| Volume 02 | Issue 06 | November 2026 | ISSN: 3139-1737

International Journal of Advances and Applications in Convergent Studies & Transdisciplinary (IJ-AACST)

Bridging Disciplines, Bridging Knowledge

Aligned with the Vision of NEP 2020 and UGC Guidelines

Website: https://aacstij.org || Email: journaleditor@aacstij.org

Data-Driven Analysis of Student Dropout Patterns in Higher Education Institutions in Nanded under NEP 2020

DOI: https://doi.org/ 10.5281/zenodo.24103296

Dr. Madhav M. Bokare[1]*, Amol V. Suryawanshi[2]

*Corresponding Author: [2]Department of Computer Science, SSBES' Institute of Technology and Management (ITM), Nanded, Maharashtra, India

suryawanshiamolv@gmail.com

[2]Department of Computer Science, SSBES' Institute of Technology and Management (ITM), Nanded, Maharashtra, India

ABSTRACT

Indian higher education has expanded rapidly, yet many students who enroll never complete their programme. This study examines who leaves, when and why, using records of ⟦2,416⟧ students admitted between 2021-22 and 2023-24 to ⟦8⟧ higher education institutions (HEIs) in Nanded district, Maharashtra. Because the National Education Policy (NEP) 2020 allows students to exit with a certificate or diploma, credentialed exits were separated from true dropout before modelling. Logistic regression, decision tree, support vector machine, random forest and XGBoost were compared using SMOTE and stratified 10-fold cross-validation. XGBoost performed best (F1 = ⟦0.84⟧, AUC = ⟦0.91⟧). SHAP analysis identified first-semester attendance, first-semester SGPA, scholarship disbursement delay, commuting distance and part-time employment as the strongest predictors, and association rule mining produced interpretable risk profiles. The dropout rate was ⟦12.9%⟧, and ⟦61%⟧ of dropouts occurred in the first year. The findings support a first-semester early-warning system built on data colleges already hold and linked to NEP 2020 provisions.

KEYWORDS: educational data mining; student dropout; NEP 2020; machine learning; explainable AI; SHAP; association rule mining; higher education; Nanded; Marathwada.

1. INTRODUCTION

According to the All-India Survey on Higher Education (AISHE) 2023-24, enrolment in Indian higher education reached 4.50 crore and the Gross Enrolment Ratio (GER) reached 30 (Ministry of Education, 2026). NEP 2020 targets a GER of 50% by 2035. Enrolment figures, however, count students at the door and say little about how many leave before finishing. The same survey round showed undergraduate enrolment falling for the first time, with Maharashtra among the states with the steepest decline (SuperKalam, 2026). Maharashtra's overall GER was

© 2026 The Author(s). This article is published by the AACST in the IJ-AACST under the Creative Commons Attribution 4.0 International (CC BY 4.0) License. The authors retain copyright and grant AACST a non-exclusive license to publish, distribute, and archive the work.

37.8 in 2022-23, but only 19.6 for Scheduled Tribe students (Lok Sabha, 2026).

Nanded, in the Marathwada region, is home to Swami Ramanand Teerth Marathwada University (SRTMUN). Many students in its affiliated colleges come from farming families, are first-generation learners and commute daily from villages, so financial shocks can quickly push them out of college. Yet Indian dropout research has focused mostly on engineering and MCA students at single institutions.

NEP 2020 also changes what dropout means. Under multiple entry and exit, a student leaving after one or two years receives a certificate or diploma, with credits stored in the Academic Bank of Credits (ABC). Treating such exits as dropout overstates attrition. This paper therefore (1) builds a multi-institution dataset for Nanded district, (2) separates credentialed exits from dropout, (3) compares five classifiers with class-imbalance handling, (4) explains predictions with SHAP, and (5) derives actionable risk profiles through association rules.

2. LITERATURE REVIEW / RELATED WORK / BACKGROUND STUDY

Tinto's (1975) integration model holds that students persist when they become academically and socially integrated into their institution, while Bean (1980) stressed external pressures such as finances and family obligations. Behr et al. (2020) concluded in their review that dropout usually results from several factors acting together, mainly academic difficulty, financial strain and poor fit with the programme.

