ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 3, 2024

ANALYZING AND PREDICTING LEARNING & LITERACY OF COLLEGE STUDENTS USING MACHINE LEARNING

[1]Manasa kotha, [2]Dr.Manoj Kumar Mahto, [3]Dr.G.Raja Vikram

M.Tech,**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**,Vignan Institute of Technology and Science,Telangana-508284.

Dr.Manoj Kumar Mahto , Professor , **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**,Vignan Institute of Technology and Science,Telangana-508284.

Dr.G.Raja Vikram ,Professor, **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**,Vignan Institute of Technology and Science,Telangana-508284.

ABSTRACT

This project focuses on leveraging machine learning techniques to analyze and predict learning outcomes and literacy levels among college students. By harnessing predictive analytics, the project aims to gain insights into factors influencing academic performance and literacy skills, facilitating informed decision-making in education. Key aspects include data collection and preprocessing, model training using decision trees, AdaBoost, XGBoost, and gradient boosting algorithms, and the development of predictive models to forecast student learning outcomes. The project emphasizes ethical considerations, ensuring data privacy and fairness in deploying predictive analytics in educational settings. Keywords: machine learning, predictive analytics, learning outcomes, literacy levels, college students, decision trees, AdaBoost, XGBoost, gradient boosting, data privacy, ethical considerations.

Keywords: *Machine learning, predictive analytics, learning outcomes, literacy levels, college students, decision trees, AdaBoost, XGBoost, gradient boosting, data preprocessing, model training, data privacy, ethical considerations.*

I. INTRODUCTION

In today's rapidly evolving knowledge-based society, the ability to navigate and critically evaluate information is essential, particularly within the context of higher education. Students must develop robust information literacy skills to succeed both academically and professionally. Information literacy involves the capacity to identify, locate, evaluate, and effectively use information from a variety of sources, including digital platforms, libraries, and academic databases. For educators, understanding the nuances of teaching information literacy is crucial to tailoring their approaches effectively and ensuring that students acquire these essential skills.

The motivation behind this research is rooted in the recognition of information literacy as a cornerstone for student success in both academic and professional realms. As societal demands and technological advancements continue to progress, the need for individuals who can proficiently navigate the vast expanse of information becomes increasingly critical. This study aims to enhance the effectiveness of information literacy teaching by investigating college students' learning behaviors and exploring predictive models for learning outcomes based on information literacy characteristics. By doing so, the research contributes to the broader dialogue on lifelong learning and seeks to improve educational practices.

The central problem addressed by this research is optimizing information literacy teaching mechanisms to better align with students' learning behaviors and academic performance. While current approaches, such as Decision Tree and Random Forest algorithms, offer valuable insights into the correlation between information thinking characteristics and learning effects, there remains significant potential for enhancement. This research proposes to bridge this gap by implementing a more advanced predictive modeling technique, specifically the XGBoost algorithm, to improve the accuracy and efficiency of learning effect predictions.

The problem statement thus focuses on refining information literacy teaching methods to better predict and understand the impact of learning behaviors on academic outcomes. It involves integrating advanced algorithms like XGBoost into existing frameworks to enhance predictive modeling and provide educators with deeper insights into students' development of information literacy skills.

II.EXISTING SYSTEM

The current state of information literacy teaching relies on Decision Tree and Random Forest

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 3, 2024

algorithms for the analysis of learning behaviour characteristics. While these algorithms offer valuable insights, there is a need for further exploration and improvement. The existing system provides a foundation for understanding the correlation between information thinking characteristics and learning outcomes, yet the potential for increased accuracy and efficiency remains untapped.

Limitations of Existing system

1. Limited predictive power of Decision Tree and Random Forest algorithms.
2. Potential inefficiencies in analysis and prediction due to exclusive reliance on these algorithms.
3. Tendency to overlook subtle nuances in learning behavior characteristics.
4. Difficulty in handling complex data relationships.
5. Lack of exploration of alternative modeling approaches.
6. Potential underestimation of uncertainty in predictive models.

