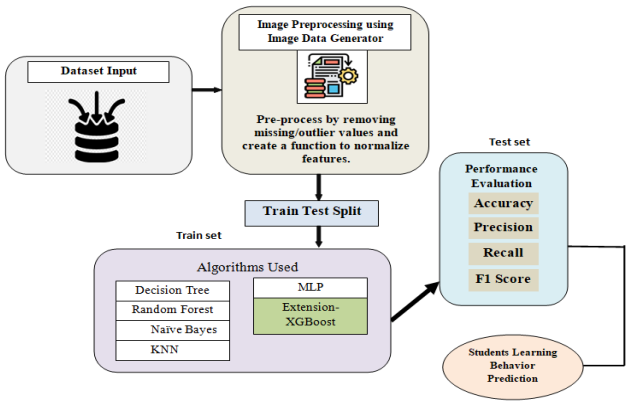

Fig 1 Proposed Architecture

The system design includes a number of interrelated parts that work together to make it easier to forecast how children will learn. First, the architecture takes in a dataset that includes pertinent data on how students learn, including things like their study habits, engagement levels, and involvement in class activities. To improve its quality and get it ready for analysis, the dataset is next subjected to image processing using Image Data Generator methods. To guarantee the reliable assessment of prediction models, the dataset is split into training and testing sets using a Train-Test-Split method after preprocessing.

A number of machine learning algorithms, such as Decision Tree[9], Naive Bayes[11], K-Nearest Neighbor (KNN)[10], Random Forest[12], Multi-Layer Perceptron (MLP), and XGBoost, are used at the heart of the design. By analyzing the preprocessed data, these algorithms are able to precisely forecast the learning behavior of pupils. Metrics for performance evaluation, including Accuracy, Precision, Recall, and F1 Score, are used to evaluate how well each algorithm captures the subtleties of students' learning styles.

In the end, the system architecture uses performance evaluation criteria and cutting-edge machine learning approaches to forecast students' learning behavior. Through the smooth integration of these elements, the architecture offers a thorough framework for comprehending and forecasting students' learning habits, enabling focused interventions and improving academic results.

C) Data Set

The Student Learning Behavior dataset is made up of an extensive range of characteristics that represent many facets of students' performance and involvement in the classroom. It contains data on the study habits, attendance histories, extracurricular activity involvement, test results, and demographics of the pupils. The dataset might also include information about how students use educational resources like online learning environments and library resources. The dataset allows for in-depth investigation and analysis of the variables impacting students' learning behaviors and academic outcomes because to its wealth of information. For scholars and educators looking to deepen understanding and encourage students' academic journeys, it is an invaluable resource.

	IPC1	IPC2	IPC3	IAC1	IAC2	LLC1	LLC2	ISK1	ISK2	ISK3	ISK4	IAS1	IT1	IT2	IT3	IB1	IE1	IE2	ILR1	label
0	69	63	78	87	94	94	87	84	61	4	4	7.9	A	1.0	0.0	0.0	0.0	0.0	no	excellent
1	78	62	73	60	71	70	73	84	91	7	2	5.4	B	2.0	0.0	0.0	0.0	0.0	no	medium
2	71	86	91	87	61	81	72	72	94	1	1	5.2	B	7.0	0.0	0.0	0.0	0.0	no	excellent
3	76	87	60	84	89	73	62	88	69	1	2	8.5	C	10.0	0.0	0.0	0.0	0.0	yes	excellent
4	92	62	90	67	71	89	73	71	73	5	6	8.8	C	6.0	0.0	0.0	0.0	0.0	no	excellent
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1008	88	85	68	84	88	66	86	76	82	2	2	7.6	A	1.0	0.0	0.0	0.0	0.0	no	excellent
1009	76	63	92	74	76	81	76	87	81	8	7	7.4	C	7.0	0.0	0.0	0.0	0.0	yes	excellent
1010	74	94	94	82	64	92	84	67	80	4	6	7.7	C	5.0	0.0	0.0	0.0	0.0	no	poor
1011	60	84	84	70	80	78	64	83	60	8	6	7.6	D	8.0	0.0	0.0	0.0	0.0	yes	excellent
1012	91	61	83	80	88	62	88	76	86	9	1	7.4	D	5.0	0.0	0.0	0.0	0.0	no	excellent

