Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Association rules and classification for recommending among equivalent courses on Indian MOOC platforms: evidence from SWAYAM and NPTEL

Iram Naim[1]

iram.naim@mjpru.ac.in

M.J.P. Rohilkhand University

Mohd Moin Kamran[2]

M.J.P. Rohilkhand University

Research Article

Keywords: MOOC, Course recommendation, Association rule mining, Associative classification, Educational data mining, SWAYAM, NPTEL

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11488802/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Association rules and classification for recommending among equivalent courses on Indian MOOC platforms: evidence from SWAYAM and NPTEL

Iram Naim[1]*, Mohd Moin Kamran[2]

[1] Department of CSIT, MJP Rohilkhand University, Bareilly, 243006, India

*Corresponding author: iram.naim@mjpru.ac.in

Abstract

Learners on India's national MOOC platforms often find one subject offered by several instructors, institutions, and course lengths, with no evidence on how earlier runs of each offering turned out. We assembled an open, run-level dataset from public pages of NPTEL and the SWAYAM course explorer: 18,344 SWAYAM course offerings from ten national coordinators and 10,347 NPTEL course runs (2016–2026) with published enrolment, exam registration, and certification counts. Outcomes were defined relative to each semester cohort, and every predictor was limited to information available before a run opens. Apriori mining produced 232 non-redundant rules linking course attributes and run history to high or low certification yield; 90.3% of the 226 rules that could be tested kept a confidence of at least 0.60 on 2025–2026 runs that were not used for mining. Eight classifiers were compared under grouped cross-validation and a temporal hold-out. A hybrid model that passes rule matches to gradient-boosted trees reached 84.4% accuracy (AUC 0.928) on 2,955 hold-out runs, against 50.1% for a majority baseline, and a rule-list classifier reached 83.2%. When used to rank equivalent courses taught by different instructors, the hybrid model ordered 75.5% of 110 hold-out pairs correctly and put the best course first in 73.6% of 72 choice sets; ranking by popularity did no better than chance. Run history carried most of the signal, and accuracy fell to 70.5% for courses never seen in training. The dataset and code are released for reuse.

Keywords: MOOC; Course recommendation; Association rule mining; Associative classification; Educational data mining; SWAYAM; NPTEL

1 Introduction

India runs its national MOOC programme through SWAYAM, a single portal on which ten national coordinators publish courses, ranging from school subjects prepared by NCERT and NIOS to university and professional courses from IGNOU, CEC, AICTE, IIM Bangalore and others [1, 2]. The largest coordinator is NPTEL, the programme run by the Indian Institutes of Technology and the Indian Institute of Science. Of the 18,344 offerings we found in the SWAYAM explorer, 10,040 (54.7%) were NPTEL runs, and NPTEL's own course pages report 41.2 million enrolments across the runs analysed in this paper. Scale

on this order makes the platforms one of the main routes through which Indian students, working engineers and teachers take credit-bearing online courses.

Completion, however, remains low. Studies of MOOCs elsewhere have reported low completion rates that vary widely with course length and assessment [3, 4], rates that stayed low across six years of MIT and Harvard courses on edX [5], and figures that depend heavily on how completion is defined [6]. For SWAYAM, a survey-based study reported an average completion rate of about 4.6% and found that many learners had not enrolled with the intention of earning a certificate [7]. Although learner intent is partly involved in low completion rates, learner choice has a large role to play.

However, the selection is not as straightforward as it seems. While searching for a subject online through the SWAYAM portal, a learner usually encounters various versions of the course. From the curated list of courses, it can be observed that there were 867 instances where one course was taught by different faculty members and 301 instances where the same course was offered with different durations. In addition to this, 1,916 new titles were obtained virtually identical to existing ones from other instructors. The explorer lists title, instructor, institute, weeks and credits, but nothing about how earlier runs went, and its enrolment field was empty for every course. NPTEL does publish run-wise statistics, but they sit on separate tabs of individual course pages and cannot be compared without collecting them first. Four gaps follow from this situation, and they motivate the present study: there is no standardised, centralised dataset of course outcomes across institutions; very few course recommendation systems have been built for Indian MOOCs; learners cannot tell which of several equivalent courses, taught by different faculty with different content and length, is likely to serve them better; and learners have no information about other students taking the same course.

Research on recommender systems for MOOCs has grown quickly and has been mapped in several recent reviews [8–10]. Much of the strongest work relies on learner-level interaction logs released by platforms such as XuetangX [11, 12]. Indian platforms have not released learner-level data, so a recommender for them has to work from what is public. Course-level outcomes published run after run are a usable substitute: they say, in aggregate, what happened to the people who enrolled. Association rules suit this kind of data because they produce readable statements about which course attributes go together with good or poor outcomes [13], while classifiers turn the same information into a ranked prediction. Combining the two has a long history in data mining [14].

The study pursues four objectives. First is identification and analysis of the data related to the gaps mentioned above, which has been done by creating a run-level dataset using the public information provided on the SWAYAM and NPTEL pages. The second is to develop a machine learning model that finds patterns with association rules and classification. The third is to study how the model behaves on the dataset and to verify

the results experimentally on the most recent, real data, for which we held out all runs from 2025 and 2026. Fourth is validation of the developed model on the accuracy criteria. The paper achieves its objectives by making three contributions:

(1) An open, cleaned dataset that joins the SWAYAM catalogue with NPTEL's run-wise statistics for 2,362 courses and 10,347 runs, with a documented schema, cohort-relative outcome labels and pre-run features.
(2) A set of association rules on course attributes and run history whose stability was checked on later runs, together with a hybrid association-rule and classification model evaluated against seven alternatives and a baseline.
(3) A recommendation procedure for choosing among equivalent courses offered by different instructors, evaluated on hold-out choice sets against random, popularity and history-only baselines.

Section 2 reviews related work. Section 3 describes the data and Section 4 the methods. Section 5 reports the results, Section 6 discusses them together with the study's limitations, and Section 7 concludes.

2 Related work

2.1 Recommender systems for MOOCs and e-learning

Recommender systems in technology-enhanced learning differ from those in retail because the goal is learning rather than consumption, and because what suits a learner depends on prior knowledge and intent [15, 16]. General surveys group the methods into collaborative, content-based, knowledge-based and hybrid families [17, 18]. Content-based methods describe items by their features and recommend items similar to those a user liked [19]; they avoid the need for other users' ratings but say little about quality. Collaborative methods need interaction histories and struggle with new items and new users, the cold-start problem [20].

Khalid and colleagues surveyed MOOC recommender studies published between 2012 and 2019 and found that most of the work implemented recommenders, with courses the most common item recommended, followed by discussion threads, peers, and learning elements [8]. A mapping review by Uddin and colleagues [9] and a later literature review [10] cover the techniques used in more detail. Various studies show that researchers like Zhang et al. have applied hierarchical reinforcement learning (HRL) to enrolment patterns from XuetangX [11], and MOOCCube has provided a significant amount of data on learning courses and concepts at XuetangX to researchers [12]. Garg and Tiwari invented a hybrid MOOC recommender system that uses several machine learning components [21]. There are no examples of any research based on platform-wide outcome data on SWAYAM or NPTEL.

2.2 2.2 Association rules and associative classification in the educational sector

Association rule mining was first created for the field of market-basket research [13], while the Apriori algorithm was applied for the processing of large databases [22]; FP-growth eliminated the necessity of obtaining candidate itemsets [23]. Since the concept of confidence is sometimes able to provide rewards to the rules related to frequently met means, measures of lift can be used to estimate the degree of the importance of the rule produced. Liu, Hsu and Ma managed to prove the efficiency of a well-pruned selection of class association rules as a classifier (CBA) that is normally comparable to decision trees.

Association rules have been employed within the field of education for the purpose of providing feedback to authors in charge of preparing course materials [26], or, when used in combination with clustering, to offer recommendations as to courses to take to students based on previous enrollment data [27]. In addition, it is also worth noting that association rule mining is an important technique among those we encounter in surveys devoted to educational data mining and learning analytics [28]. The mentioned applications operate in the context of a particular institution's learning management system, where student data is available and utilized. This paper aims to apply the methods of this family to public aggregate data covering different institutions.

