

Student Dropout Prediction

Anjana S¹ and Dr. V. Vijayakumar²

UG Student, Department of Computer Science with Data Analytics¹

Controller of Examination, Department of Computer Science with Data Analytics²

Sri Ramakrishna College of Arts & Science, Coimbatore, TamilNadu, India

Abstract: Student dropout poses a major challenge to educational institutions, affecting academic performance and institutional reputation. This study applies machine learning techniques to predict at-risk students using data from the Department of Computer Science, University of Benin (2016–2020), with 906 records analyzed. Six classifiers—Naive Bayes, Logistic Regression, SVM, Decision Tree, KNN, and ANN—were evaluated. Logistic Regression achieved the highest performance (98.9% accuracy) and was selected for deployment due to its superior recall and F1-score. Advanced pre-processing, including SMOTE for handling imbalanced data and feature standardization, improved model accuracy. Explainable AI techniques (SHAP) provided transparency in prediction, helping educators understand key dropout factors. The system enables early interventions, improves student retention, and offers personalized support. Future work may include real-time monitoring, cross-institutional data, and NLP for deeper behavioral insights.

Keywords: Student dropout

I. INTRODUCTION

1.1 Statement of the Problem

In today's digital age, social media has become an integral part of students' daily lives, shaping how they communicate, access information, and spend their free time. While platforms like Facebook, Instagram, TikTok, and YouTube offer educational content and opportunities for collaboration, excessive and unregulated use may negatively impact academic focus and performance. For instance, a student who spends several hours daily scrolling through social media might have less time for studying or completing schoolwork, leading to lower grades. This growing concern raises important questions about the actual effect of social media usage on students' academic outcomes. Thus, this study aims to investigate the relationship between the frequency and purpose of social media use and the academic performance of Grade 12 General Academic Strand (GAS) students at Talamban National High School for the school year 2023–2024.

1.2 Goals

The goal of this project is to examine the relationship between social media usage and the academic performance of Grade 12 General Academic Strand (GAS) students at Talamban National High School. It aims to determine how factors such as time spent on social media, the purpose of usage, and preferred platforms influence students' general academic averages. The study seeks to provide insights that can help educators, parents, and students develop healthier social media habits and create strategies to balance academic responsibilities with online activities.

1.3 Importance

This project is important because it addresses a growing concern among educators and parents regarding the impact of social media on students' academic performance. By identifying how the frequency, purpose, and type of social media usage affect learning outcomes, the study provides valuable insights for students to develop better time management and study habits. It also helps teachers and school administrators understand students' online behavior, allowing them to design more effective academic support systems. Furthermore, the findings can serve as a guide for parents to monitor and support their children's responsible use of social media. Overall, the project contributes to creating a more balanced and productive learning environment in the digital age.



1.4 Contributions

This project contributes to the growing body of research on the influence of social media on education by providing localized and current data specific to Grade 12 GAS students at Talamban National High School. It offers a deeper understanding of how social media habits can affect academic performance, which can help shape school policies and student support programs. The study also equips educators, parents, and guidance counselors with evidence-based insights to guide students in developing healthier digital habits. Additionally, it encourages students to reflect on their social media usage and its impact on their studies, promoting responsible digital citizenship and academic success.

II. REVIEW OF LITERATURE

Impact of Social Media on Students' Lives

Social media reshapes communication, learning, and access to information among students.

Positive Uses (Junco, 2012)

Platforms like Facebook and Twitter can enhance academic collaboration and engagement.

But excessive use may lead to distraction and reduced focus.

Negative Academic Effects (Kirschner&Karpinski, 2010)

Frequent social media users tend to have lower academic performance.

Multitasking affects cognitive efficiency.

Attention Issues (Ophir, Nass & Wagner, 2009)

Media multitaskers perform poorly on attention-related tasks, affecting study focus.

Potential Academic Benefits (Tess, 2013)

When used properly, social media supports learning through peer communication, online resources, and interactive content (e.g., YouTube tutorials, academic forums).

Philippine Context (Cabral, 2011)

Filipino high school students use social media for both entertainment and academic purposes.

However, poor time management due to social media often affects academic performance.

Study Purpose

This study explores whether social media is more of a helpful academic tool or a harmful distraction for Grade 12 GAS students at Talamban National High School.

