Advances in Consumer Research
https://acr-journal.com/
Volume-3 | Issue-2 | Feb 2026
Original Researcher Article

A Machine Learning–Based Model for Early Identification of Slow Learners in the Bachelor of Computer Applications Program

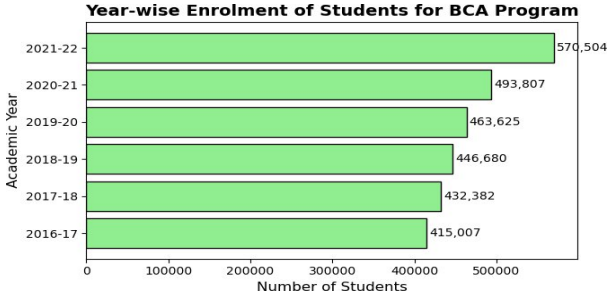
Santosh P. Nalawade[1], Dr. Rajendra S. Pujari[2]
[1]Research Scholar, Bharati Vidyapeeth (Deemed to be University), Institute of Management and Rural Development Administration, Sangli, Maharashtra, India
Email:ID: nalawadesantosh011@gmail.com;
[2]Head, Department of Computer Applications, Bharati Vidyapeeth (Deemed to be University), Institute of Management and Rural Development Administration, Sangli, Maharashtra, India
Email:ID:rajendraspujari@gmail.com

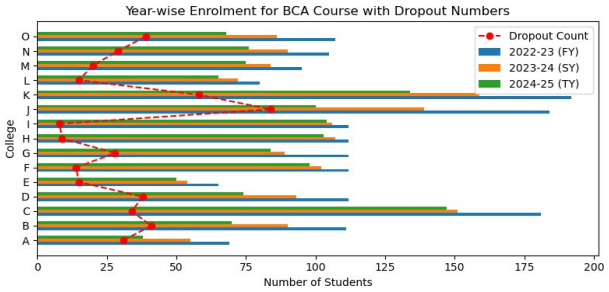
ABSTRACT
The rapid growth in enrollment in professional computing programs such as the Bachelor of Computer Applications (BCA) has been accompanied by a notable increase in student dropout rates. Early identification of academically slow learners is therefore essential for timely intervention and improved retention. This study proposes a machine learning–based model for the early identification of slow learners in the BCA program using academic, demographic, behavioral, psychological, and technological factors. Primary data were collected from BCA students enrolled in 11 colleges affiliated with Shivaji University, Kolhapur, using a structured questionnaire. Following a pilot study (n = 580), correlation analysis and Chi-square tests were applied to identify significant predictors, resulting in the selection of 20 influential variables. Multiple classification algorithms were implemented using the WEKA tool, and their performance was compared. The REPTree algorithm demonstrated an optimal balance between accuracy, recall, computational efficiency, and interpretability. The findings confirm that machine learning techniques can effectively support the early identification of slow learners and provide a data-driven basis for academic interventions aimed at reducing dropout rates and improving student performance.
Keywords: *Slow learners; At-risk students; Machine learning; BCA program; Student retention*

1. INTRODUCTION

Professional and technology-oriented programs such as the Bachelor of Computer Applications (BCA) have experienced significant growth in student enrollment due to expanding opportunities in the

information technology sector. Despite this growth, higher education institutions continue to face high dropout rates, particularly in computing programs that demand strong analytical and programming skills [2][3]. High dropout rates negatively impact academic quality, institutional reputation, and workforce readiness, making early identification and intervention critical for sustainable educational outcomes [4][5].





According to the All India Survey on Higher Education (AISHE), BCA enrollment increased by more than 155,000 students between 2016–17 and 2021–22. [1]

However, data collected from 11 colleges affiliated with Shivaji University revealed that only 1,003 out of 1,362 students enrolled in 2022–23 continued their

studies in 2024–25, corresponding to a dropout rate of 26.35%. This trend highlights the urgent need for early identification of academically slow (at-risk) learner's students.

Research Gap and Objectives

Most existing studies on student dropout and slow learner identification rely on historical academic records or qualitative approaches, limiting their applicability for

early-stage prediction [6][7]. In addition, many studies do not incorporate course- specific factors relevant to computing education or integrate behavioral, psychological, and technological indicators within a unified model. [8]- [10]

The objectives of this study are:

To identify significant factors influencing slow learning in the BCA program.

To develop a machine learning–based model for early identification of slow learners.

To evaluate and compare the performance of different classification algorithms.

Data Source and Variables

This study is based on primary data collected from students enrolled in the BCA program across 11 colleges affiliated with Shivaji University, Kolhapur. Data were gathered using a structured questionnaire administered through Google Forms to ensure ease of access, accuracy, and efficient data collection.