2.3 Forecasting success and completion in MOOCs

There is considerable literature predicting which MOOC students will graduate from the course. Kizilcec, Piech, and Schneider distinguished the patterns of activity followed by completions and sampler students [29]. Gardner and Brooks produced a paper about the prediction of students' success in MOOCs, featuring the types of characteristics, models, and evaluation used in the research [30]. Feng, Tang and Liu modelled dropout on XuetangX data [31], and Dass, Gary and Cunningham used random forests to predict dropout in a self-paced course [32]. Course-level factors also matter: Jordan found that longer courses had lower completion rates [4], and reviews of why learners join and leave MOOCs point to motivations and difficulties that course design can affect [33]. Most of this work predicts outcomes for individual learners within one or a few courses. Predicting the outcome of a whole course run from information available before it starts, which is what a course recommender needs, has received much less attention.

2.4 Studies of Indian MOOC platforms

Published work on SWAYAM and NPTEL is mostly descriptive or survey-based. Ali, Bhat and Ganaie [1] and Panda, Mahapatra and Saibabu [2] described the structure and coverage of SWAYAM. Singh surveyed learners who dropped out of SWAYAM courses and listed the reasons they gave [7], and Dhamdhere examined the use of additional reading material on the platform [34]. None of these studies used the outcome statistics

that NPTEL publishes for each run, and none built a recommender. Table 1 summarises how the present study responds to each research gap.

Table 1 Research gaps and how the study addresses them

Research gap	What this study does	Section
No standardised, centralised student or course dataset across institutions	Builds an open run-level dataset joining 18,344 SWAYAM offerings from ten coordinators with NPTEL statistics for 10,347 runs from 114 institutes	3, 5.1
Few course recommendation systems for Indian MOOCs	Develops association rules, eight classifiers and a hybrid model on Indian platform data, validated on 2025–2026 runs	4, 5.2, 5.3
Same course offered by different faculty with different content and duration; no way to tell which is better	Identifies equivalent courses by text similarity and ranks them by predicted outcome, with a rule as explanation	3.5, 4.4, 5.4
No information about students studying the same course	Uses published enrolment, registration and certification counts of earlier runs as aggregate peer information; learner-level co-enrolment is not public and is discussed as a limitation	5.1, 6

3 Data

3.1 Sources and collection

Two public sources were used. The first was the NPTEL course portal [35]. Its course catalogue listed 3,484 courses with title, instructor, institute, discipline, and content type. Each course page has a statistics tab that reports, for every run of the course, the run period, the run identifier, the numbers of learners who enrolled, registered for the proctored examination, and were certified, the breakdown of certificates into Gold, Silver, Elite, and completed categories, the number of toppers, and the minimum, maximum, mean, and standard deviation of exam marks. The course page also gives the course abstract and the outline of units and lectures. Of the listed courses, 2,409 had at least one run with published statistics, for 10,456 runs in total. The statistics page of NPTEL provides aggregate data from the previous years [36].

The SWAYAM course explorer was the second source of data in this study. We were able to retrieve all the courses offered by the explorer for future, ongoing, and past courses, which resulted in 18,344 results containing the course name, instructor, college, national coordinator, number of weeks, credits, category, and dates. The explorer also had an enrolment field, but it was left blank for all courses.

Both sources were accessed on 23rd October 2026 via the same public pages visible to learners without logging in and with low requests. No personal or enrolment data were obtained. It was possible to link 9,110 out of 10138 cleaned NPTEL runs (88.0%) to their official number of weeks using the SWAYAM course codes.

3.2 Cleaning and the unit of analysis

The unit of analysis is defined as a course run, which refers to one instance of a course in one semester. The following cases have been removed from the analysis: 91 runs from 2014 and 2015, when the program was still in early development; 13 runs that had fewer than 100 students; and 5 runs with no examination results. The cleaned-up dataset contains 10,347 runs of 2,362 different courses from 114 universities of 26 disciplines. Each run is categorized into a specific cohort based on its semester of conducting: January or July. Thus, 22 different cohorts of runs are created over the period from January 2016 until July 2026. The runs from the years 2016-2024 are classified as the development set (7,392 cases), while the runs from 2025-2026 are the leave-out set, making 2,955 runs of 1,882 courses. The final cohort represents the most up-to-date information, as it incorporates only the runs whose outcomes were published by the date of data collection.

The period of operation indicated by NPTEL goes from enrollment to examination, thus exceeding the teaching time span. In 89.1% of the runs connecting to SWAYAM, a period of two months is equal to a 4-week course, whereas a period of three months corresponds to an 8-week course, and a period of four months equals a 12-week course. Therefore, we have treated the operating period in months as a duration feature for all the runs and official weeks for descriptive comparisons.

3.3 Outcome definitions

Three rates were computed for each run: the registration rate (registered divided by enrolled), the pass rate (certified divided by registered), and the certification yield, defined as

$$\textit{certification yield} = \textit{certified learners} / \textit{enrolled learners} \quad (1)$$

The yield consists of the decision to take the exam and the probability of getting a passing grade. As perceived by a potential learner, it is the portion of individuals who took the course and walked away with a certificate, and therefore it can be used as a legitimate target by a recommender. However, it is not an indicator of any learning process, as we will see in Section 6.

Please note that absolute yields cannot be compared over the years. The pooled yield rose from 5.4% in 2016 to 21.1% in 2025 as the share of enrollees registering for the examination grew (Section 5.1). Each run's yield was therefore converted to a percentile within its semester cohort, and a run was labelled *high yield* if its yield was at or above the cohort median and *low yield* otherwise. The label is balanced by construction (50.1% high in the development set and 50.1% in the hold-out set), so accuracy can be read directly against a 50% baseline.

3.4 Features

Every feature had to be known before a run opens for enrolment, to avoid leaking the outcome into the inputs [38]. The current run's own enrolment, registrations and marks were excluded; enrolment in particular is the denominator of the yield. Table 2 lists the 23 features, grouped into catalogue attributes, run calendar, run history and competition. History features use only earlier runs of the same course: the cohort percentile of the previous run's yield and registration rate, its enrolment and mean mark, the mean percentile of all earlier runs, and the number of semesters since the last run. For a course's first run, these values are missing, which the models handle through missing-value indicators.

Table 2. Features used by the classifiers. All are known before a run opens for enrolment

Group	Feature	Description
Catalogue	Discipline	26 NPTEL disciplines (one-hot; rare levels pooled)
	Institute group	Eight largest IIT/IISc institutes, other IITs, IISERs and research institutes, IIITs/NITs, others
	Coordinating institute	IIT or IISc that coordinates the run
	Units, lectures	Number of units and lectures in the course outline
	Instructor load	Number of NPTEL courses listed under the same instructor
	Multiple instructors	More than one instructor named
	Regional language	Course delivered in an Indian language other than English
	Title type	Title marks the course as introductory, advanced, computing-related, or language and communication skills (four flags)
Calendar	Semester	January or July semester
	Run period	Months from enrolment to examination (2, 3, 4 or 6); a proxy for 4-, 8- and 12-week courses
History	Run number, first run	Number of the run within the course; flag for a first offering
	Previous outcome	Cohort percentile of the previous run’s certification yield and registration rate
	Previous scale and marks	Enrolment (log10) and mean exam mark of the previous run
	Mean prior outcome	Mean cohort percentile of all earlier runs
	Gap	Semesters since the previous run
Competition	Alternatives	Number of equivalent courses by other instructors offered in the same semester

3.5 Identifying equivalent courses

Two courses were treated as equivalent alternatives if their texts were similar and they were taught by different instructors. Each course was represented by two term frequency–inverse document frequency (TF-IDF) vectors [39], one built from the title with word

bigrams and one from the abstract plus lecture titles. The cosine similarities were combined as S = 0.6 S_title + 0.4 S_content. We read samples of pairs in several similarity bands and set the threshold at τ = 0.55; above it, pairs were mostly genuine alternatives such as *Analog Circuits* offered by two IITs, or *Introduction to Operating Systems* and *Operating Systems* from different institutes. Pairs taught by the same instructor were excluded because they were usually sequels (Part I and Part II). With these rules, 298 NPTEL courses that had published runs had at least one equivalent.