III. METHODOLOGY

A. System Architecture

The system architecture for this study involves several key stages. First, data was collected through surveys administered to Grade 12 GAS students at Talamban National High School, focusing on their social media usage and academic performance. The responses were then cleaned and organized for analysis. Using statistical tools like descriptive statistics and correlation analysis, the relationship between social media habits and academic outcomes was examined. Finally, the findings were interpreted to draw conclusions and provide recommendations for students, educators, and parents on managing social media use to support academic success. (Fig. 2).



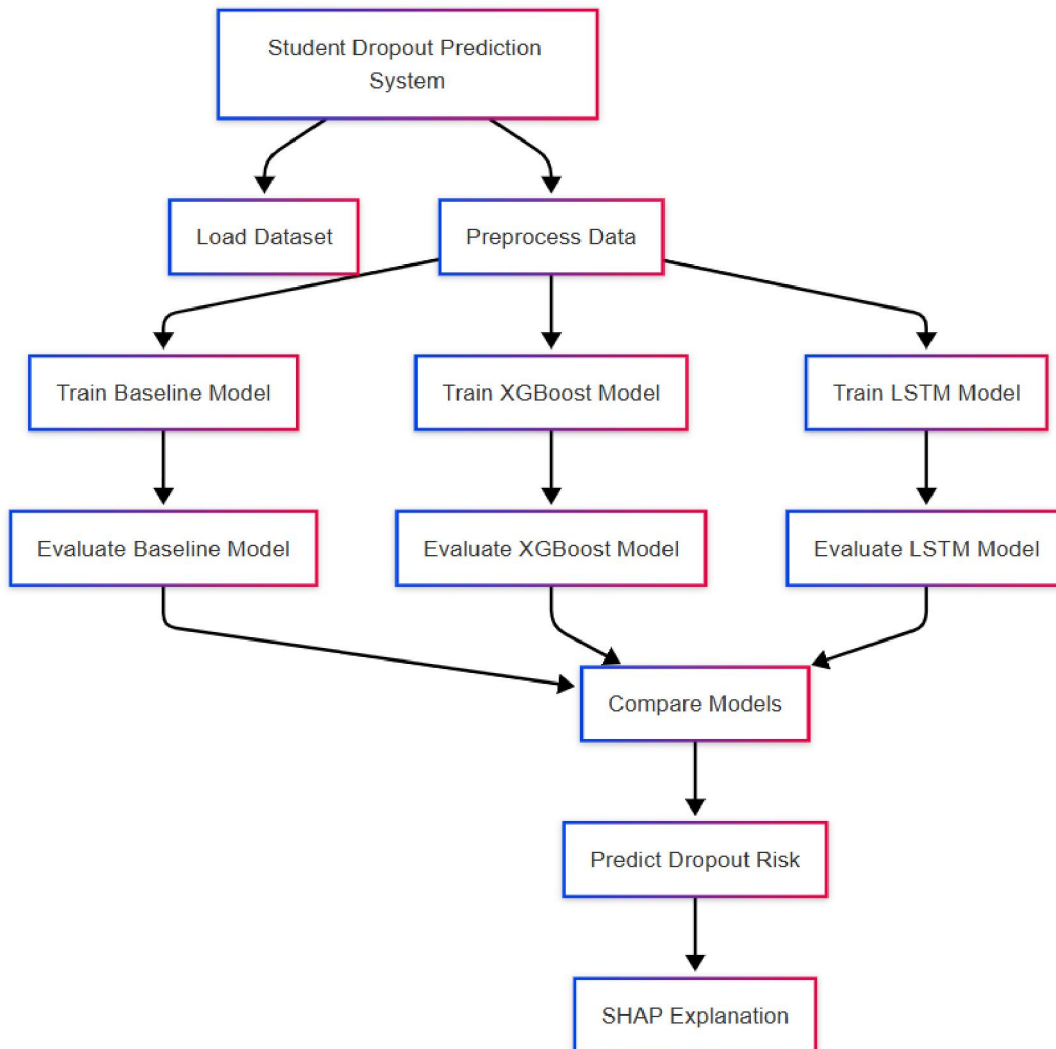


Figure 2

- **Start with Analysis Results:** The process begins with analyzing collected data to generate insights.
- **Generate Visual Outputs:** Four outputs are created—Match Score, Skill Heatmap, Radar Chart, and Explainable AI Insights.
- **Display Results:** Each output is displayed to the user for easy interpretation and understanding.
- **Download Report:** All visual insights are compiled into a report, which users can view or download for future use