The full study was conducted on a total sample of N = 2028 BCA students.

The following variables were considered in the study:

(1) age, (2) gender, (3) family income, (4) father's

education, (5) mother's education, (6) home environment, (7) 12th stream, (8) 12th percentage,

(9) mode of instruction in the 12th class, (10) daily study hours, (11) study preference, (12) coaching or tuition classes, (13) examination preparation strategy,

(14) frequency of seeking help from instructors, (15) practice of programming outside the classroom, (16) use of LMS tools, (17) availability of teaching infrastructure, (18) part-time employment, (19) first- semester percentage, (20) mathematics proficiency,

(21) basic computer knowledge, (22) understanding of basic programming concepts, (23) class attendance, (24) timely submission of assignments,

(25) attention level in class, (26) frequency of

examination stress, (27) confidence during oral presentations, (28) internal class test performance,

(29) improvement in programming concepts, and (30) use of technology in teaching.

Initially, 30 variables representing demographic, socio-economic, academic, behavioral, psychological, and technological dimensions were considered. Following statistical validation using correlation analysis and Chi-square tests, 20 influential variables were selected for model development.

Data Preprocessing

All categorical variables were encoded into numerical form to ensure compatibility with machine learning algorithms [11]. Correlation analysis and Chi-square tests were conducted to identify significant relationships between student attributes and academic performance. Strong positive correlations were observed for first-semester performance, internal test scores, attendance, assignment submission, programming practice, mathematics proficiency, and LMS usage [12].

METHODOLOGY

Pilot Study

A pilot study involving 580 students was conducted to validate the questionnaire and refine variables.

Machine Learning Implementation To ensure better generalization performance and to minimize sampling bias, 10-fold cross-validation was employed [13]. The dataset was randomly divided into ten equal subsets. In each iteration, nine subsets were used for training the model and the remaining subset was used for testing. This process was repeated ten times, and the average performance across all folds was considered as the final result. The WEKA machine learning toolkit was used for preprocessing, classification, and evaluation [14]. The dataset was converted into CSV format, and several classifiers—including Naive Bayes, Logistic Regression, SMO, J48, Random Forest, Random Tree, and REPTree—were applied [15]

Performance Evaluation:

In addition to classification accuracy, the performance of the models was evaluated using Precision, Recall, and F1-score [16][17]. Precision measures the proportion of correctly predicted positive cases out of all predicted positive cases. Recall (Sensitivity) measures the proportion of correctly predicted positive cases out of all actual positive cases. The F1-score is the harmonic mean of Precision and Recall and provides a balanced measure, especially when the class distribution is imbalanced [18].

Feature Selection:

The decision tree generated using the REPTree algorithm is presented in Figure 1.1. The tree illustrates the hierarchical structure of significant factors influencing the identification of slow learners. The root node and subsequent branches indicate the relative importance of variables used in classification. The final model was constructed using more than ten statistically significant factors obtained after feature selection.

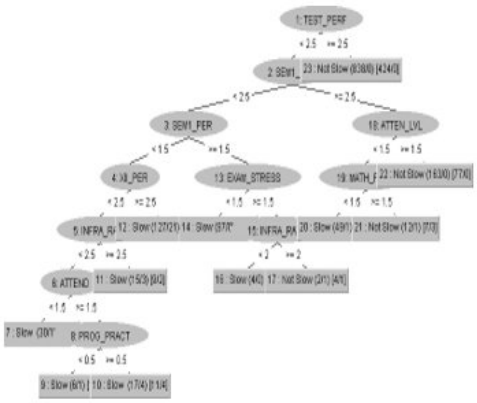


Figure 1.1

Classification Accuracy (%) of Different Algorithms

Sr. No.	Algorithms	All Data %	Training %	Testing %
1	NaiveBayes	0.948	0.945	0.979
Sr. No.	Algorithms	All Data %	Training %	Testing %
2	NaiveBayesMultinomial	0.948	0.942	0.986
3	Logistic	0.941	0.971	0.99
4	NiveBayesUpdateable	0.948	0.945	0.979
5	SimpleLogistic	0.948	0.952	0.972
6	SMO	0.955	0.962	0.993
7	LWL	0.918	0.902	0.965
8	AdaBoostM1	0.918	0.902	0.965
9	ClassificationViaRegression	0.952	0.94	0.979
10	Decision Table	0.955	0.962	0.993
11	JRip	0.95	0.971	0.972
12	DecisionStump	0.918	0.902	0.965
13	J48	0.957	0.988	0.936
14	LMT	0.948	0.993	0.943
15	RandomForest	0.961	1	0.993
16	RandomTree	0.939	1	0.965
17	REPTree	0.937	0.942	0.979