4 Methods

Figure 1 shows the overall framework. The data are integrated into a run table and a course table; association rules and classifiers are learned from the development runs; a hybrid model combines the two; and its scores are used to rank equivalent courses. Every step is evaluated on runs that the step did not see.

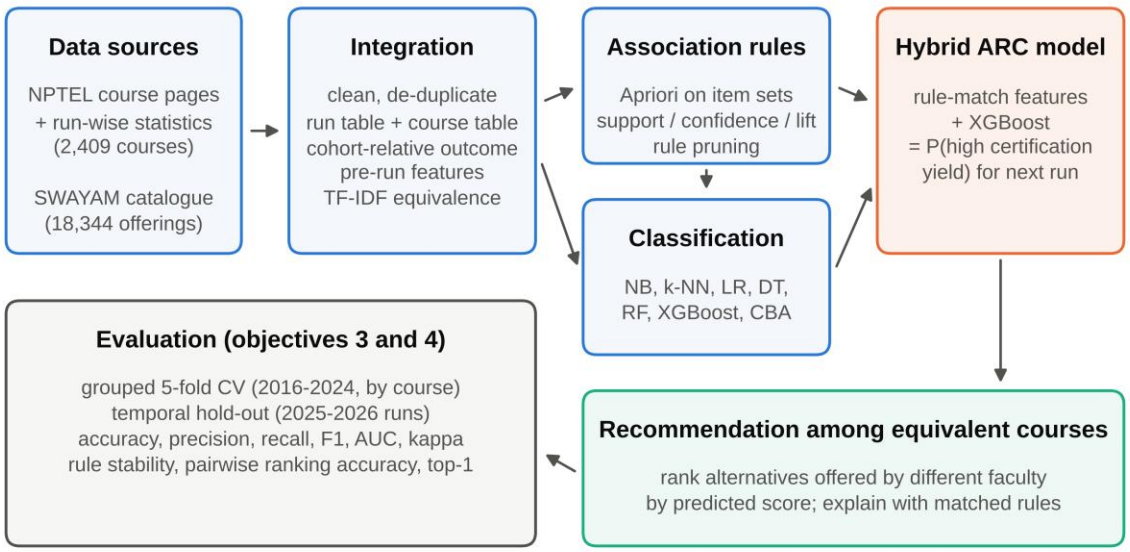


Fig. 1 Framework of the study. Course data from NPTEL and SWAYAM are integrated into run and course tables; association rules and classifiers are learned on 2016–2024 runs and combined in the hybrid ARC model, whose scores rank equivalent courses. Every component is evaluated on runs it did not see

4.1 Association rule mining

Each run was described as a transaction of binary items. The 45 items covered discipline (the twelve largest disciplines), institute group, semester, run period, lecture count (few, medium or many, cut at 40 and 60 lectures, the tertiles of the development set), history (first run, previous run high or low), previous registration rate (top or bottom 40% of its cohort), at least five earlier runs, more than one instructor, regional-language delivery,

and four title types (introductory, advanced, computing, language and communication skills). Two outcome items, HighYield and LowYield, were added.

For a rule X ⇒ Y over N transactions, support, confidence, and lift were computed as

$$support(X \Rightarrow Y) = |\{t : X \cup Y \subseteq t\}| / N \quad (2)$$

$$confidence(X \Rightarrow Y) = support(X \Rightarrow Y) / support(X) \quad (3)$$

$$lift(X \Rightarrow Y) = confidence(X \Rightarrow Y) / support(Y) \quad (4)$$

Frequent itemsets were found with Apriori [22] as implemented in MLxtend [40], using a minimum support of 0.02 (about 150 development runs), antecedents of at most three items, and a minimum confidence of 0.60. Only rules whose consequent was an outcome item were kept, and only those with lift of at least 1.15 [24, 25]. A rule was then dropped if a shorter rule with the same consequent had equal or higher confidence. To test stability, each remaining rule was re-evaluated on the hold-out runs. A rule was *testable* if at least 20 hold-out runs satisfied its antecedent and *stable* if its hold-out confidence was still at least 0.60 with lift above 1.

4.2 Classification models

Eight classifiers were compared with a majority-class baseline: Gaussian naive Bayes; k-nearest neighbours with k = 25; L2-regularised logistic regression; a decision tree with depth at most 6 and at least 20 runs per leaf; a random forest of 500 trees [41]; gradient-boosted trees with XGBoost (400 trees of depth 4, learning rate 0.05, row and column subsampling of 0.8) [42]; an associative classifier in the style of CBA [14]; and the hybrid model described below. Categorical features were one-hot encoded, with rare levels pooled; numeric features were median-imputed with missing-value indicators and, for the distance-based and linear models, standardised. Models were built with scikit-learn [43]. Hyperparameters were set once from common defaults and were not tuned on the hold-out set, since tuning on test data biases error estimates [44].

The CBA-style classifier ranks the mined rules by confidence, then support, then antecedent length, and keeps a rule only if it correctly covers at least one training run not already covered by a higher-ranked rule. A run is labelled by the first rule it satisfies, and runs matched by no rule receive the majority class of the uncovered training runs. Its score for ranking is the confidence of the matching rule, expressed as the probability of high yield.

4.3 Hybrid association-rule and classification model

The hybrid model, called ARC (association-rule-augmented classifier) in what follows, joins the two approaches. In every training dataset, rules are extracted according to Section 4.1, and then 60 rules with the highest lift are stored. For the current run, a binary

indicator shows if the run matches any rule's antecedent. This information about rule matches is added to the input variables of Table 2, and the model is built based on the combined input. Rule extraction matters in the training set since if rules were extracted from the complete dataset, the rule variables could contain information about test runs.

4.4 Recommendations for similar courses

The recommendations process addresses a specific question: having selected the course that the learner is interested in, which of its similar courses should be suggested first in the list? The process retrieves all courses having similarity $S \geq \tau$ and taught by another lecturer. Then every course is evaluated according to the model's predicted probability of its next being successful and ranked accordingly. Each suggestion also has the matching rule with the best lift, which makes it possible to understand the reason for giving the recommendation besides knowing its rank.

4.5 Experimental design and metrics

Two experiments have been conducted. The first one was five-fold cross-validation that was carried out on development data, with folds being grouped according to the courses so that no course would contribute data to both training and test folds. The second one was a time hold-out, meaning that all models were trained on the development runs and tested on 2,955 hold-out runs just once. Of the given runs, 454 belong to 357 courses that did not take part in the development set, thus becoming a cold-start case.

Classification was assessed with accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve (AUC) [45], and Cohen's kappa [46], using a 0.5 threshold. Ninety-five percent confidence intervals for hold-out accuracy and AUC came from 1,000 bootstrap resamples of runs. Feature influence was measured by permutation importance on the hold-out set (20 repeats, AUC as the score).

For the ranking evaluation, we took every pair of equivalent courses in which both courses ran during the hold-out period. The ground truth for a course was the cohort percentile of its first hold-out run, and its score was the model's prediction for that run, made with information available before the run. Three measures were used: pairwise accuracy, the share of pairs in which the better course received the higher score (ties counted as half); the top-1 hit rate in choice sets formed by connected groups of equivalent courses; and normalised discounted cumulative gain (NDCG) [47]. Baselines were random ordering, popularity measured by the previous run's enrolment [48], and the previous run's outcome alone. Pairwise accuracy was compared with 0.5 using an exact binomial test on decided pairs, with bootstrap confidence intervals over pairs.

All analyses used Python 3.13 with pandas 3.0, scikit-learn 1.9, XGBoost 3.4 and MLxtend 0.25. The random seed was fixed at 42.