Fig 2 Dataset

D) Data Processing

Data Loading with Pandas Dataframe: The process of processing data begins with loading the dataset into a pandas data frame, which is a vital tool that is well-known for its effectiveness in managing structured data. By utilizing the features of the dataframe, the dataset's contents are arranged into a tabular structure for easy access and manipulation during the stages of processing that follow.

Column Dropping: In an effort to refine the data, unnecessary or duplicate columns are carefully found and removed from the data frame. Column dropping is a selected method that helps to simplify the dataset by removing extraneous information and simplifying the computation. Column dropping simplifies the dataset by keeping only the most pertinent features, guaranteeing that next studies concentrate on the most important variables.

Normalization of Training Data: The training data is normalized to promote fair comparisons and lessen the impact of different feature scales. The numerical feature values are standardized by this transformative process, which usually rescales them to a common range like [0, 1] or [-1, 1]. Normalization encourages fairness in model training and evaluation by standardizing feature magnitudes, making it easier to make accurate and dependable predictions across a variety of datasets.

E) Visualization

Data visualization is made into an art form by combining the potent capabilities of the Seaborn and Matplotlib tools. Built on top of Matplotlib, Seaborn provides a simple-to-use interface for writing little to no code while producing visually striking charts. A broad range of high-level functions are available in Seaborn for dataset exploration and comprehension, ranging from basic histograms and scatter plots to complex heatmaps and violin plots. Matplotlib, on the other hand, provides more precise control over customizing plots, enabling the production of visualizations suitable for publishing. Seaborn and Matplotlib work together to enable analysts and data scientists to effectively communicate insights through eye-catching and educational images.

F) Label Encoding and Feature Selections

Label encoding converts categorical variables into a numerical format that makes them easier to understand by machine learning algorithms. Through this method, every category inside a feature is given a distinct numerical label.

Finding and keeping characteristics that have significant linear correlations with the target variable is a key component of feature selection based on high correlation values. Highly associated features are found and chosen for the prediction model by calculating correlation coefficients between the features and the target variable. By concentrating on the most significant traits and eliminating superfluous or unnecessary ones, this selective method maximizes interpretability and forecast accuracy while improving model efficiency.

G) Training and Testing

In order to ensure that the performance of the machine learning model can be precisely evaluated on unknown data, it is imperative that the data be split into training and testing subsets. The supplied dataset is divided into two separate subsets for this process: the training set and the testing set. The model is trained on the patterns and relationships found in the data using the training set, which usually consists of a higher percentage of the data. The testing set, on the other hand, is a smaller subset of the data that is used to assess the performance of the trained model. The testing set acts as an impartial gauge of the model's capacity for generalization by excluding some of the data during training, giving information about how well the model performs when applied to fresh, untested data.

To ensure that the training and testing subsets of the data are representative of the entire dataset, the splitting of the data into these subsets is usually done at random. Allocating a specific percentage of the data, say 70–80%, to the training set and the remaining amount to the testing set is a common approach. This guarantees a balance between maintaining a suitable evaluation dataset and offering enough data for model training.

Furthermore, methods like cross-validation could be used to evaluate model performance even more and lessen possible biases brought about by the data splitting procedure. Generally, robust model building and evaluation in machine learning applications depend on the meticulous division of data into training and testing subsets.

H) Algorithms Used

Basic machine learning algorithms like Random Forest, Decision Tree, Naive Bayes, K-Nearest Neighbors, and Multi-Layer Perceptron have a wide range of uses in many fields.