Classification Accuracy (%) of Different Algorithms using Jupyter Notebook

Sr. No.	Algorithms	Jupyter Notebook All Data %	Weka Tools All Data %
1	NaiveBayes	0.946	0.948
2	NaiveBayesMultinomial	0.937	0.948
3	Logistic	0.92	0.941
4	NiveBayesUpdateable	0.93	0.948
5	SimpleLogistic	0.93	0.948
6	SMO	0.955	0.955
7	LWL	0.92	0.918
8	AdaBoostM1	0.937	0.918
9	ClassificationViaRegression	0.94	0.952
10	Decision Table	0.94	0.955
11	JRip	0.95	0.95
12	DecisionStump	0.92	0.918
13	J48	0.96	0.957
14	LMT	0.95	0.948
Sr. No.	Algorithms	Jupyter Notebook All Data %	Weka Tools All Data %
15	RandomForest	0.94	0.961
16	RandomTree	0.94	0.939
17	REPTree	0.94	0.937

Correlation Coefficient (r)

The correlation coefficient (**r**) ranges between **-1 and +1**. Each row represents the strength and direction of the relationship between a student-related factor and academic result (marks, grades, or performance).

The correlation coefficient values range from 0.5651 to 0.9787, showing positive correlations — meaning that as a factor improves, student results also improve.

Correlation of all factors with Result

Factors	Result
SEM1_PER	0.909736
PROG_IMPROVE	0.806844
TEST_PERF	0.798501
ATTEN_LVL	0.782494
ORAL_CONF	0.768372
ATTEND	0.757949
EXAM_STRESS	0.752969

PROG_CONC	0.750794
ASSIGN_TIME	0.731053
MATH_PROF	0.701897
XII_PER	0.655605
COMP_KNOW	0.63603
EXAM_PREPARE	0.629464
INFRA_RATE	0.624677
STUDY_HRS	0.620823
PROG_PRACT	0.611992
JOB_RESP	0.549836
LMS_COMF	0.526377
HOME_ENV	0.506619
TECH_USE	0.479027

Chi-Square Test in Educational Data Analysis: The Chi-Square Test is used to analyze whether a significant relationship exists between two categorical variables in educational datasets. It helps determine if factors such as study hours and result category or attendance level and learning outcome are statistically related

Hypotheses, Formula & p-value Interpretation

Sr. No.	Variable	Chi2	p_value
1	SEM1_PER	552.324803	4.11E-114
2	XII_PER	462.200891	8.98E-95
3	ORAL_CONF	461.070254	1.57E-94
4	PROG_IMPROVE	446.964973	1.65E-91
5	TEST_PERF	425.357495	7.01E-87
6	PROG_CONC	420.290385	8.53E-86
7	ATTEN_LVL	394.966954	2.23E-80
8	ATTEND	394.6733	2.58E-80
9	ASSIGN_TIME	374.49808	5.31E-76
10	EXAM_STRESS	358.587968	1.33E-72
11	STUDY_HRS	353.82971	1.38E-71
12	MATH_PROF	345.276022	9.23E-70
13	EXAM_PREPARE	327.695966	5.19E-66
14	COMP_KNOW	318.319991	5.17E-64
15	LMS_COMF	314.510067	3.35E-63
16	INFRA_RATE	289.773126	6.17E-58
17	PROG_PRACT	270.733204	6.87E-54
18	HELP_FREQ	191.112316	4.75E-37
19	JOB_RESP	169.267195	1.75E-37
20	HOME_ENV	166.677473	6.41E-32

Interpretation of Heatmap:

The heatmap visually represents the correlation strength between different student factors. Dark red shades indicate a strong positive correlation, showing variables that increase together. Dark blue shades represent a strong negative correlation, meaning when one factor increases, the other decreases. Light- colored areas indicate weak or no correlation, helping identify which variables have minimal influence on each other for the study [19][20].

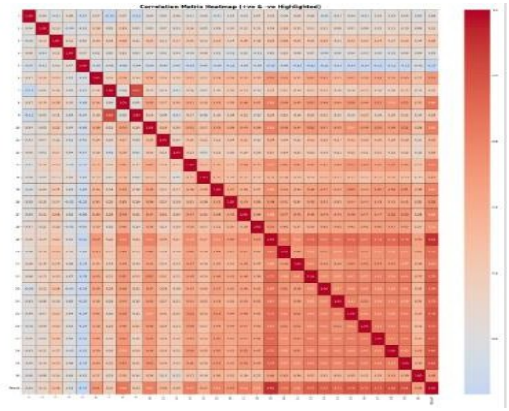


Figure 1.2

Correlation analysis and Chi-square tests were employed to identify significant relationships between student factors and academic performance. Based on these analyses, the most influential factors were selected for model development.