5 Results

5.1 What the data show

The SWAYAM catalogue held 18,344 offerings of 4,772 distinct titles from 1,390 institutions. Table 3 breaks these down by national coordinator. NPTEL accounted for more than half of all offerings, followed by IGNOU and CEC. Courses of 12 weeks were the most common (50.0% of offerings with a stated duration), followed by 8 weeks (22.1%), and 96.3% of offerings carried credits. Repetition of subjects was common: 3,335 titles appeared more than once, 867 titles (6,190 offerings) had been taught by more than one instructor, 823 by more than one institute, and 143 under more than one coordinator. Among the 1,916 near-duplicate title pairs taught by different instructors, 21.9% differed in duration, and 18.9% came from different coordinators. Outcome information was available only for NPTEL; for the other nine coordinators, the catalogue gave no way to compare courses beyond their descriptions.

Table 3 SWAYAM course offerings by national coordinator, 2016–2026

Coordinator	Offerings	Distinct titles	Institutes	Median weeks	With credits (%)
NPTEL	10,040	2,392	275	12	99.9
IGNOU	2,560	423	101	12	96.3
CEC	2,194	727	603	12	99.5
IIMB	902	273	120	6	94.8
NIOS	634	96	10	24	89.7
NITTTR	618	282	67	8	99.4
INI	469	302	45	12	100.0
AICTE	437	328	150	12	62.9
NCERT	350	39	26	24	17.7
UGC	140	74	63	15	95.7
All	18,344	4,772	1,390	12	96.3

Offerings are course runs listed in the SWAYAM explorer as upcoming, ongoing, or completed on 23 October 2026. The titles and organizations accumulated under each of the coordinators indicate that column totals are greater than the respective overall figures for entries involved in each category

The revised chart of NPTEL indicates that the total number of enrolments is equal to 41.2 million people, the total number of registrations for examinations is 8.85 million, and the number of certifications is 5.67 million; it should also be noted that the number of enrolments is not a direct count of individual learning since one individual can be enrolled in multiple courses. By median yardstick, the number of enrolments constitutes 1,647 packages; the registration yield is 12.6%; the number of passing rates of those whose registrations succeeded is 59.2% and lastly, the certification rate is 6.5% (interquartile range 2.6–13.8%).

Figure 2 shows how the programme changed. The number of runs with published statistics grew from 166 in 2016 to 1,777 in 2025. The pooled registration rate rose from

7.6% to 32.7% over the same years, while the pass rate among registered learners stayed in a band of roughly 56–77%. As a result, the pooled yield roughly quadrupled, from 5.4% to 21.1%. The one sharp break is 2020, when the yield fell to 4.1% and the pass rate to 56.4%, which coincides with the disruption of proctored examinations during the COVID-19 pandemic. This trend is the reason the outcome label is defined within cohorts.

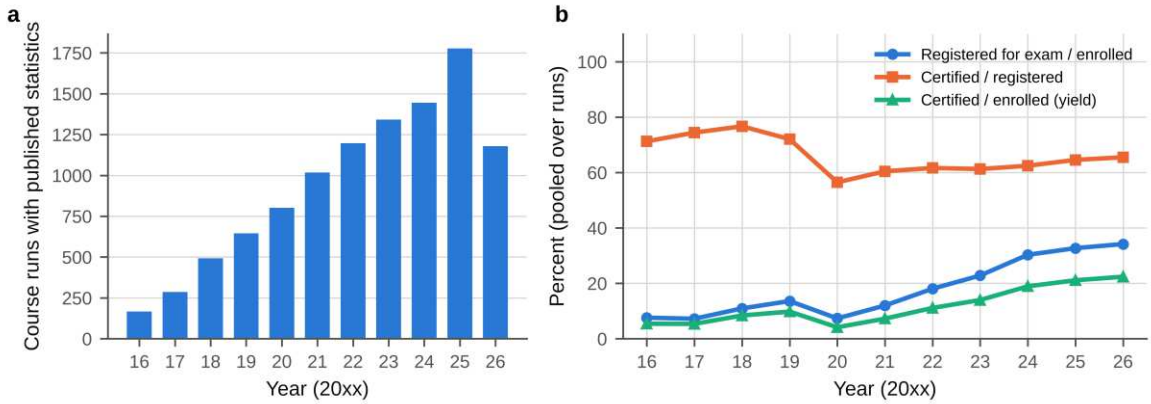


Fig. 2 Growth and outcomes of NPTEL course runs, 2016–2026. Several runs with published statistics per year. b Pooled registration rate, pass rate among registered learners, and certification yield. The 2026 values include only runs whose results had been published by 23 October 2026

Shorter courses converted more enrolees into certificates (Fig. 3a). The median yield was 8.7% for 4-week runs (n = 980), 8.2% for 8-week runs (n = 2,719) and 5.7% for 12-week runs (n = 5,409); the share of runs in the high-yield half fell from 61.4% to 57.4% and 44.3% (Kruskal–Wallis $H = 172.1$, $p < 0.001$). Both components moved in the same direction: 12-week runs had lower median registration (12.5% against 15.4%) and lower median pass rates (53.9% against 62.3%). Discipline mattered at least as much (Fig. 3b). Median yields ranged from 15.8% in multidisciplinary courses and 11.9% in management to 2.7% in physics and 2.6% in mathematics.

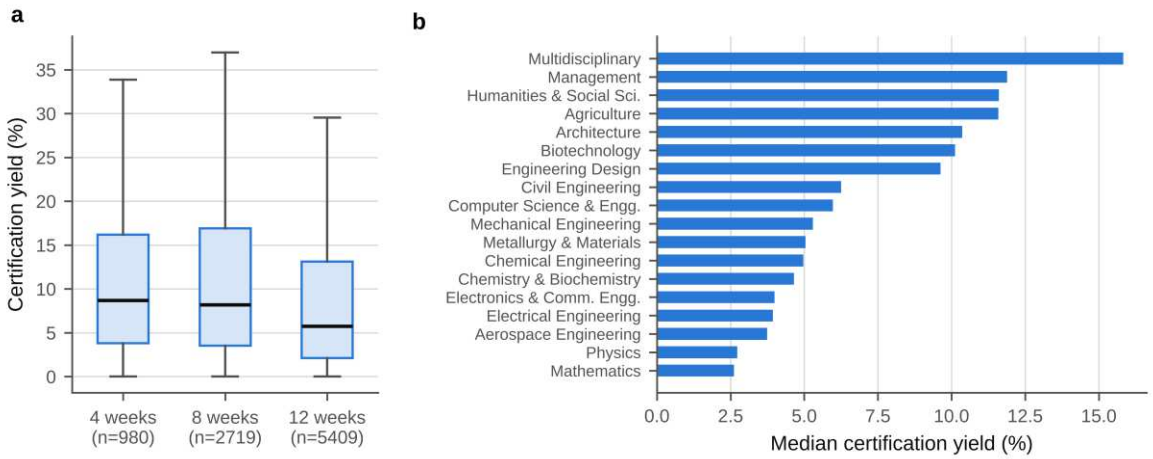


Fig. 3 Certification yield by course length and discipline. a Yield of NPTEL runs linked to their official length in SWAYAM (boxes show quartiles and medians; whiskers extend to 1.5 times the interquartile range; outliers are not drawn). b Median yield by discipline, for the 18 disciplines with at least 100 runs

Two further results bear directly on the recommendation problem. First, outcomes were a stable property of courses. For the 1,893 courses with at least two runs, the intraclass correlation of the cohort percentile was 0.77, so about three-quarters of the variation lay between courses rather than between runs of the same course; and across 7,985 consecutive pairs of runs, the label of the next run matched that of the previous run 83.4% of the time. Second, the choice between equivalent courses had real consequences. For the 198 equivalent pairs with published runs, the median difference in mean cohort percentile was 12.1 points, and in 23.2% of pairs it exceeded 25 points. The alternatives in a pair differed in institute in 74.7% of cases and in run length in 36.9%, and the longer of the two had a median of 1.41 times as many lectures.