Random Forest: Random Forest is an ensemble learning technique that builds many decision trees during training and produces the mean prediction (regression) or mode of the classes (classification) for each individual tree. It is resistant to overfitting and performs well with big, highly dimensional datasets.

Decision Tree: Decision Tree is a straightforward yet effective technique that creates a tree-like structure by iteratively dividing the dataset into subsets according to the most important attribute. [9] Because of its great interpretability and intuitiveness, it can be used to clarify the decision-making process and comprehend the significance of features.

Naive Bayes: With an assumption of predictor independence, Naive Bayes is a probabilistic classifier based on the Bayes theorem. It frequently works effectively in text categorization and other areas, especially when working with high-dimensional data, despite [11]its simplicity and the "naive" assumption.

K-Nearest Neighbors (KNN): KNN is an instance-based, non-parametric learning method that groups new data points in the feature space according to how close they are to the majority class of their K nearest neighbors [10]. It is simple to use and adaptable, especially with smaller datasets.

Multi-Layer Perceptron: An artificial neural network called an MLP is made up of several layers of nodes, or neurons, coupled to one another at each layer. MLPs are frequently employed for tasks like pattern recognition, regression, and classification because they can understand intricate correlations in data.

These methods, each with unique strengths and limitations based on the particular issue domain and dataset characteristics, are fundamental components of a data scientist's toolkit and the basis of many machine learning applications.

EXPERIMENTAL RESULTS

Precision: Precision measures the percentage of correctly categorized samples or instances among the positive samples. Consequently, the following is the formula to determine the precision:
True positives/(True positives + False positives) = TP/(TP + FP) is the formula for precision.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

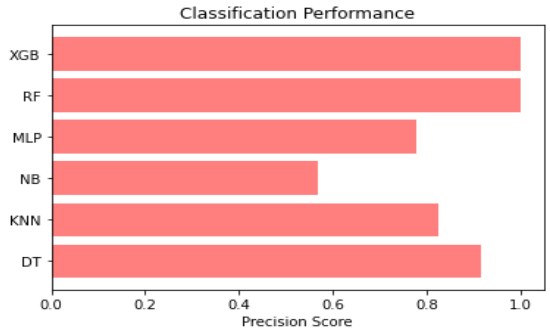

Fig 3 Precision Comparison Graph

Recall: In machine learning, recall is a metric that assesses a model's capacity to locate all pertinent instances of a given class. It is a measure of how well a model captures examples of a particular class: the ratio of correctly predicted positive observations to the total number of real positives.

$$Recall = \frac{TP}{TP + FN}$$

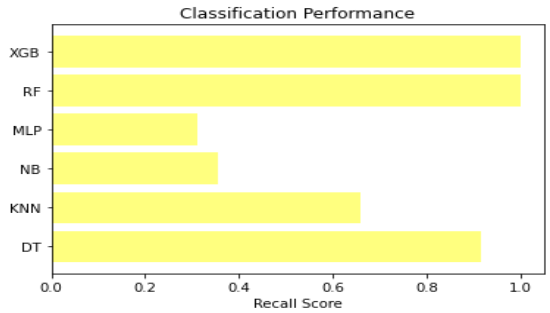

Fig 4 Recall Comparison Graph

F1-Score: An evaluation statistic for machine learning called the F1 score quantifies the accuracy of a model. It integrates a model's precision and recall ratings. The number of times a model correctly predicted throughout the whole dataset is calculated by the accuracy metric.

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

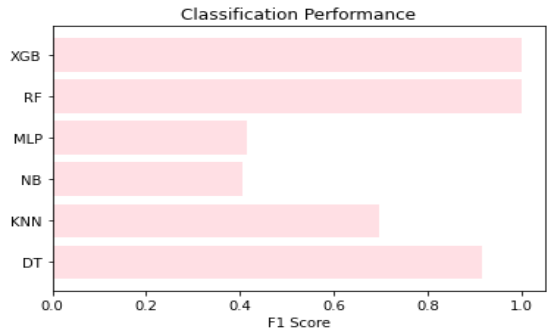

Fig 5 F1 Score Comparison Graph

Accuracy: A test's accuracy is determined by how well it can distinguish between patient and healthy cases. We should compute the percentage of true positive and true negative in each analyzed case to assess the accuracy of a test. This can be expressed mathematically as follows:

Accuracy = TP + TN / (TP + TN + FP + FN)

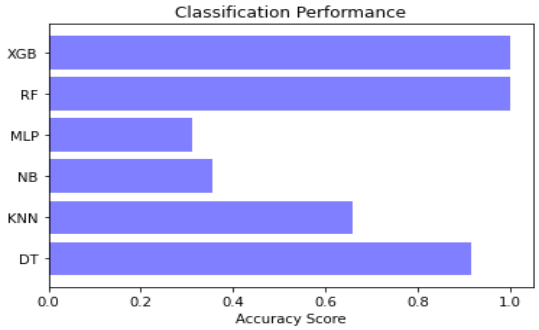

Fig 6 Accuracy Comparison Graph

	ML Model	Accuracy	Precision	f1_score	Recall
0	DT	0.916	0.918	0.916	0.916
1	KNN	0.658	0.825	0.696	0.658
2	NB	0.355	0.569	0.404	0.355
3	MLP	0.310	0.778	0.416	0.310
4	RF	1.000	1.000	1.000	1.000
5	Extension-XGB	1.000	1.000	1.000	1.000

Fig 7 Performance Evaluation Table

Fig 13 Upload Input Data

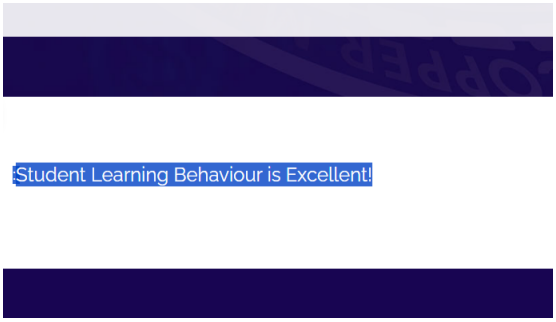

Fig 14 Final Outcome

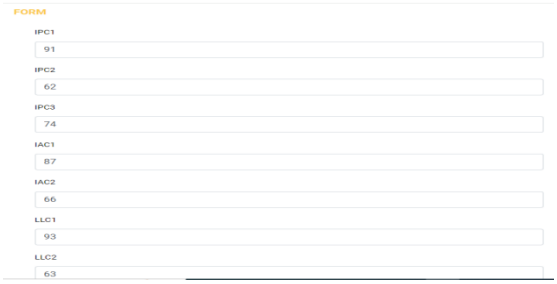

Fig 15 Upload Input Data

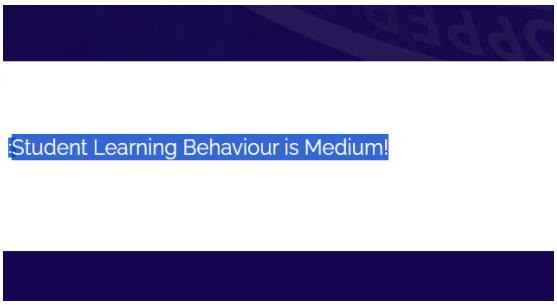

Fig 16 Predicted Result

CONCLUSION

In conclusion, today's information-rich environment, information literacy is essential for success. It goes beyond academic performance to become a lifetime learning tool and a means of navigating the intricacies of contemporary life. Teachers can customize their teaching approaches to meet the needs of each individual student and create a more inclusive and productive learning environment by understanding the complex interactions that exist between student learning behaviors and outcomes. Using predictive models like Random Forest,

Decision Tree, KNN, Naive Bayes, Neural Network, and Random Forest—plus the potent Extension-XGBoost—improves teachers' capacity to recognize and respond to differences in students' information literacy competency levels. By enabling educators and administrators to convert these insights into workable methods, the practical integration of XGBoost within Flask promotes informed decision-making and leads to observable gains in educational results.