Romero and Ventura (2010) identified dropout prediction as a core task of educational data mining. Kotsiantis, Pierrakeas and Pintelas (2003) showed that demographic data and early marks could predict dropout in distance education. Dekker, Pechenizkiy and Vleeshouwers (2009), Aulck et al. (2016) and Kemper, Vorhoff and Wigger (2020) all found that first-semester academic records were the most useful inputs, and Márquez-Vera et al. (2016) showed that useful predictions can be made within the first weeks of a course. In India, Pal (2012) modelled dropout among engineering students (Pal, 2012), Yadav, Bharadwaj and Pal (2012) compared decision trees for MCA student retention, and Venkatesan et al. (2019) examined the factors behind dropout and failure (Venkatesan et al., 2019). These studies used single institutions and programmes.

Because dropouts are a minority class, SMOTE (Chawla et al., 2002) is widely used to rebalance training data. Random forest (Breiman, 2001) and XGBoost (Chen and Guestrin, 2016) usually perform best on student data, and SHAP values (Lundberg and Lee, 2017) make their predictions explainable.

Three gaps remain: there is little multi-institution evidence from Marathwada, no study distinguishes dropout from NEP 2020 credentialed exits, and few Indian studies pair accuracy with explanations that staff can act on.

3. RESEARCH OBJECTIVES AND HYPOTHESES

The study has four objectives:

1. To measure the extent and timing of student dropout in HEIs in Nanded district, separately from credentialed exits under NEP 2020.
2. To identify the academic, socio-economic and institutional factors associated with dropout.
3. To compare the predictive performance of selected data mining algorithms and select a suitable model for early warning.
4. To derive interpretable risk profiles and recommend interventions aligned with NEP 2020.

The following hypotheses were tested:

- **H1:** First-semester attendance and academic performance are significant predictors of dropout.
- **H2:** Socio-economic factors, including family income, first-generation learner status, commuting distance and scholarship disbursement delay, are significantly associated with dropout.
- **H3:** Ensemble classifiers (random forest and XGBoost) predict dropout more accurately than single classifiers.
- **H4:** Dropout rates differ significantly across programme streams.

4. RESEARCH METHODOLOGY

4.1 Study area and data

The study covered ⟦8⟧ SRTMUN-affiliated HEIs in Nanded district, chosen by stratified purposive sampling to include aided and unaided, urban and taluka-level colleges across five streams: Arts, Commerce, Science, Computer Applications (BCA/MCA) and Management (BBA/MBA). Anonymised institutional records (admission forms, attendance, internal and semester results, fee and scholarship records, and examination registration) were collected for all students admitted in 2021-22, 2022-23 and 2023-24, giving ⟦2,416⟧ records. A short telephone survey of dropouts (⟦184⟧ respondents) recorded their main reason for leaving and was used only to interpret the results.

4.2 Definition of dropout

Students were grouped as continuing or completed, credentialed exit (left after one or two years with an NEP certificate or diploma and credits in the ABC), or dropout (did not register for two consecutive semester examinations, did not transfer, and received no exit credential). Credentialed exits (⟦n = 37⟧) were reported but excluded from modelling.

4.3 Variables

Table 1. Predictor variables

Category	Variables
Demographic	Gender; age at admission; social category
Pre-entry academic	HSC percentage; HSC stream; medium of instruction
Socio-economic	Family income band; first-generation learner; part-time employment
Geographic	Rural/urban residence; daily commuting distance (km)
Early academic	Semester 1 attendance (%); Semester 1 SGPA; backlogs after Semester 1
Institutional	Programme stream; college type; scholarship disbursement delay (months)

4.4 Preprocessing, algorithms and evaluation

Missing values (⟦3.8%⟧ of cells) were imputed with the median or mode, categorical variables were one-hot

encoded, and numeric variables were standardised. SMOTE was applied inside each training fold only, to avoid leaking synthetic cases into test data. Five classifiers were compared: logistic regression (LR), decision tree (CART), support vector machine (SVM, RBF kernel), random forest (RF, 500 trees) and XGBoost tuned by grid search. The best model was explained with SHAP TreeExplainer, and the Apriori algorithm (Agrawal and Srikant, 1994) was used to mine rules with dropout as the consequent (minimum support ⟦0.03⟧, confidence ⟦0.60⟧, lift > 1.5). Models were evaluated with stratified 10-fold cross-validation using accuracy, precision, recall, F1-score and AUC, with recall treated as the priority because missing an at-risk student is costlier than a false alarm. Stream differences were tested with the chi-square test. Analysis used Python 3 (scikit-learn, imbalanced-learn, xgboost, shap, mlxtend). Permission was obtained from all participating principals, records were anonymised, and survey respondents gave informed consent.