5.2 Association rules and their stability on recent runs

Mining the development runs produced 413 rules with an outcome consequent and confidence of at least 0.60. After the lift filter and redundancy pruning, 232 rules remained: 138 predicting high yield and 94 predicting low yield. Table 4 lists a selection.

Table 4 Selected association rules mined on 2016–2024 runs and re-tested on 2025–2026 runs

IF (antecedent)	THEN	Runs 2016–24	Conf. 2016–24	Lift 2016–24	Runs 2025–26	Conf. 2025–26	Lift 2025–26
Discipline: biotechnology; previous run high yield; previous registration rate in top 40%	High	166	0.96	1.91	66	0.94	1.88
Discipline: humanities and social sciences; previous run high yield; at least five earlier runs	High	160	0.96	1.91	145	0.90	1.79
Discipline: humanities and social sciences; previous registration rate in top 40%; July semester	High	196	0.95	1.90	81	0.89	1.78
Discipline: mathematics; previous run low yield; four-month run period (12-week course)	Low	163	0.94	1.89	108	0.95	1.91
Discipline: mathematics; previous registration rate in bottom 40%; four-month run period (12-week course)	Low	163	0.93	1.87	107	0.95	1.91
Previous registration rate in bottom 40%; introductory title; four-month run period (12-week course)	Low	165	0.93	1.86	89	0.93	1.87
Discipline: multidisciplinary	High	236	0.82	1.64	126	0.81	1.62
Discipline: mathematics; four-month run period (12-week course)	Low	292	0.86	1.73	144	0.88	1.75
Title marks course as advanced	Low	199	0.76	1.52	112	0.77	1.54
First run of the course; four-month run period (12-week course)	Low	1,153	0.60	1.21	270	0.51	1.03

Rows 1–6: the three high-yield and three low-yield rules with the highest lift, excluding rules that name a single institute. Rows 7–9: rules using catalogue attributes only. Row 10: a first-run rule that did not hold on later runs. Runs, runs satisfying the antecedent; Conf., confidence

Of the 232 rules, 226 were testable on the hold-out runs; the other 6 involved six-month run periods, which NPTEL used almost only in 2022. Of the testable rules, 204 (90.3%) were stable; that is, they kept a confidence of at least 0.60 with lift above 1 on runs from 2025 and 2026. Stability was 92.6% for high-yield rules and 86.8% for low-yield rules. The median confidence was 0.85 in the development runs and 0.87 in the hold-out runs, the median lift was 1.70 and 1.74, and confidences in the two periods were strongly correlated (r = 0.93; Fig. 4).

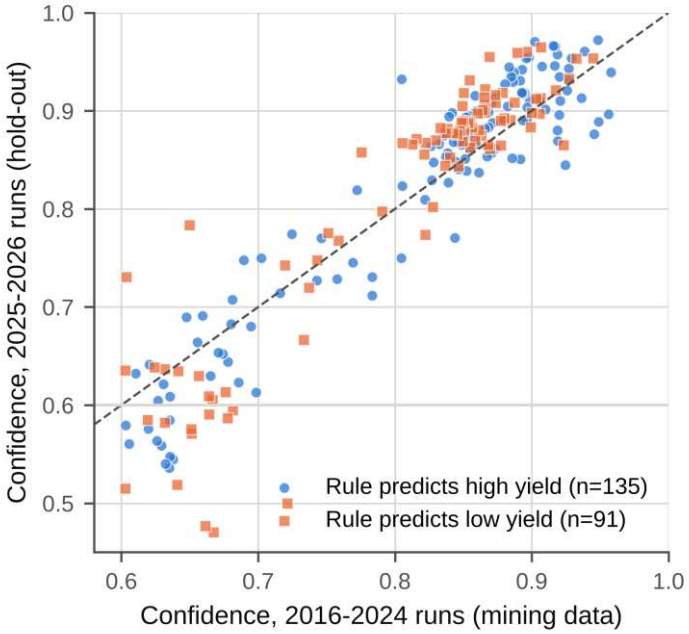


Fig. 4 Stability of association rules. Each point is one of the 226 testable rules mined on 2016–2024 runs; its confidence on those runs is plotted against its confidence on 2025–2026 hold-out runs. The dashed line marks equal confidence, and the horizontal line marks the 0.60 threshold

The strictest rules combine discipline with history. A biotechnology course whose previous run was high yield and whose previous registration rate was in the top 40% of its cohort had a high-yield run in 95.8% of development cases and 93.9% of hold-out cases. On the low side, a mathematics course with a four-month run period and a low previous run was low yield in 94.5% of development runs and 95.4% of hold-out runs. Rules that use only catalogue attributes are weaker but still informative, which matters for courses without history: multidisciplinary courses were high yield with confidence 0.82, mathematics courses with four months were low yield with confidence 0.86, and courses whose title marks them as advanced were low yield with confidence 0.76. Of the 60 testable catalogue-only rules, 85% were stable.

The rules that failed are instructive too. All 6 rules involving a course's first run predicted low yield in the development data, with confidence between 0.60 and 0.72, but none of them held in the hold-out period, where first runs were as likely to be high yield as low. New courses launched in 2025 and 2026 appear to have started on a stronger footing than new courses did in earlier years, and a recommender that relied on those rules would have penalised them unfairly.

5.3 Classification accuracy

Table 5 reports the cross-validated and hold-out performance of all models. Under grouped cross-validation, logistic regression, XGBoost and ARC were the most accurate, at 79.0%, 78.9% and 78.9%, against 48.2% for the majority baseline. All models scored higher on the hold-out set than in cross-validation. The main reason is the mix of runs: 27% of development runs were first offerings, for which no history exists, compared with 12% of hold-out runs.

Table 5 Classification performance under grouped cross-validation (2016–2024) and on the temporal hold-out (2025–2026)

Model	CV accuracy (%)	CV AUC	Hold-out accuracy (%) (95% CI)	Precision	Recall	F1	AUC	Kappa	Accuracy, unseen courses (%)
Majority class (baseline)	48.2 ± 1.5	0.500	50.1 (48.3–51.8)	0.501	1.000	0.667	0.500	0.000	50.0
Naive Bayes	72.9 ± 1.5	0.788	77.3 (75.8–78.8)	0.766	0.787	0.777	0.824	0.547	67.8
k-nearest neighbours (k = 25)	76.6 ± 1.1	0.854	82.5 (81.1–83.8)	0.844	0.799	0.820	0.913	0.650	64.8
Logistic regression	79.0 ± 1.6	0.875	84.0 (82.6–85.3)	0.850	0.825	0.837	0.922	0.679	69.8
Decision tree (depth 6)	76.9 ± 0.8	0.849	83.5 (82.1–84.7)	0.886	0.769	0.823	0.913	0.669	67.0
Random forest	78.4 ± 1.0	0.878	84.4 (83.0–85.6)	0.870	0.809	0.838	0.929	0.687	70.0
XGBoost	78.9 ± 0.9	0.877	84.5 (83.1–85.9)	0.871	0.810	0.839	0.928	0.689	70.9
Associative classifier (CBA)	76.7 ± 1.7	0.832	83.2 (81.8–84.6)	0.832	0.832	0.832	0.886	0.664	69.8
Hybrid ARC (rules + XGBoost)	**78.9 ± 1.1**	**0.879**	**84.4 (83.0–85.7)**	**0.868**	**0.811**	**0.839**	**0.928**	**0.687**	**70.5**

CV values are means ± standard deviations over five folds grouped by course. Hold-out values are for 2,955 runs; confidence intervals are from 1,000 bootstrap resamples. Unseen courses are the 454 hold-out runs of 357 courses absent from the development set. The positive class is a high-yield run

On the hold-out set, ARC reached an accuracy of 84.4% (95% CI 83.0–85.7%), an F1 score of 0.839 and an AUC of 0.928 (95% CI 0.919–0.936). XGBoost alone (84.5%, AUC 0.928) and the random forest (84.4%, AUC 0.929) performed the same within the confidence intervals, and logistic regression was close behind (84.0%, AUC 0.922). Adding rule features therefore did not raise accuracy, but it did not cost any either, and it gives each prediction a rule-based explanation. The CBA-style rule list, which uses only 185 rules and can be printed on a few pages, reached 83.2% accuracy with an AUC of 0.886; its lower AUC reflects the coarse scores a rule list produces. Naive Bayes was the weakest model (77.3%). The ARC confusion matrix contained 1,200 true positives, 1,293 true negatives, 183 false positives, and 279 false negatives, so its errors leaned slightly towards missing high-yield runs. Accuracy was similar in each of the four hold-out cohorts, between 83.4% and 85.1% (Fig. 5).