III.PROPOSED SYSTEM

In the proposed system, we introduce the XGBoost algorithm as a more sophisticated approach to analyze and predict learning effects. XGBoost is chosen for its enhanced capabilities in handling complex relationships within data and its robust performance in predictive modelling. This shift aims to elevate the accuracy and efficiency of the learning effect prediction model, providing a more nuanced understanding of the correlation between information thinking characteristics and academic outcomes. The proposed system seeks to contribute to the evolution of information literacy teaching, refining the predictive modelling process for more precise and insightful results.

Proposed system Advantages

1. Enhanced Predictive Accuracy: XGBoost improves prediction accuracy.
2. Efficient Handling of Complex Relationships: XGBoost handles complex data relationships effectively.
3. Improved Efficiency: XGBoost offers faster analysis and prediction.

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 3, 2024

4. Nuanced Understanding: Provides deeper insight into the correlation between learning behaviors and academic outcomes.

5. Contribution to Teaching Evolution: Demonstrates advancement in educational practices.

IV.LITERATURE REVIEW

1. Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method

Published by Mingjie Tan and Peiji Shao from the University of Electronic Science and Technology of China and Sichuan Open University in 2015, this study addresses the issue of high dropout rates in e-learning programs by employing various machine learning models. The research utilized a dataset comprising 62,375 students and focused on incorporating personal characteristics and academic performance as key input attributes. The study applied three machine learning models—Artificial Neural Network (ANN), Decision Tree (DT), and Bayesian Networks (BNs)—to predict student dropout rates. The findings demonstrated that all three models were effective, with the Decision Tree model showing superior performance in terms of accuracy, precision, recall, and F-measure metrics. While the study provided high predictive accuracy and actionable insights for early intervention, it also highlighted challenges such as the complexity of models, potential data bias affecting generalizability, and significant resource requirements in terms of time and computing power. This research underscores the importance of proactive measures in educational settings to mitigate dropout rates and informs decision-making processes.

2. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning

Authored by Tal Yarkoni and Jacob Westfall and published in 2017, this study advocates for a shift in psychology from a predominant focus on explaining the causes of behavior to a more predictive approach utilizing machine learning principles. The researchers conducted a review of fundamental concepts and tools from machine learning, assessing their application to predictive research questions in psychology. The study contrasts the traditional emphasis on tightly controlled experiments with the potential benefits of machine learning techniques. It found that the conventional focus on causal explanations often leads to complex theories with limited predictive accuracy. In contrast, integrating machine learning principles can significantly improve predictive capabilities and enhance the understanding of behavior. While the study highlights the potential for improved predictive accuracy and a deeper understanding of psychology through machine learning, it also notes challenges such as resistance to paradigm shifts within the field, difficulties in integrating new techniques into traditional research practices, and ethical concerns regarding data privacy. This research suggests that a shift towards predictive modeling could offer valuable insights into psychological phenomena.

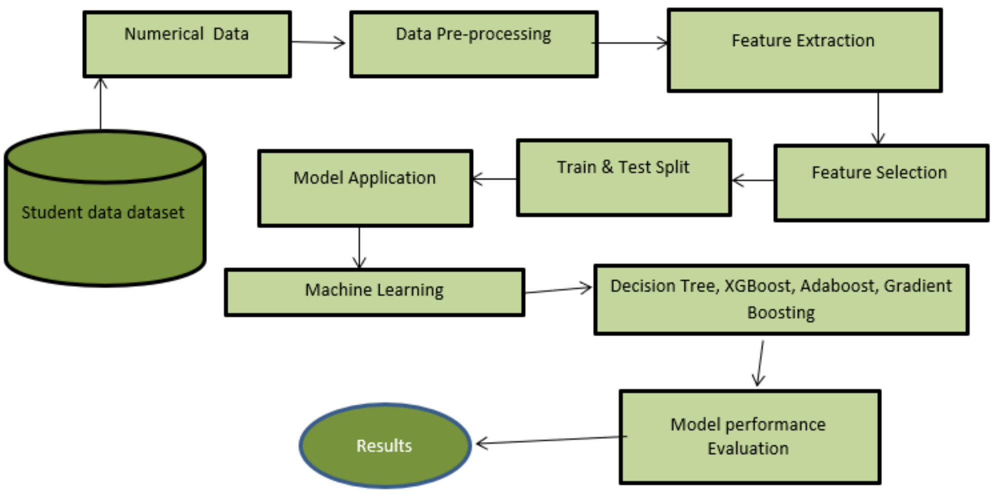


Fig1 : system architecture

V. METHODOLOGY

Problem Definition and Scope Identification:

- Define the scope of the project by identifying the specific learning and literacy metrics to be analyzed and predicted, such as academic performance, reading comprehension, writing proficiency, etc.
- Determine the target audience for the predictive analytics system, including educators, administrators, policymakers, and students themselves.
- Establish clear objectives and

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

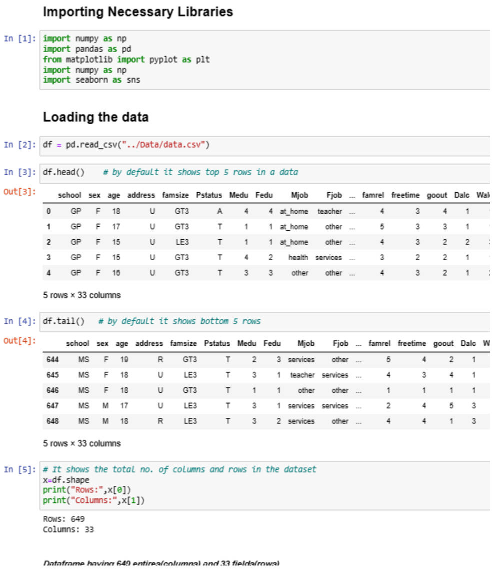
www.ijasem.org

Vol 18, Issue 3, 2024

research questions to guide the project's direction and focus.

Data Collection and Preparation:

- Gather relevant data sources, including academic records, standardized test scores, demographic information, self-reported learning behaviors, reading habits, and writing samples from college students.



Importing Necessary Libraries

```
In [1]: import numpy as np
        import pandas as pd
        from matplotlib import pyplot as plt
        import numpy as np
        import seaborn as sns
```

Loading the data

```
In [2]: df = pd.read_csv("../Data/data.csv")
In [3]: df.head()    # by default it shows top 5 rows in a data
```

Out[3]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1

5 rows × 33 columns

```
In [4]: df.tail()   # by default it shows bottom 5 rows
```

Out[4]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc
644	MS	F	19	R	GT3	T	2	3	services	other	...	5	4	2	1
645	MS	F	18	U	LE3	T	3	1	teacher	services	...	4	3	4	1
646	MS	F	18	U	GT3	T	1	1	other	other	...	1	1	1	1
647	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3
648	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3

5 rows × 33 columns

```
In [5]: # It shows the total no. of columns and rows in the dataset
        x=df.shape
        print("Rows:",x[0])
        print("Columns:",x[1])

Rows: 649
Columns: 33
```

Dataframe having 649 entires(columns) and 33 fields(rows)

- Cleanse and preprocess the collected data to handle missing values, outliers, and inconsistencies.
- Perform exploratory data analysis (EDA) to gain insights into the distribution, correlations, and patterns within the dataset.

Feature Engineering and Selection:

- Identify and extract meaningful features from the preprocessed data that are indicative of learning outcomes and literacy levels, such as GPA, SAT scores, reading/writing assessments, study habits, etc.
- Conduct feature selection techniques to prioritize relevant features and reduce dimensionality if necessary, considering factors such as predictive power, multicollinearity, and interpretability.

Model Selection and Training:

- Choose appropriate machine learning algorithms for analyzing and predicting learning and literacy metrics, such as regression, classification, or clustering models.
- Split the dataset into training and validation sets to train and evaluate the performance of the selected models.
- Experiment with different algorithms, hyperparameters, and ensemble techniques to optimize model performance and generalization ability.

Model Evaluation and Validation:

- Evaluate the trained models using appropriate evaluation

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 3, 2024

metrics tailored to the specific learning and literacy prediction tasks, such as mean squared error (MSE), accuracy, precision, recall, F1-score, etc.

- Validate the models using cross-validation techniques and test datasets to assess their robustness and generalizability across different student populations and educational contexts.