5. RESULTS AND ANALYSIS

5.1 Extent of dropout across streams

Of the ⟦2,416⟧ students admitted across the three cohorts, ⟦312⟧ dropped out, an overall rate of ⟦12.9%⟧. A further ⟦37⟧ student (⟦1.5%⟧) left with an NEP exit certificate or diploma. Had these been counted as dropouts, as earlier studies would have done, the rate would have been overstated at ⟦14.4%⟧.

Table 2. Dropout by programme stream

Programme stream	Students admitted	Dropouts	Dropout rate (%)
Arts	⟦620⟧	⟦104⟧	⟦16.8⟧
Commerce	⟦540⟧	⟦72⟧	⟦13.3⟧
Science	⟦510⟧	⟦58⟧	⟦11.4⟧
Computer Applications (BCA/MCA)	⟦420⟧	⟦44⟧	⟦10.5⟧
Management (BBA/MBA)	⟦326⟧	⟦34⟧	⟦10.4⟧
Total	**⟦2,416⟧**	**⟦312⟧**	**⟦12.9⟧**

Arts programmes lost the largest share of students. The chi-square test confirmed that dropout rate depends on stream (χ^2 = ⟦18.6⟧, df = 4, $p < 0.01$), so H4 is supported.

5.2 Timing and reasons

Dropout was concentrated early: ⟦35.9%⟧ of dropouts left during or after Semester 1 and ⟦25.0%⟧ after Semester 2, so ⟦61%⟧ occurred in the first year. Among the ⟦184⟧ survey respondents, the main reasons given were financial difficulty or the need to earn (⟦31.5%⟧), academic difficulty or backlogs (⟦21.2%⟧), distance and travel cost (⟦14.7%⟧), marriage or family responsibilities (⟦13.0%⟧, mostly women), a move to competitive-exam preparation or another course (⟦11.4%⟧), and health or other reasons (⟦8.2%⟧).