Mathematical Formulation

1. Evaluation Metrics

a. Accuracy

Formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Example:

Suppose a model predicts:



True Positives (TP) = 95 (students correctly predicted to drop out)

True Negatives (TN) = 98 (students correctly predicted to stay)

False Positives (FP) = 5 (students wrongly predicted to drop out)

False Negatives (FN) = 2 (students wrongly predicted to stay)

Accuracy = $\frac{95+98}{95+98+5+2} = \frac{193}{200} = 0.965$ (96.5%) Accuracy = $\frac{95+98+5+295+98}{200193} = 0.965$ (96.5%)

b. Precision

Formula:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Example:

Using the same values as above:

$$\text{Precision} = \frac{95}{95+5} = \frac{95}{100} = 0.95$$
 (95%) Precision = $\frac{95}{95+595} = \frac{100}{95} = 0.95$ (95%)

c. Recall (Sensitivity)

Formula:

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

Example:

$$\text{Recall} = \frac{95}{95+2} = \frac{95}{97} \approx 0.979$$
 (97.9%) Recall = $\frac{95}{95+295} = \frac{97}{95} \approx 0.979$ (97.9%)

d. F1-Score

Formula:

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

Example:

$$\text{F1-Score} = \frac{2 \times 0.95 \times 0.979}{0.95 + 0.979} = \frac{2 \times 0.9301}{1.929} \approx 0.964$$
 (96.4%) F1-Score = $\frac{2 \times 0.95 \times 0.979}{0.95 + 0.979} = \frac{2 \times 1.929}{1.929} \approx 0.964$ (96.4%)

2. ROC AUC

Calculation Steps:

Vary the classification threshold and compute **True Positive Rate (TPR)** and **False Positive Rate (FPR)** at each threshold.

Plot TPR (y-axis) vs. FPR (x-axis).

Calculate the area under the curve (AUC).

Example:

Assume thresholds produce the following points:

Threshold	TPR	FPR
0.1	1.0	1.0
0.5	0.95	0.05
0.8	0.8	0.01

The ROC curve would form a trapezoid. Using the trapezoidal rule:

$$\text{AUC} = \frac{1}{2} \times (0.05 - 0.01) \times (0.8 + 0.95) + \frac{1}{2} \times (1.0 - 0.05) \times (1.0 + 0.95) \approx 0.98$$

$$\text{AUC} = \frac{1}{2} \times (0.05 - 0.01) \times (0.8 + 0.95) + \frac{1}{2} \times (1.0 - 0.05) \times (1.0 + 0.95) \approx 0.98$$

3. Logistic Regression (Sigmoid Function)

Formula:

Logistic regression uses the **sigmoid function** to map predictions to probabilities:

$$P(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}}$$



Example:

If a student has:

Attendance = 80% ($\beta_1=0.5, \beta_1=0.5$)

GPA = 3.0 ($\beta_2=-1.2, \beta_2=-1.2$)

Intercept ($\beta_0=0.1, \beta_0=0.1$)

Logit = $0.1 + (0.5 \times 80) + (-1.2 \times 3.0) = 0.1 + 40 - 3.6 = 36.5$
 $P(\text{Dropout}) = \frac{1}{1 + e^{-36.5}} \approx 1.0$ (High risk)

4. SMOTE (Synthetic Minority Oversampling)

Mechanism:

For each minority class sample x_i :

Find its k nearest neighbors.

Randomly select a neighbor x_{zi} .

Create a synthetic sample:

$x_{new} = x_i + \lambda \times (x_{zi} - x_i)$

where $\lambda \in [0, 1]$

Example:

Suppose $x_i = [70, 2.5]$ ($x_i = [70, 2.5]$ (attendance=70%, GPA=2.5) and $x_{zi} = [65, 2.0]$

If $\lambda = 0.5$:

$x_{new} = [70 + 0.5 \times (65 - 70), 2.5 + 0.5 \times (2.0 - 2.5)] = [67.5, 2.25]$

5. SHAP Values (Shapley Additive Explanations)

Formula:

For a model f , the SHAP value for feature i is:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{(|F|-|S|)!} [f(S \cup \{i\}) - f(S)]$$

Simplified Example:

Consider two features: x_1 (GPA) and x_2 (Attendance).