Interpretation and Analysis:

- Interpret the trained models to gain insights into the factors influencing learning outcomes and literacy levels among college students.
- Analyze the feature importance and contribution of various predictors to understand their impact on predictive performance and identify actionable insights for educational interventions.

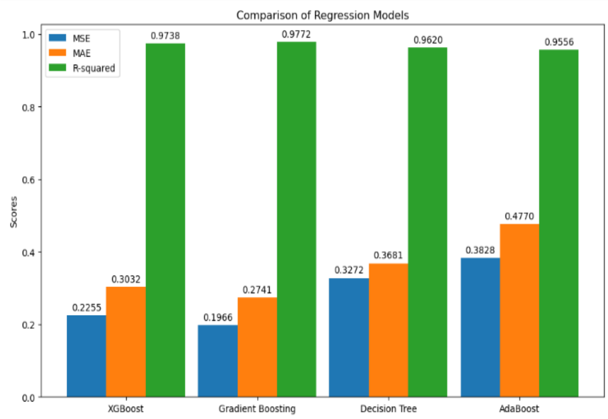
Model Deployment and Integration:

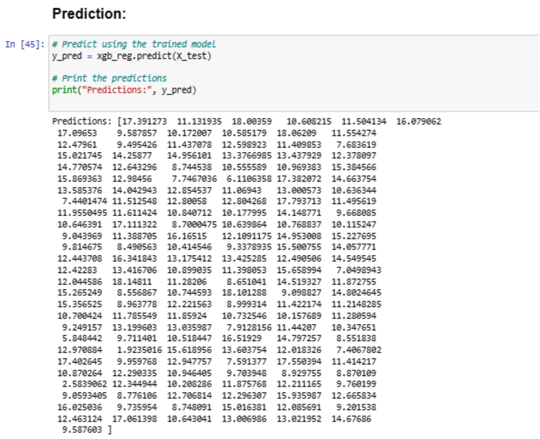
- Deploy the trained machine learning models into a production environment, either as standalone applications or integrated within existing educational systems and platforms.
- Develop user interfaces and dashboards to facilitate user interaction with the predictive analytics system, allowing stakeholders to input data, visualize analysis results, and interpret predictions.

Monitoring and Maintenance:

- Establish monitoring mechanisms to track the performance and effectiveness of the deployed models over time, incorporating feedback loops for continuous improvement.
- Implement maintenance procedures to address issues, update models with new data, and adapt to evolving educational trends and requirements.

VI.RESULTS





Prediction:

```
In [45]: # Predict using the trained model
y_pred = xgb_reg.predict(X_test)

# Print the predictions
print("Predictions:", y_pred)

Predictions: [17.391273  11.131935  18.00359   10.608215  11.504134  16.079062
 17.09653    9.587857  10.172007  10.585179  18.06209   11.554274
 12.47961    9.495426  11.437078  12.598923  11.409853   7.683619
 15.021745  14.25877   14.956101  13.3766985 13.437929  12.378097
 14.770574  12.643296   8.744538  10.555589  10.969383  15.384566
 15.869363  12.98456    7.7467036  6.1106358 17.382072  14.663754
 13.585376  14.042943  12.854537  11.06943   13.000573  10.636344
  7.4401474 11.512548  12.80058   12.804268  17.793713  11.495619
 11.9550495 11.611424  10.840712  10.177995  14.148771   9.668085
 10.646391  17.111322   8.7000475 10.639864  10.768837  10.115247
  9.043969  11.388705  16.16515   12.1091175 14.953008  15.227695
  9.814675   8.490563  10.414546   9.3378935 15.500755  14.057771
 12.443708  16.341843  13.175412  13.425285  12.490506  14.549545
 12.42283   13.416706  10.899035  11.398053  15.658994   7.0498943
 12.044586  18.14811   11.28206    8.651041  14.519327  11.872755
 15.265249   8.556867  10.744593  18.101288   9.098827  14.8024645
 15.356525   8.963778  12.221563   8.999314  11.422174  11.2148285
 10.700424  11.785549  11.85924   10.732546  10.157689  11.280594
  9.249157  13.199603  13.035987   7.9128156 11.44207   10.347651
  5.848442   9.711401  10.518447  16.51929   14.797257   8.551838
 12.970884   1.9235016 15.618956  13.603754  12.018326   7.4067802
 17.402645   9.959768  12.947757   7.591377  17.550394  11.414217
 10.870264  12.290335  10.946405   9.703948   8.929755   8.870109
  2.5839062 12.344944  10.208286  11.875768  12.211165   9.760199
  9.0593405  8.776106  12.706814  12.296307  15.935987  12.665834
 16.025036   9.735954   8.748091  15.016381  12.085691   9.201538
 12.463124  17.061398  10.643041  13.006986  13.021952  14.67686
  9.587603 ]
```

