



Analyzing Factors Influencing Student Dropout in North Gujarat Higher Education Institutions Using Apriori Association Rule Mining

Het T. Trivedi¹ and Badal K. Kothari²

Department of Computer Science & IT, Hemchandracharya North Gujarat University (HNGU), Patan, Gujarat, India

¹htrivedi@ngu.ac.in; ²bkkothari@ngu.ac.in

Abstract— Student dropout from higher education has economic, social and academic roots, and in developing regions such as North Gujarat it erodes the human capital on which regional growth depends. This paper applies the Apriori association rule mining algorithm to uncover the combinations of factors that co-occur with dropout across the five main districts of North Gujarat: Patan, Mehsana, Banaskantha, Sabarkantha and Aravalli. Data on demographic background, socio-economic status, academic history, institutional environment and geographical constraints were collected from 1,245 dropout and active students of regional higher education institutions and universities, of which 1,210 complete records were analysed. Of 42 strong rules generated, the five with the highest confidence and lift for the dropout outcome were examined in detail. The strongest rule links low annual family income with a daily commute of more than 15 km (confidence 84.5%, lift 2.12); weak attendance combined with poor prior performance, and rural residence combined with low parental literacy and low income, form further high-risk profiles, as does English-medium instruction for students with weak prior attainment. Because these rules identify specific, observable combinations of risk, they can inform early-warning systems, transport provision and financial support designed for the circumstances of North Gujarat's students.

Keywords— Apriori algorithm, association rule mining, educational data mining (EDM), student dropout, North Gujarat, HNGU, socio-economic factors

I. INTRODUCTION

Higher education drives socio-economic progress and innovation and raises living standards in regional economies. In India, the spread of universities and regional college campuses over the past two decades has lifted gross enrolment ratios considerably. Dropout is the persistent counterweight to this expansion: each student who leaves represents institutional resources spent without result, potential left unrealized and a widening of economic inequality. The problem is especially acute in North Gujarat, which comprises the districts of Patan, Mehsana, Banaskantha, Sabarkantha and Aravalli. Mehsana has strong industrial and agricultural economies, whereas Banaskantha and Patan contend with semi-arid conditions, rural remoteness and lower baseline literacy, so the region's student population is highly heterogeneous.

Statistical studies of dropout have commonly relied on linear regression or descriptive statistics, which struggle to capture non-linear relationships among many interacting factors. Low income, for instance, may not on its own lead a student to withdraw, yet in combination with a long commute and a change in the medium of instruction, from Gujarati in school to English in college, it may become decisive. Educational data mining (EDM) offers methods for uncovering such hidden patterns in academic data [1-5]. Among them, association rule mining with the Apriori algorithm is notable for its interpretability: it finds frequent itemsets and generates conditional rules stating precisely which combinations of factors tend to occur together with dropout.

This study analyses dropout in the higher education institutions of North Gujarat, with particular reference to colleges affiliated with Hemchandracharya North Gujarat University (HNGU), Patan [4]. Organizing the analysis around demographic, academic, financial and institutional variables, it uses Apriori to extract rules that can be acted upon. These rules are intended as the basis of an automated early-warning system that flags students at risk before they formally withdraw, so that administrators can intervene in time.

II. LITERATURE REVIEW

EDM has developed into a discipline concerned with methods for analysing the distinctive kinds of data generated in educational settings [5]. Researchers have applied predictive modelling, classification and clustering to student retention and academic success [2]. Several studies, for example, have used decision trees (C4.5 and random forest) and naive Bayes classifiers to predict student failure from mid-term marks and attendance [6, 7]. Such models predict a binary outcome, dropout or retention, effectively, but they do not by themselves explain how demographic and environmental attributes combine to produce that outcome.

Association rule mining (ARM) addresses this limitation. The Apriori algorithm of Agrawal and Srikant, the foundation of market-basket analysis, has been adapted to education to relate student behaviour to learning outcomes. Earlier work indicates that socio-economic variables such as parental occupation and family income are closely linked to academic performance, and regional studies in rural India point to



seasonal agricultural labour, unreliable public transport in arid areas and language barriers as major obstacles to continuing in higher education. Little of this work, however, applies data mining to North Gujarat specifically, a region whose heavy dependence on pastoral and agricultural livelihoods coexists with emerging industrial centres. The present study fills that gap by mapping the demographic attributes of North Gujarat onto a formal Apriori framework.