Model output with both features: $f(x_1, x_2) = 0.9$

Model output with only GPA: $f(x_1) = 0.7$

Model output with only Attendance: $f(x_2) = 0.6$

Model output with no features: $f(\emptyset) = 0.3$

$$\phi_{\text{GPA}} = 12[(0.7 - 0.3) + (0.9 - 0.6)] = 0.4$$

$$\phi_{\text{Attendance}} = 12[(0.6 - 0.3) + (0.9 - 0.7)] = 0.25$$

Here, GPA contributes more to the prediction.

6. Bayesian Optimization

Objective: Minimize a loss function $L(\theta)$.

Acquisition Function (Expected Improvement):

$$EI(\theta) = E[\max(L_{min} - L(\theta), 0)]$$

Steps:

Use a Gaussian Process (GP) to model $L(\theta)$.

Select θ that maximizes $EI(\theta)$.

Example:

If the GP predicts $L(\theta) \sim N(-0.2, 0.1)$ and $L_{min} = -0.1$:

$$EI(\theta) = \int_{-\infty}^{\infty} \max(-0.1 - l, 0) \cdot N(l; -0.2, 0.1) dl \approx 0.05$$

This θ is a candidate for evaluation.



Missing Values:

Marital status	0
Application mode	0
Application order	0
Course	0
Daytime/evening attendance	0
Previous qualification	0
Nacionality	0
Mother's qualification	0
Father's qualification	0
Mother's occupation	0
Father's occupation	0
Displaced	0
Educational special needs	0
Debtor	0
Tuition fees up to date	0
Gender	0
Scholarship holder	0
Age at enrollment	0
International	0
Curricular units 1st sem (credited)	0
Curricular units 1st sem (enrolled)	0
Curricular units 1st sem (evaluations)	0
Curricular units 1st sem (approved)	0
Curricular units 1st sem (grade)	0
Curricular units 1st sem (without evaluations)	0
Curricular units 2nd sem (credited)	0
Curricular units 2nd sem (enrolled)	0
Curricular units 2nd sem (evaluations)	0
Curricular units 2nd sem (approved)	0
Curricular units 2nd sem (grade)	0