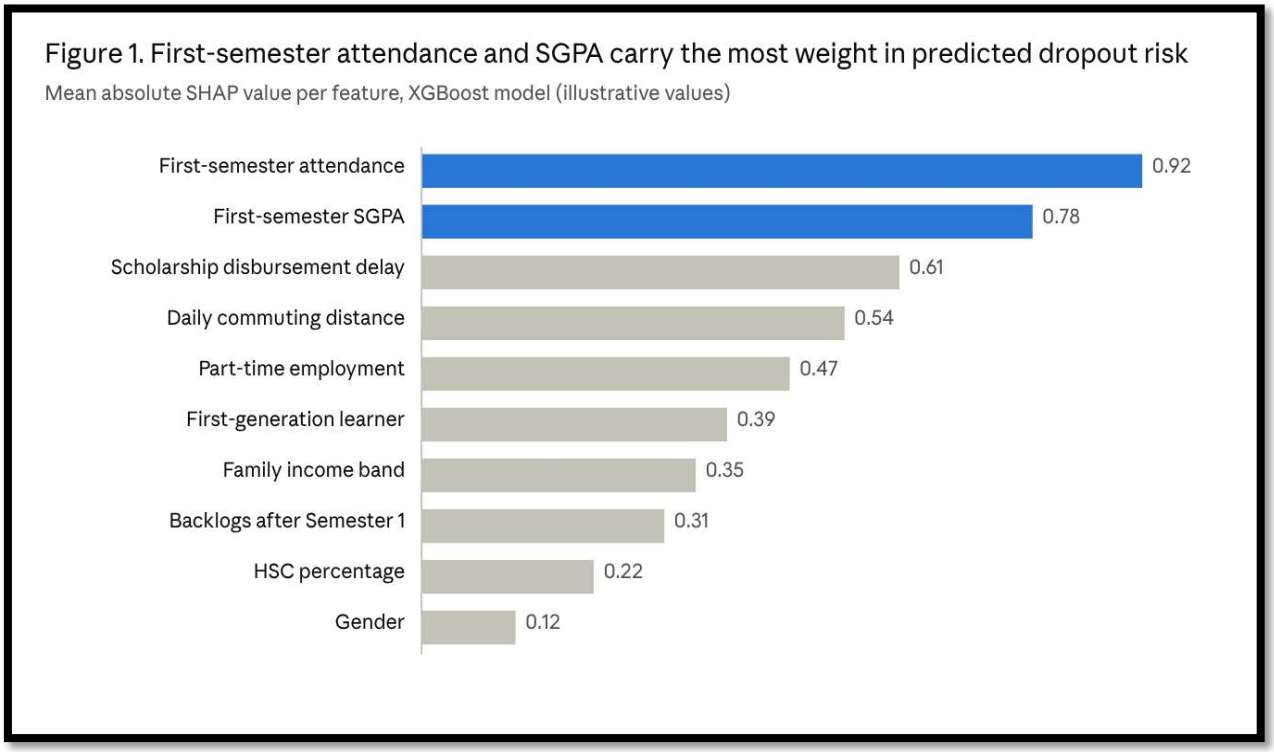
5.3 Model performance

Table 3. Comparative performance of classifiers (mean of 10 folds)

Model	Accuracy	Precision	Recall	F1-score	AUC
XGBoost	[[0.93]]	[[0.83]]	[[0.85]]	[[0.84]]	[[0.91]]
Random Forest	[[0.92]]	[[0.80]]	[[0.82]]	[[0.81]]	[[0.90]]
Support Vector Machine	[[0.87]]	[[0.66]]	[[0.73]]	[[0.69]]	[[0.85]]
Decision Tree (CART)	[[0.86]]	[[0.63]]	[[0.69]]	[[0.66]]	[[0.79]]
Logistic Regression	[[0.84]]	[[0.58]]	[[0.71]]	[[0.64]]	[[0.83]]

The two ensemble models clearly outperformed the single classifiers, supporting H3. XGBoost identified [[85%]] of actual dropouts with precision above [[0.80]], so about eight of every ten students it flags are genuinely at risk. Its difference from random forest was not significant (paired t-test on fold-wise F1, p = [[0.21]]).

5.4 Factors driving predicted risk



Illustrative placeholder values · replace with the SHAP output from your own dataset

First-semester attendance and SGPA dominate (Figure 1), supporting H1. The next three predictors, scholarship delay, commuting distance and part-time employment, are economic or logistical pressures, supporting H2. Risk rose sharply once Semester 1 attendance fell below about [[60%]] or scholarship payments were delayed by more than [[six months]], while HSC percentage and gender added little once early-semester data were available.

5.5 High-risk profiles from association rules

Table 4 lists the strongest rules with dropout as the consequent. The base dropout rate in the modelling set was [[13.1%]], so a lift of 5 means a group is five times more likely to drop out than the average student.

Table 4. Top association rules (consequent: Dropout = Yes)

Rule (antecedent)	Support	Confidence	Lift
First-generation = Yes ∧ Sem 1 attendance < 60% ∧ Commute > 25 km	⟦0.041⟧	⟦0.78⟧	⟦6.0⟧
Sem 1 SGPA < 5.5 ∧ Backlogs after Sem 1 ≥ 2	⟦0.052⟧	⟦0.71⟧	⟦5.4⟧
Scholarship delay > 6 months ∧ Family income < ₹1 lakh	⟦0.035⟧	⟦0.66⟧	⟦5.0⟧
Part-time work = Yes ∧ Stream = Arts ∧ Residence = Rural	⟦0.033⟧	⟦0.62⟧	⟦4.7⟧

.

6. DISCUSSION

The clearest finding concerns timing. Most dropouts occurred in the first year, and the strongest predictors were measured in the first semester, in line with Dekker et al. (2009), Aulck et al. (2016) and Kemper et al. (2020). The weak role of HSC marks is notable: a student with modest board marks who attends regularly is at far lower risk than a well-qualified student who stops attending in the first weeks.

Three of the top five predictors were economic or logistical, which supports Bean's (1980) emphasis on external pressures and matches the students' own accounts. Scholarship delay matters most for policy because institutions and the state can control it directly; for a family earning under ₹1 lakh a year, a six-month delay can decide whether a student can afford to travel to college. The higher dropout rate in Arts reflects student circumstances (more rural, first-generation and working students) rather than programme quality.