III. METHODOLOGY

The study follows the standard EDM sequence: data collection, preprocessing, discretization, rule mining with Apriori, filtering of rules by support, confidence and lift, and interpretation of the resulting knowledge.

A. Data Source and Attributes

Primary and secondary data were gathered from 1,245 students enrolled in undergraduate programmes (BA, BCom, BSc, BCA and BBA) at colleges in North Gujarat. The attributes cover five dimensions: demographic (gender, location), socio-economic (family income, parental literacy), academic (previous marks, first-year attendance), institutional (medium of instruction, course type) and geographical (commuting distance). The attributes used for mining and their discretized values are listed in Table I.

ATTRIBUTE MAPPING AND DISCRETIZATION

Category	Attribute	Discretized values (items)
Demographic	LOC	Rural, Semi-urban, Urban
Socio-economic	INC	Low (<1.5L INR), Medium (1.5L–5L INR), High (>5L INR)
Socio-economic	PLIT	Low (illiterate/primary), Medium (secondary), High (graduate)
Academic	ATT	Low (<60%), Medium (60–75%), High (>75%)
Academic	PREV_PERF	Poor (<45%), Average (45–65%), Good (>65%)
Geographical	DIST	Short (<5 km), Moderate (5–15 km), Long (>15 km)
Institutional	MED	Gujarati medium, English medium
Target	STATUS	Dropout, Active

B. Apriori Formulation

Apriori finds frequent itemsets in a transactional dataset and derives implications of the form $X \rightarrow Y$, where the antecedent X and consequent Y are disjoint itemsets. Here each student record is a transaction, and Y is restricted to the outcome $\{\text{STATUS} = \text{Dropout}\}$. With N the total number of transactions, rules are evaluated by three measures, support (Supp), confidence (Conf) and lift:

$$\text{Supp}(X \rightarrow Y) = \text{Count}(X \cup Y) / N \quad (1)$$

$$\text{Conf}(X \rightarrow Y) = \text{Count}(X \cup Y) / \text{Count}(X) \quad (2)$$

$$\text{Lift}(X \rightarrow Y) = \text{Supp}(X \cup Y) / [\text{Supp}(X) \cdot \text{Supp}(Y)] \quad (3)$$

Support measures how common the full pattern is, confidence estimates the conditional probability of dropout given the antecedent, and lift compares that probability with the overall dropout rate. A rule is retained only if its support exceeds a minimum support threshold and its confidence a minimum confidence threshold. A lift greater than 1 means that dropout is more frequent among students with the antecedent than in the sample as a whole, that is, the two are positively associated; the further lift exceeds 1, the more informative the rule.

IV. DATA PREPROCESSING AND IMPLEMENTATION

The raw survey data contained missing values and inconsistent categories. Records lacking any of the primary socio-economic variables were removed, leaving 1,210 complete records. Continuous variables, such as commuting distance in kilometres and examination marks, were discretized into the intervals shown in Table I. Apriori was run in Python using the open-source mlxtend library. The minimum support was set at 0.05 (5%), low enough to retain meaningful subgroups but high enough to exclude very specific, noisy patterns, and the minimum confidence at 0.65 (65%) to ensure that retained rules are reliable predictors.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The algorithm generated 42 strong association rules. After restricting the consequent to $\{\text{STATUS} = \text{Dropout}\}$, the five rules with the highest lift and confidence were selected for detailed analysis; they are listed in Table II.

TOP ASSOCIATION RULES FOR STUDENT DROPOUT

ID	Antecedent (X) → {STATUS = Dropout}	Supp.	Conf.	Lift
R01	{INC = Low, DIST = Long}	0.082	0.845	2.12
R02	{ATT = Low, PREV_PERF = Poor}	0.114	0.792	1.98
R03	{LOC = Rural, PLIT = Low, INC = Low}	0.068	0.751	1.88
R04	{MED = English, PREV_PERF = Poor}	0.055	0.713	1.79
R05	{INC = Low, ATT = Low}	0.095	0.688	1.72

A. Analysis of the Discovered Rules

The five rules share a consistent structure. Since lift equals confidence divided by the overall dropout rate, all five imply a baseline dropout rate of about 40% in the analysed sample (for



R01, $0.845/2.12 \approx 0.40$). Against this baseline, each antecedent raises the likelihood of dropout by a factor of between 1.72 and 2.12. The rules also fall into two groups: R01, R03 and R05 describe economic and geographical disadvantage, while R02 and R04 describe academic difficulty. Low income appears in three of the five rules, which marks it as the most pervasive single risk factor, but in none of them does it act alone.