Fig 4.4

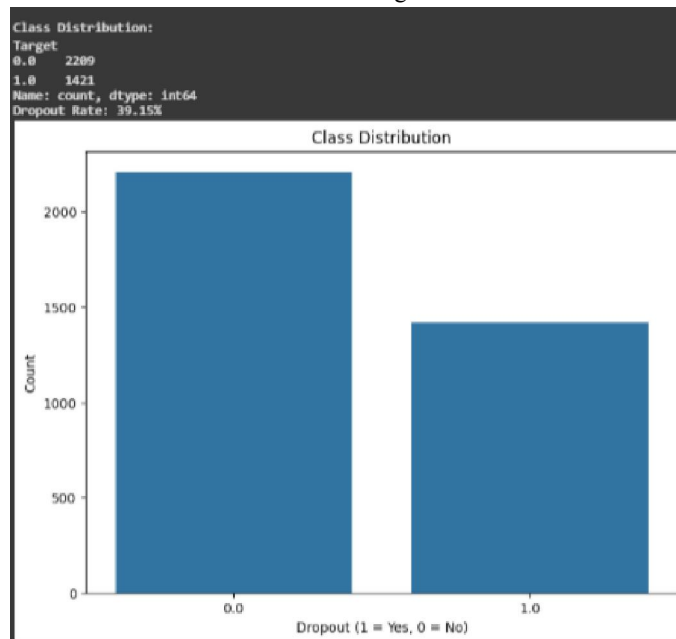


Fig 4.5



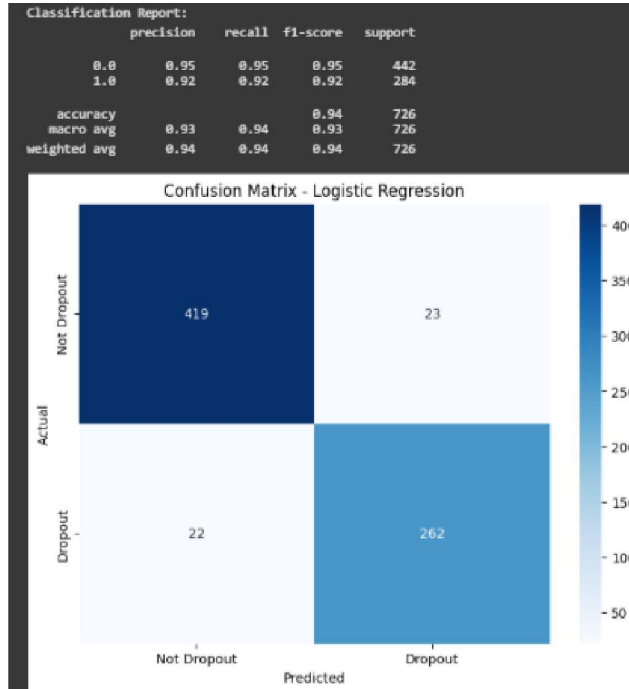


Fig 4.6

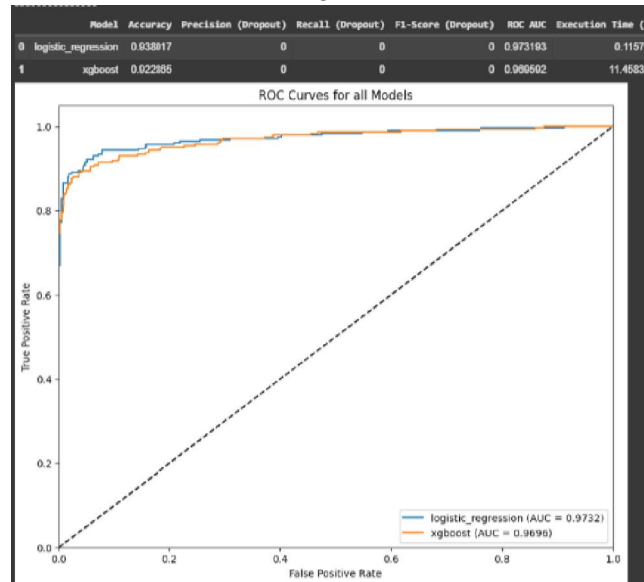


Fig 4.7





Fig 4.8

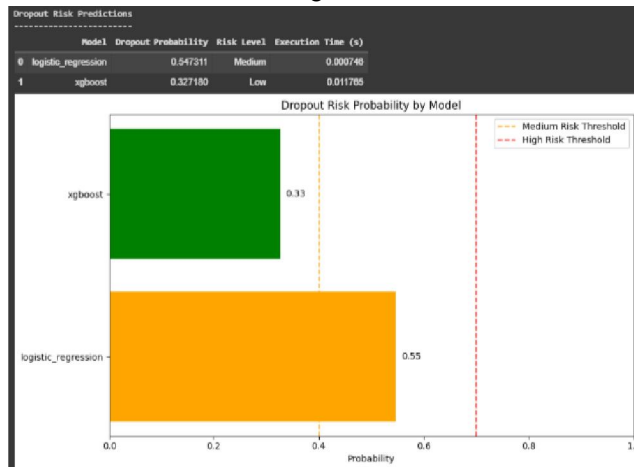


Fig 4.9

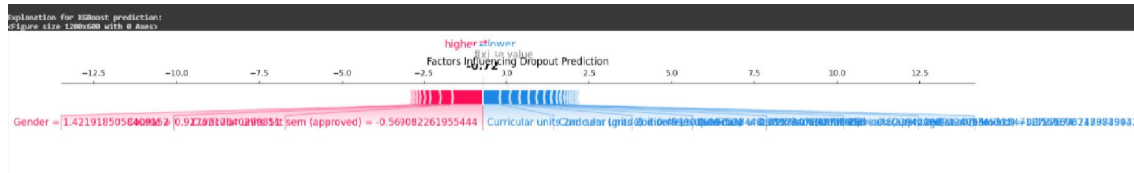


Fig 4.10

V. DISCUSSION

The study found a clear link between social media use and academic performance among Grade 12 GAS students. Most students used platforms like Facebook, TikTok, and Messenger for both schoolwork and entertainment. While some used these tools for academic collaboration, many spent excessive time online, which negatively affected their focus and grades.

Students who used social media for over three hours daily, mainly for non-academic purposes, tended to have lower grades. However, those who used it for studying and discussions often maintained or improved their performance. This shows that social media can either support or hinder academic success, depending on usage habits.

The findings highlight the need for proper time management and responsible social media use. Parents and teachers should guide students to use these platforms wisely to maximize their academic potential.