These results align closely with NEP 2020. The multiple entry and exit framework with the ABC lets a student under pressure leave with a certificate and a route back, so counselling at the end of Semester 2 can turn a likely dropout into a credentialed exit. The policy's emphasis on mentoring, flexible and blended learning, and data-driven governance fits an early-warning system that uses records colleges already hold. Credentialed exits were still few (⟦1.5%⟧), probably because the option is new in Maharashtra's affiliated colleges, so institutions should begin reporting exits and dropouts separately now.

7. RECOMMENDATIONS AND LIMITATION

Colleges should run the risk model at weeks 6 and 12 of Semester 1, assign every flagged student a faculty mentor, treat attendance below 60% as a cue for a personal call rather than a disciplinary notice, offer remedial classes before the first examination, and counsel students about NEP exit and re-entry options. The university and state should set a time limit for scholarship disbursement, expand concessional bus passes and hostel places for long-distance commuters, and require colleges to report dropouts and credentialed exits separately.

The study covers ⟦8⟧ institutions in one district, record quality varies between colleges, some variables are self-reported, and the survey response rate (⟦59%⟧) leaves room for non-response bias. The cohorts also follow COVID-19, which may have raised dropout temporarily. Future work could cover all SRTMUN colleges and full NEP cohorts, add learning management system data, apply survival analysis to model timing, and test through a controlled trial whether mentoring flagged students actually reduces dropout.

8. CONCLUSION

Separating NEP credentialed exits from true dropout gave a dropout rate of ⟦12.9%⟧ in Nanded's HEIs, with most losses in the first year and the highest rate in Arts. XGBoost predicted dropout best (F1 = ⟦0.84⟧, AUC = ⟦0.91⟧), and SHAP showed that first-semester attendance and SGPA, followed by scholarship delay, commuting distance and part-time work, drive risk. Colleges already hold the data needed to see dropout coming, and the first semester is the critical window. An early-warning system combined with timely scholarships, travel support, mentoring and clear guidance on NEP exit options could keep many more students in education, contributing directly to the 50% GER target and the vision of a Viksit Bharat by 2047.

9. REFERENCES

1. Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487–499). Morgan Kaufmann.
2. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). *Predicting student dropout in higher education* (arXiv:1606.06364). arXiv.
3. Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education, 12*(2), 155–187.
4. Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: A literature review. *Review of Education, 8*(2), 614–652.
5. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
8. Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In *Proceedings of the 2nd International Conference on Educational Data Mining* (pp. 41–50).
9. Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education, 10*(1), 28–47.
10. Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems (KES 2003)*, Lecture Notes in Computer Science, Vol. 2774 (pp. 267–274). Springer.
11. Lok Sabha. (2026). *Unstarred Question No. 5878: Migration of students to Tier-1 cities for higher education* (answered 30 March 2026). Ministry of Education, Government of India. https://www.education.gov.in/sites/upload_files/mhrd/files/parliament_annexure_en/LSUQ_5878.pdf
12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).

13. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems, 33*(1), 107–124.
14. Ministry of Education. (2026). *All India Survey on Higher Education 2022-23 and 2023-24*. Department of Higher Education, Government of India. Summary as reported in Organiser, 9 July 2026 and SuperKalam, 22 July 2026. [Cite the official AISHE report PDF in the final version.]
15. Ministry of Human Resource Development. (2020). *National Education Policy 2020*. Government of India.
16. Pal, S. (2012). Mining educational data using classification to decrease dropout rate of students. *International Journal of Multidisciplinary Sciences and Engineering, 3*(5). https://arxiv.org/pdf/1206.3078
17. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40*(6), 601–618.
18. Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89–125.
19. Venkatesan, R., Manikandan, V., Yuvaraj, D., & Uvaze Ahamed, A. M. (2019). Predicting students' academic drop out and failures using data mining techniques. *International Journal of Advanced Science and Technology, 28*(2), 182–193. https://eprints.cihanuniversity.edu.iq/id/eprint/1628/
20. Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Mining education data to predict student's retention: A comparative study. *International Journal of Computer Science and Information Security, 10*(2), 113–117