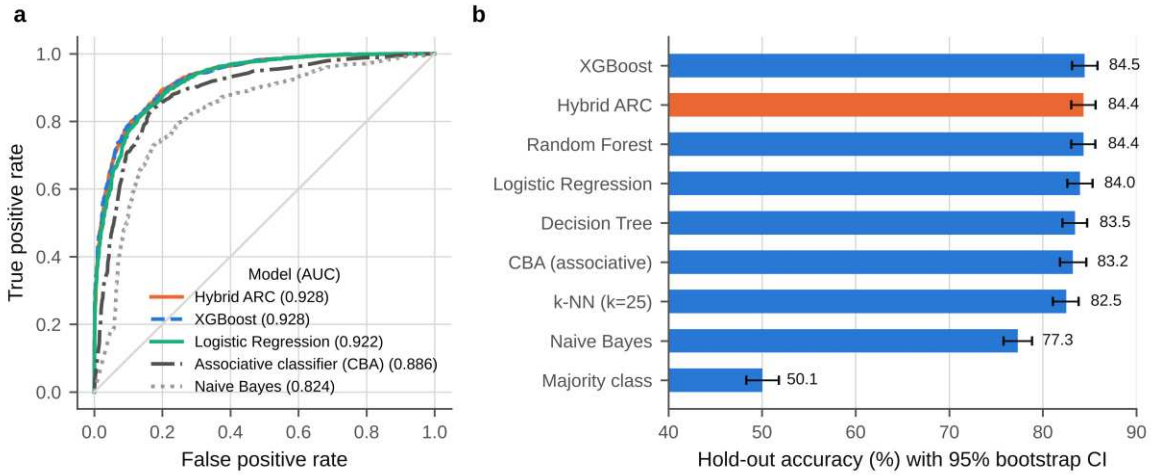


Fig. 5 Classification performance on the 2025–2026 hold-out runs. a Receiver operating characteristic curves for five models, with AUC in parentheses. b Accuracy of all models with 95% bootstrap confidence intervals

For the 454 hold-out runs of courses never seen in development, ARC accuracy fell to 70.5% (AUC 0.792). This is well above chance, but it shows clearly that the model's strength comes from history.

5.4 Ranking equivalent courses

The hold-out period contained 110 pairs of equivalent courses that both ran, grouped into 72 choice sets (58 of them pairs, the rest sets of three to six courses). Table 6 and Fig. 6 summarise the ranking results.

Table 6 Ranking equivalent courses offered by different instructors on hold-out runs

Scoring method	Pairs ordered correctly (%) (95% CI)	p (binomial)	Best course ranked first (%)	NDCG
Random ordering	50.0		45.8	0.916
Popularity (previous-run enrolment)	50.0 (41.8–58.2)	1.00	47.2	0.923
Previous-run outcome only	70.0 (62.3–76.8)	< 0.001	66.7	0.956
Associative classifier (CBA)	58.2 (49.5–66.4)	0.08	54.2	0.926
Logistic regression	74.5 (66.4–82.7)	< 0.001	68.1	0.972
XGBoost	73.6 (65.5–81.8)	< 0.001	69.4	0.974
Hybrid ARC	**75.5 (67.3–82.7)**	**< 0.001**	**73.6**	**0.971**

In the study, 110 pairs of equivalent courses were evaluated from 72 choice sets, while their cohort ranks were determined during their first hold-out run. In the analysis, ties and missing scores were assigned half weight in pairs counted as correct; p-values were used to determine the significance of students ranking pairs against a 0.5 expectation. Thus, the values in random ordering can be interpreted as expectations

ARC ordered 75.5% of pairs correctly (95% CI 67.3–82.7%; exact binomial p < 0.001) and placed the best course first in 73.6% of choice sets, against 45.8% expected by chance. Logistic regression (74.5%) and XGBoost (73.6%) were close. Ranking by the previous run's outcome alone ordered 70.0% of pairs correctly but could not rank 26 pairs in which a course had no earlier run. Popularity was no better than chance: courses with more enrolments in their previous run were as likely to have the lower yield as the higher (50.0%). The CBA rule list ordered 58.2% of pairs correctly (p = 0.08), held back by ties between runs matched by the same rule. When the ground truth was taken as the mean over all of a course's hold-out runs rather than its first, ARC's pairwise accuracy was 74.5%.

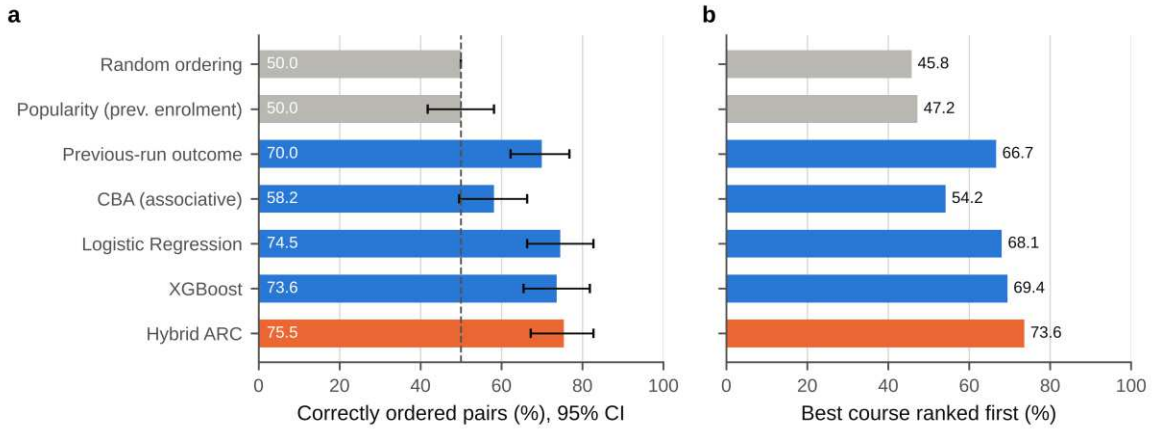


Fig. 6 Ranking equivalent courses on hold-out runs. a Share of 110 equivalent-course pairs ordered correctly, with 95% bootstrap confidence intervals; the dashed line marks chance. b Share of 72 choice sets in which the course with the highest observed yield was ranked first

Two choice sets illustrate what the ranking does. *Software Testing* was offered by IIIT Bangalore as a 12-week course and by IIT Kharagpur as a 4-week course. ARC scored the first at 0.96 and the second at 0.45, citing the rule that a course with a high previous yield, high previous registration, and at least five earlier runs is high yield (confidence 0.90). The hold-out yields were 38.9% and 12.4%, so the ranking was right even though it went against the general tendency of shorter courses to do better, and even though the IIT Kharagpur course had the larger enrolment. *Theory of Computation* was offered by IIT Kanpur as an 8-week course and by IIT Hyderabad as a 12-week course; ARC scored them at 0.64 and 0.04, and the observed yields were 11.8% and 3.9%.

5.5 What drives the predictions

Permutation importance (Fig. 7) confirms the role of history. Shuffling the mean percentile of all earlier runs lowered hold-out AUC by 0.042, and shuffling the previous run's percentile lowered it by 0.036; discipline came next (0.014), followed by the previous run's mean mark (0.007). History has already borne witness to the minor contribution of the institute grouping, number of lectures, type of title, and number of tutors. An ablation supports the same conclusion. With catalogue and calendar features only, XGBoost reached 74.2% accuracy (AUC 0.818); with history features only, 82.2% (AUC 0.920); and with both, 84.5% (AUC 0.928).