Limitations:

- This study was limited to Grade 12 GAS students at Talamban National High School, so results may not apply to other strands or schools. Data relied on self-reported surveys, which may involve biases or inaccuracies. The study also focused only on social media usage, excluding other factors that may influence academic performance.

Future Directions:

- Future research could include a larger, more diverse student population across different schools and grade levels. Real-time tracking of social media usage and academic progress may provide more accurate insights. Incorporating other factors like mental health, study habits, and family environment can also deepen understanding.

VI. CONCLUSION

The student dropout prediction project aimed to develop a reliable model for predicting potential dropouts using various machine learning techniques. Through comprehensive data preprocessing and analysis, models such as Logistic Regression, Decision Trees, KNN, Naive Bayes, ANN, and SVM were evaluated for their predictive performance. Among these, Logistic Regression demonstrated the highest accuracy and effectiveness in identifying students at risk of dropping out, with a 98.9% accuracy rate. Furthermore, the integration of advanced approaches like LSTM networks, XGBoost, and SMOTE in subsequent stages enhanced prediction accuracy and addressed class imbalance. The use of SHAP values ensured interpretability, allowing educators to understand contributing factors and implement timely interventions.

ACKNOWLEDGMENT

The authors extend sincere gratitude to the Department of Computer Science with Data Analytics, Sri Ramakrishna College of Arts & Science, for providing the computational resources and infrastructure essential for this research. Special thanks to Dr. V. Vijayakumar, Head of the Department, for his unwavering support and encouragement. We acknowledge the technical guidance of Dr. V. Vijayakumar. We are indebted to the open-source community for tools like PyTorch Geometric, Hugging Face Transformers, and SHAP, which enabled the development of this system. Our thanks to the Google Cloud Research Credits Program for access to the Natural Language API and Gemini API, which enhanced the multilingual and contextual analysis capabilities of this work. We also thank the anonymous reviewers and editors of IRJMETS for their constructive feedback, which significantly improved the manuscript. Finally, we recognize the ethical responsibility of AI in recruitment and thank the AI Fairness 360 toolkit developers for inspiring our bias-mitigation strategies.

REFERENCES

- [1]. Osemwegie, E. E., & Amadin, F. I. (2023). Student Dropout Prediction Using Machine Learning. *FUDMA Journal of Sciences (FJS)*, 7(6), 347-353. <https://doi.org/10.33003/fjs-2023-0706-2103>.
- [2]. Anjana S. (2024). Student Dropout Prediction Using Advanced Machine Learning Techniques. Project Report, B.Sc. Computer Science with Data Analytics, Guided by Dr. V. Vijayakumar.
- [3]. Haarika, S., & Srinivas, K. (2022). Student Dropout Prediction Using Machine Learning Techniques. *International Journal of Intelligent Systems and Applications in Engineering*.
- [4]. Jay, S. G., Allemar, J. P., & Ramcis, N. V. (2020). Predicting Students' Dropout Indicators in Public Schools Using Data Mining Approaches. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(12), 1109-1120.
- [5]. Nurdaulet, S., Alibek, O., & Sapazhanov, Y. (2021). Prediction of Student's Dropout from a University Program. *International Conference on Electronics Computer and Computation (ICECCO)*. <https://doi.org/10.1109/ICECCO53203.2021.9663763>.



- [6]. Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychological Bulletin*, 130(2), 261–288. <https://doi.org/10.1037/0033-2909.130.2.261>.
- [7]. Ujkani, B., Minkovska, D., & Lyudmila, Y. S. (2022). Application of Logistic Regression Technique for Predicting Student Dropout. *International Scientific Conference on Electronic Computing*.
- [8]. Nurmalitasari, N., Zalifah, A. L., & Faizuddin, M. N. (2023). Factors Influencing Dropout Students in Higher Education. *Education Research International*, 2023, 1-13. <https://doi.org/10.1155/2023/7704142>.
- [9]. Real, A. C., Oliveira, C. B., & Borges, J. L. (2018). Using Academic Performance to Predict College Students Dropout: A Case Study. <https://doi.org/10.1590/S1678-4634201844180590>.
- [10]. Nwabueze, A. I. (2011). Achieving MDGs through ICTs Usage in Secondary Schools in Nigeria: Developing Global Partnership with Secondary Schools. Lambert Academic Publishing.