VII.CONCLUSION

The project has successfully achieved its objectives by leveraging machine learning techniques to analyze and predict student learning outcomes and literacy levels. Through extensive data analysis and predictive modeling, the project has gained valuable insights into the factors influencing academic performance among college students. These insights provide educators, administrators, and policymakers with a deeper understanding of student learning behaviors, enabling them to design targeted interventions and support mechanisms to maximize student success.

The developed machine learning models demonstrated with XGBoost as, 97.3% Adaboost as 95.5% ,Decision Tree as 96.2%, Gradient boosting as 97.2% accuracy, robustness, and generalization ability in predicting learning outcomes and literacy levels of college students. By identifying key factors that significantly influence student learning, such as academic background, study habits, and socio-economic status, the project highlights areas for targeted interventions and personalized support services. Moving forward, there are opportunities for further research and development, including refining predictive models, incorporating additional data sources, and exploring advanced machine learning techniques to enhance the effectiveness of predictive analytics in education. Ethical considerations, such as data privacy and fairness, remain paramount in deploying predictive analytics in educational practice to ensure trust and integrity in the educational system.

VIII.REFERENCES

[1] Hellas, A.; Ihantola, P.; Petersen, A.; Ajanovski, V.V.; Gutica, M.; Hynninen, T.; Knutas, A.; Leinonen, J.; Messom, C.; Liao, S.N. Predicting academic performance: A systematic literature review. In Proceedings of the Companion of the 23rd Annual ACM Conference on Innovation and

Technology in Computer Science Education, Larnaca, Cyprus, 2–4 July 2018; pp. 175–199.

[2] Zhang, L.; Li, K.F. Education analytics: Challenges and approaches. In Proceedings of the 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), Krakow, Poland, 16–18 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 193–198.

[3] Martin, A.J.; Nejad, H.G.; Colmar, S.; Liem, G.A.D. Adaptability: How students' responses to uncertainty and novelty predict their academic and non-academic outcomes. *J. Educ. Psychol.* **2013**, *105*, 728.

[4] Pang, Y.; Judd, N.; O'Brien, J.; Ben-Avie, M. Predicting students' graduation outcomes through support vector machines. In Proceedings of the 2017 IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, USA, 18–21 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8.

[5] Walsh, K.R.; Mahesh, S. Exploratory study using machine learning to make early predictions of student outcomes. In Proceedings of the Twenty-third Americas Conference on Information Systems, Data Science and Analytics for Decision Support (SIGDSA), Boston, MA, USA, 10–12 August 2017; AIS: Atlanta, GA, USA, 2017; pp. 1–5.

[6] Kumari, P.; Jain, P.K.; Pamula, R. An efficient use of ensemble methods to predict students academic performance. In Proceedings of the 2018 4th International Conference on Recent Advances in Information Technology (RAIT), Dhanbad, India, 15–17 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6

[7] Alejandro Pena-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works" in Expert Systems with Applications, pp. 1432-1462, 2014.

[8] A. F. ElGamal, "An educational data mining model for predicting student performance in the programming course", *International Journal of Computer Applications*, vol. 70, no. 17, 2013.

[9] C. Romero and S. Ventura, "Educational Data Mining: a review of the state art Systems Man and Cybernetics", *Part C: Applications and Reviews IEEE Transactions on*, vol. 40, no. 6, pp. 601-618, 2010.

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 3, 2024

[10] C. Romero, M. I. Lopez, J. M. Luna and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums", *Computers & Education*, vol. 68, pp. 458e472, 2013.