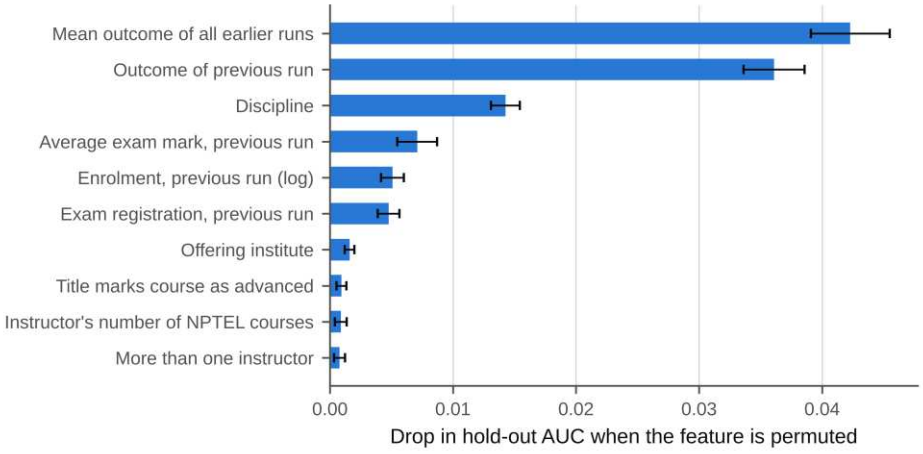


Fig. 7 Permutation importance of the ten main features of XGBoost, measured as the average decline in the AUC, is presented in the hold-out runs over 20 permutations (standard deviations are indicated with the error bars)

6 Discussion

The research aimed to accommodate four missing components, and the findings explain each of them. Regarding the first, the absence of an organized and centralized database, the unified data provided by the SWAYAM-NPTEL table demonstrates how it could be created using public data available, as well as the obstacles faced. NPTEL supplies

enough information for comparing achievements across institutions and years; the other nine coordinators share only information about catalogs, not results, while there is no data in the enrollment column of the SWAYAM explorer. If a uniform outcome schema were developed among coordinators, including the same fields that NPTEL provides, an analysis like this one could be performed in all other institutions.

In terms of what was demonstrated by the data for the second and third gaps, a recommender based on course-level outcomes can indeed work effectively. The outcomes of previous iterations are stable; about 75% of the variance attributable to cohort relative yield is due to different courses, and consecutive courses are labelled in the same way in more than 80% of cases. Thus, because of the stability of the outcomes, it is possible to make predictions, and that explains the domination of historical features in course outcome modeling. It also implies that course selection is not unimportant; the median difference in outcomes for similar courses taught by different people was 12 percentiles, while almost 25% of pairs had differences greater than 25 percentiles. A learner choosing between them on the catalogue description alone has little to go on, and our results suggest that the obvious shortcut, choosing the more popular course, is no better than a coin toss.

The association rules add something the classifiers do not: statements that a learner, an instructor or a coordinator can read and check. Several agree with earlier findings. The lower yield of longer courses matches Jordan's observation that completion falls with course length [4], and the gap between disciplines suggests that mathematics and physics courses may ask more of learners than courses in management or the humanities. The rules also proved useful as a check on drift. The first-run rules, mined from nine years of development data, failed on the latest runs. Without the hold-out test, we would have built a recommender that discouraged learners from new courses for reasons that no longer applied. Re-mining and re-testing rules each semester is cheap and should be part of any deployment.

The hybrid ARC model matched the best pure classifiers in accuracy and had the highest pairwise accuracy and top-1 hit rate when ranking equivalent courses, although XGBoost had a marginally higher NDCG and the differences among ARC, XGBoost and logistic regression were within their confidence intervals. Its practical advantage is that each recommendation arrives with a rule that explains it. A rule-list classifier alone was only about one point less accurate on the hold-out set, which makes it a reasonable choice where transparency matters more than the last point of accuracy, although it ranked equivalent courses less well.

The fourth gap, the lack of information about other students taking the same course, is the one the public data address least. Enrolment and registration counts show how many people took a course, but not who they were or what else they took, so collaborative

filtering of the usual kind is impossible. If the platforms released anonymised co-enrolment counts, for example, how many learners of course A also took course B, association rules could be mined on learner baskets in the way they were originally designed for [13, 27], and the course-level model presented here could serve as a prior for courses with few learners.

6.1 Limitations

Several limitations should be kept in mind. The outcomes are aggregates for whole runs, so they describe what happened to a course's enrollees as a group and cannot be read as predictions for any individual learner. Enrolment refers to learners who acquire the course just to browse it without any intention of taking the exam [7]. Thus, low yield might reveal somebody's interest in the course rather than its quality. The certification yield reflects the measure of completion rather than that of learning. The label is relative and therefore attaches to the low-yield courses and does not show the number of hours. The course outlines and abstracts were taken from the current course pages and used for extending earlier runs. Although the presentation of the course may have changed from one run to another. The similarity threshold was set through sampling, and if there were different thresholds, it would lead to the collection of data and to the analysis. Results, however, are obtained for NPTEL only, and thus courses cannot be assessed from the data collection of other coordinators. Finally, the data were collected from web pages whose structure may change; the released files preserve the state of the sources on 23 October 2026.

7 Conclusion

Public pages of SWAYAM and NPTEL contain enough information to build a useful course recommender, once the scattered run-wise statistics are collected and joined to the catalogue. On 10,347 NPTEL runs, association rules on course attributes and history were readable and mostly stable on later runs; a hybrid rule-and-boosting model predicted whether a run would be in the high-yield half of its cohort with 84.4% accuracy on 2025–2026 data, and the same model ranked equivalent courses offered by different instructors correctly about three times in four. Course history carried most of the signal, which limits accuracy for new courses and shows how much value the platforms could add by publishing outcome data for every coordinator in a common format. Future work should extend the outcome data beyond NPTEL, add learner-level or co-enrolment information if it becomes available, and test the recommender with learners.

Declarations

Acknowledgments

The authors thank NPTEL and SWAYAM for making course statistics and catalogue information publicly available.

Funding

No funding was received for conducting this study.

Author contributions

Iram Naim and Mohd Moin Kamran made contributions to the conception and design of the study, data collection, analysis, and writing of the manuscript. Both the authors read and approved the final copy.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable. The study used only aggregate, course-level statistics and catalogue information that NPTEL and SWAYAM publish openly. It involved no human participants and no individual-learner data. Instructor names, which appear in the public catalogues, are replaced by pseudonymous identifiers in the released dataset, and the paper does not evaluate individual instructors.

Consent for publication

Not applicable.

Data availability

The dataset obtained and used in this study includes the cleaned NPTEL run table, the course table, the SWAYAM catalogue extract, the equivalent course pairs, and the data dictionary made available on Zenodo [49]. The data was obtained from the public site of NPTEL and SWAYAM [35, 37] on October 23, 2026.

Code availability

The Python code used for data cleaning, rule mining, model training, evaluation, and the figures is included in the same deposit [49].

Use of generative AI tools

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with data collection from public NPTEL and SWAYAM web pages, data analysis code, and drafting of the manuscript text. The authors reviewed and edited all content and take full responsibility for the content of the publication.