FUTURE SCOPE

Going forward, there is a great deal of promise in combining cutting-edge machine learning methods with teaching approaches. Predictive models have a great deal of room for improvement as technology develops in order to better comprehend and assist students' learning journeys. Furthermore, current research and development initiatives in the field of educational data analytics present chances to investigate novel approaches and broaden the application of predictive modeling to tackle newly developing issues in education. Through adoption of these developments and promoting cooperation among scholars, instructors, and tech creators, we can keep utilizing data-driven insights to influence the course of education and enable students all over the world.

REFERENCES

[1] Z. Changhai, ‘‘Research on the information literacy education model of Chinese college students based on critical thinking and creativity,’’ J. China Library, vol. 4, no. 15, pp. 15–16, Aug. 2016, doi: 10.13530/j.cnki.jlis.164008.

[2] S. Hui, ‘‘Information literacy education for college students,’’ Educ. Theory Pract., vol. 30, no. 10, pp. 38–39, Oct. 2008.

[3] G. Yang, B. Wen, and W. Lin, ‘‘Research status, hot spots and enlightenment of college students’ information literacy: Based on bibliometric analysis of CNKI from 2000 to 2021,’’ in Proc. 4th World Symp. Softw. Eng., Sep. 2022, pp. 161–166, doi: 10.1145/3568364.3568389.

[4] L. Yu, D. Wu, H. H. Yang, and S. Zhu, ‘‘Smart classroom preferences and information literacy among college students,’’ Australas. J. Educ. Technol., vol. 38, no. 2, pp. 144–163, Feb. 2022, doi: 10.14742/ajet.7081.

[5] Y. Ying, ‘‘Research on college students’ information literacy based on big data,’’ Cluster Comput., vol. 22, no. S2, pp. 3463–3470, Mar. 2019, doi: 10.1007/s10586-018-2193-0.

[6] X. Ouyang, Y. Xiao, and J. Zhong, ‘‘Research on the influencing factors and the promotion measures of college students’ information literacy,’’ in Proc. 8th Int. Conf. Inf. Technol. Med. Educ. (ITME), Dec. 2016, vol. 12, no. 12, pp. 728–732, doi: 10.1109/ITME.2016.0169.

[7] T. Nishikawa and G. Izuta, ‘‘The information technology literacy level of newly enrolled female college students in Japan,’’ Hum. Social Sci. Rev., vol. 7, no. 1, pp. 1–10, Mar. 2019, doi: 10.18510/hssr.2019.711.

[8] Y. Sun, Z. Tan, Z. Li, and S. Long, ‘‘Predicting and analyzing college students’ performance based on multifaceted data using machine learning,’’ in Proc. 4th Int. Conf. Adv. Comput. Technol., Inf. Sci. Commun. (CTISC), Apr. 2022, pp. 1–6, doi: 10.1109/CTISC54888.2022.9849815.

[9] A. Weipeng and S. Jiaze, ‘‘Improvement and analysis of decision tree C4.5 algorithm,’’ Comput. Eng. Appl., vol. 55, no. 12, pp. 169–173, Jun. 2019, doi: 10.3778/j.issn.1002-8331.1805-0482.

[10] Z. Yu, H. Chen, J. Liu, J. You, H. Leung, and G. Han, ‘‘Hybrid k-nearest neighbor classifier,’’ IEEE Trans. Cybern., vol. 46, no. 6, pp. 1263–1275, Jun. 2015, doi: 10.1109/TCYB.2015.2443857.

[11] M. S. Roobini, K. Babu, J. Joseph, and G. Ieshwarya, ‘‘Predicting stock price using data science technique,’’ in Proc. 2nd Int. Conf. Artif. Intell. Smart Energy (ICAIS), no. 2, Feb. 2022, pp. 1013–1020, doi: 10.1109/ICAIS53314.2022.9742772.