Finalized List of 20 Factors:

1. Home Environment	11. Basic Computer Knowledge
2. 12th Percentage	12. Basic Programming Concepts
3. Daily Study Hours	13. Class Attendance
4. Exam Preparation Strategy	14. Timely Assignment Submission
5. Practice of Programming Outside Class	15. Students' Attention Level
6. Use of LMS Tools (Moodle, Google Classroom)	16. Exam Stress Frequency
7. Availability of Teaching Infrastructure	17. Confidence During Oral Presentations
8. Part-Time Job	18. Internal Class Test Performance
9. First Semester Percentage	19. Improvement in Programming Concepts
10. Mathematics Proficiency	20. Use of Technology in Teaching

RESULTS AND DISCUSSION

The comparative analysis indicated that the REPTree algorithm achieved balanced and robust performance,

characterized by high recall (0.979), stable testing accuracy, rapid execution time, and an effectively pruned decision tree structure. Compared to more complex ensemble-based methods, REPTree offers greater interpretability, providing transparent decision rules that are highly valuable for academic decision- making and early intervention planning.

Prior to model construction, Correlation Analysis and Chi-Square statistical tests were employed as feature selection techniques to identify the most significant variables influencing student performance. Through this statistical screening process, 20 key factors were selected from the original dataset. These factors demonstrated strong associations with students' academic outcomes and contributed substantially to model effectiveness.

The findings confirm that slow learning in the BCA program is not driven by a single parameter but is influenced by a combination of academic preparedness, learning behavior, psychological confidence, and technological exposure. The integration of statistical feature selection with the REPTree algorithm enhanced both the predictive accuracy and interpretability of the model, making it suitable for practical implementation in educational institutions.

CONCLUSION AND FUTURE WORK:

This study proposes an effective machine learning– based approach for early identification of slow (at- risk) learners in the BCA program, supporting proactive interventions and dropout reduction. The REPTree algorithm demonstrated balanced predictive performance, fast execution, and interpretability, making it suitable for educational data mining. Correlation Analysis and Chi-Square tests were used to select 20 significant factors capturing academic, behavioral, psychological, and technological influences. Future work will focus on training an enhanced REPTree-based model using these 20 factors and validating it on larger, multi-institutional datasets. Integration into institutional early-warning systems and longitudinal performance analysis will

further strengthen slow learners - early-risk prediction.

REFERENCES

1. Ministry of Education, Government of India, "All India Survey on Higher Education (AISHE) 2021–22 Report," New Delhi, India, 2023.
2. V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," Rev. Educ. Res., vol. 45, no. 1, pp. 89–125, 1975.
3. K. Mishra and S. Jha, "Student retention in higher education: A systematic review," Int. J. Educ. Dev., vol. 87, pp. 102–113, 2021.
4. J. R. Bean and B. S. Metzner, "A conceptual model of nontraditional undergraduate student attrition," Rev. Educ. Res., vol. 55, no. 4, pp. 485– 540, 1985.
5. UNESCO, "Reducing Dropout and Improving Completion in Higher Education," Paris, France, UNESCO Publishing, 2020.
6. T. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," Rev. Educ. Res., vol. 45, no. 1, pp. 89–125, 1975.
7. V. Raju and R. Kumar, "Student dropout prediction in higher education using machine learning techniques," IEEE Access, vol. 8, pp. 1–12, 2020.
8. R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in Learning Analytics, New York, NY, USA: Springer, 2014, pp. 61–75.
9. H. Huang and S. Fang, "Early prediction of college student dropout using machine learning," Int.J. Emerging Technol. Learn., vol. 17, no. 3, pp. 4–18, 2022.
10. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," in Proc. 2017 ACM SIGKDD Int. Conf. Learn. Analytics Knowl., 2017, pp. 1–8.
11. H. Witten, E. Frank, M. A. Hall, and C. J. Pal,Data Mining: Practical Machine Learning Tools andechniques*, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2017.
12. R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," J. Educ. Data Mining, vol. 1, no. 1, pp. 3– 17, 2009.
13. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI), 1995, pp. 1137–1143.
14. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2017.
15. M. Hall et al., "The WEKA data mining software: An update," SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10–18, 2009.
16. C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.
17. D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," J. Mach. Learn. Technol., vol. 2, no. 1, pp. 37–63, 2011.
18. N. Japkowicz and M. Shah, Evaluating Learning Algorithms: A Classification Perspective. Cambridge, U.K.: Cambridge Univ. Press, 2011.

19. R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021.

20. Agresti, *An Introduction to Categorical Data Analysis*, 3rd ed. Hoboken, NJ, USA: Wiley, 2019

.