Rule R01, combining severe financial constraint ($\{INC = Low\}$) with a heavy commuting burden ($\{DIST = Long\}$), has the highest confidence (84.5%) and lift (2.12). In North Gujarat many higher education centres are concentrated in district towns such as Patan, Mehsana and Palanpur. Students from outlying village clusters in Banaskantha or rural Patan must travel long distances each day on limited state transport buses. Without the means to pay for bus fares or for hostel accommodation, they often fall behind in attendance, and this is frequently the path to dropping out. Rule R05 ($\{INC = Low, ATT = Low\}$; confidence 68.8%) is consistent with this account: it links the same financial disadvantage directly to low attendance.

Rule R02 is the most frequent of the five (support 0.114) and has the second-highest confidence (79.2%): low first-year attendance together with poor prior performance. It identifies students who arrive academically under-prepared and then disengage early, a pattern visible within the first semester and therefore well suited to early warning. Rule R03 adds a household dimension: rural residence, low parental literacy and low income together yield a dropout confidence of 75.1%, pointing to students whose families have limited means and limited experience of higher education with which to support them.

Rule R04 reveals a linguistic barrier. Students who enter professional or technical courses taught entirely in English, many of them from Gujarati-medium secondary schools, and whose prior academic performance is weak, drop out with a likelihood of 71.3%. This points to a clear need for language bridge courses and bilingual teaching aids during the first two semesters of college.

These rules describe associations, not causes: they show which combinations of circumstances co-occur with dropout in the sample, and the causal mechanisms proposed above are interpretations that further study should test. Their practical value, however, does not depend on establishing causation, because each antecedent is observable at or soon after admission and can therefore be used to identify students for support.

VI. PROPOSED INTERVENTIONS AND POLICY RECOMMENDATIONS

The parameters of the mined rules suggest the following targeted actions for universities and administrative bodies in North Gujarat:

- *Targeted transit subsidies.* Institutions should partner with the Gujarat State Road Transport Corporation (GSRTC) to provide fully subsidized point-to-point travel passes for rural students who match the $\{DIST = Long, INC = Low\}$ profile of R01.
- *Automated attendance alerts.* An automated SMS warning should be sent when a student's attendance falls below 60% in the first semester, triggering proactive academic counselling to counter the pattern of R02 (and of R05).
- *Bilingual learning resources.* Parallel reference materials in English and Gujarati should be developed for foundational technical programmes (BCA, BSc IT) to reduce the linguistic friction identified in R04.

VII. CONCLUSION AND FUTURE WORK

This study applied Apriori association rule mining to identify the structural patterns associated with student dropout in higher education institutions across North Gujarat. The results show that dropout seldom follows from a single cause; it is associated with compound risk, such as the interplay of low family income, long commutes and a change in the language of instruction. The mined rules give regional institutions such as HNGU clear guidance on where preventive counselling and financial support are most needed. Future work will extend the model with real-time student data from learning management systems (LMS) and apply sequential pattern mining to trace the precise path a student follows before disengaging completely from higher education.

REFERENCES

- [1] G.-J. Hwang and Y.-F. Tu, "Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review," *Mathematics*, vol. 9, p. 584, 2021. <https://doi.org/10.3390/math9060584>
- [2] E. Alyahyan and D. Düşteğör, "Predicting academic success in higher education: Literature review and best practices," *International Journal of Educational Technology in Higher Education*, 2021. <https://doi.org/10.1186/s41239-020-0177-7>
- [3] Yağcı, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* 9, 11 (2022).
- [4] Hemchandracharya North Gujarat University (HNGU), Annual Academic Enrollment and Statistical Report, Patan, Gujarat, 2024.
- [5] C. Chaka, "Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: An overview of reviews," *Journal of e-Learning and Knowledge Society*, vol. 18, no. 2, pp. 58–69, 2022.
- [6] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, p. 11, 2022. <https://doi.org/10.1186/s40561-022-00192-z>
- [7] D. Thakur and N. Kapoor, "Predicting student's performance using data mining algorithm," *IEEE*, 2022. <https://doi.org/10.1109/ICACTA54488.2022.9753265>