References

1. Ali A, Bhat MA, Ganaie SA. India on the cyber learning platform: an insight about SWAYAM MOOC platform. KIIT J Libr Inf Manag. 2020;7(1). https://doi.org/10.5958/2455-8060.2020.00003.8
2. Panda SP, Mahapatra RK, Saibabu K. A study of Webs of Active Learning for Young Aspiring Minds (SWAYAM): a MOOC platform of India. KIIT J Libr Inf Manag. 2019;6(1). https://doi.org/10.5958/2455-8060.2019.00001.6
3. Jordan K. Initial trends in enrolment and completion of massive open online courses. Int Rev Res Open Distrib Learn. 2014;15(1). https://doi.org/10.19173/irrodl.v15i1.1651
4. Jordan K. Massive open online course completion rates revisited: assessment, length and attrition. Int Rev Res Open Distrib Learn. 2015;16(3). https://doi.org/10.19173/irrodl.v16i3.2112
5. Reich J, Ruipérez-Valiente JA. The MOOC pivot. Science. 2019;363(6423):130–1. https://doi.org/10.1126/science.aav7958
6. Celik B, Cagiltay K. Uncovering MOOC completion: a comparative study of completion rates from different perspectives. Open Praxis. 2024;16(3):445–56. https://doi.org/10.55982/openpraxis.16.3.606
7. Singh G. Why do MOOCs fail on completion rate? An analysis of SWAYAM courses. Indian J Educ Technol. 2022;4(1):111–23. https://journals.ncert.gov.in/IJET/article/view/529
8. Khalid A, Lundqvist K, Yates A. Recommender systems for MOOCs: a systematic literature survey (January 1, 2012 – July 12, 2019). Int Rev Res Open Distrib Learn. 2020;21(4):255–91. https://doi.org/10.19173/irrodl.v21i4.4643
9. Uddin I, Imran AS, Muhammad K, Fayyaz N, Sajjad M. A systematic mapping review on MOOC recommender systems. IEEE Access. 2021;9:118379–405. https://doi.org/10.1109/ACCESS.2021.3101039
10. Najmani K, Benlahmar EH, Sael N, Zellou A. A systematic literature review on recommender systems for MOOCs. Ingénierie des Systèmes d'Information. 2022;27(6):895–902. https://doi.org/10.18280/isi.270605
11. Zhang J, Hao B, Chen B, Li C, Chen H, Sun J. Hierarchical reinforcement learning for course recommendation in MOOCs. Proc AAAI Conf Artif Intell. 2019;33(1):435–42. https://doi.org/10.1609/aaai.v33i01.3301435
12. Yu J, Luo G, Xiao T, Zhong Q, Wang Y, Feng W, et al. MOOCCube: a large-scale data repository for NLP applications in MOOCs. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. p. 3135–42. https://doi.org/10.18653/v1/2020.acl-main.285
13. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. ACM; 1993. p. 207–16. https://doi.org/10.1145/170035.170072

14. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98). AAAI Press; 1998. p. 80–6.
15. Manouselis N, Drachsler H, Vuorikari R, Hummel H, Koper R. Recommender systems in technology enhanced learning. In: Ricci F, Rokach L, Shapira B, Kantor PB, editors. Recommender systems handbook. Boston: Springer; 2011. p. 387–415. https://doi.org/10.1007/978-0-387-85820-3_12
16. Drachsler H, Verbert K, Santos OC, Manouselis N. Panorama of recommender systems to support learning. In: Ricci F, Rokach L, Shapira B, editors. Recommender systems handbook. 2nd ed. Boston: Springer; 2015. p. 421–51. https://doi.org/10.1007/978-1-4899-7637-6_12
17. Ricci F, Rokach L, Shapira B. Recommender systems: introduction and challenges. In: Ricci F, Rokach L, Shapira B, editors. Recommender systems handbook. 2nd ed. Boston: Springer; 2015. p. 1–34. https://doi.org/10.1007/978-1-4899-7637-6_1
18. Bobadilla J, Ortega F, Hernando A, Gutiérrez A. Recommender systems survey. Knowl Based Syst. 2013;46:109–32. https://doi.org/10.1016/j.knosys.2013.03.012
19. Lops P, de Gemmis M, Semeraro G. Content-based recommender systems: state of the art and trends. In: Ricci F, Rokach L, Shapira B, Kantor PB, editors. Recommender systems handbook. Boston: Springer; 2011. p. 73–105. https://doi.org/10.1007/978-0-387-85820-3_3
20. Schein AI, Popescul A, Ungar LH, Pennock DM. Methods and metrics for cold-start recommendations. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2002. p. 253–60. https://doi.org/10.1145/564376.564421
21. Garg V, Tiwari R. Hybrid massive open online course (MOOC) recommendation system using machine learning. In: International Conference on Recent Trends in Engineering, Science & Technology (ICRTEST 2016). IET; 2016. https://doi.org/10.1049/cp.2016.1479
22. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases. Morgan Kaufmann; 1994. p. 487–99.
23. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. ACM; 2000. p. 1–12. https://doi.org/10.1145/342009.335372
24. Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data. ACM; 1997. p. 255–64. https://doi.org/10.1145/253260.253325
25. Tan PN, Kumar V, Srivastava J. Selecting the right objective measure for association analysis. Inf Syst. 2004;29(4):293–313. https://doi.org/10.1016/S0306-4379(03)00072-3

26. García E, Romero C, Ventura S, de Castro C. An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. User Model User-Adapt Interact. 2009;19(1–2):99–132. https://doi.org/10.1007/s11257-008-9047-z
27. Aher SB, Lobo LMRJ. Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. Knowl Based Syst. 2013;51:1–14. https://doi.org/10.1016/j.knosys.2013.04.015
28. Romero C, Ventura S. Educational data mining and learning analytics: an updated survey. WIREs Data Min Knowl Discov. 2020;10(3):e1355. https://doi.org/10.1002/widm.1355
29. Kizilcec RF, Piech C, Schneider E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. ACM; 2013. p. 170–9. https://doi.org/10.1145/2460296.2460330
30. Gardner J, Brooks C. Student success prediction in MOOCs. User Model User-Adapt Interact. 2018;28(2):127–203. https://doi.org/10.1007/s11257-018-9203-z
31. Feng W, Tang J, Liu TX. Understanding dropouts in MOOCs. Proc AAAI Conf Artif Intell. 2019;33(1):517–24. https://doi.org/10.1609/aaai.v33i01.3301517
32. Dass S, Gary K, Cunningham J. Predicting student dropout in self-paced MOOC course using random forest model. Information. 2021;12(11):476. https://doi.org/10.3390/info12110476
33. Hew KF, Cheung WS. Students' and instructors' use of massive open online courses (MOOCs): motivations and challenges. Educ Res Rev. 2014;12:45–58. https://doi.org/10.1016/j.edurev.2014.05.001
34. Dhamdhere S. Study of use of an additional reading material in SWAYAM MOOC platform in India and need of learner centric personalized adaptive real-time reading material in MOOCs. J Educ Online. 2024;21(1). https://doi.org/10.9743/jeo.2024.21.1.4
35. National Programme on Technology Enhanced Learning (NPTEL). Courses and course statistics. https://nptel.ac.in/courses. Accessed 23 Oct 2026.
36. National Programme on Technology Enhanced Learning (NPTEL). Statistics. https://archive.nptel.ac.in/resources/statistics/. Accessed 23 Oct 2026.
37. SWAYAM, Ministry of Education, Government of India. Course explorer. https://swayam.gov.in/explorer. Accessed 23 Oct 2026.
38. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data. 2012;6(4):15. https://doi.org/10.1145/2382577.2382579
39. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag. 1988;24(5):513–23. https://doi.org/10.1016/0306-4573(88)90021-0

40. Raschka S. MLxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack. J Open Source Softw. 2018;3(24):638. https://doi.org/10.21105/joss.00638
41. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324
42. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
44. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. https://doi.org/10.1186/1471-2105-7-91
45. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74. https://doi.org/10.1016/j.patrec.2005.10.010
46. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104
47. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst. 2002;20(4):422–46. https://doi.org/10.1145/582415.582418
48. Cremonesi P, Koren Y, Turrin R. Performance of recommender algorithms on top-N recommendation tasks. In: Proceedings of the Fourth ACM Conference on Recommender Systems. ACM; 2010. p. 39–46. https://doi.org/10.1145/1864708.1864721
49. Iram Naim*, Mohd Moin Kamran: Course-run outcomes and catalogue data from NPTEL and SWAYAM, 2016–2026 [dataset]. Zenodo; 2026. https://doi.org/10.5281/zenodo.23594146