Formal Representation of STEM Academic Trajectories Using a Deterministic Finite Automaton: A Gender Based Analysis of Academic Progression

Melissa Moreno Novoa, Edwin Puertas and Juan Carlos Martinez-Santos
Universidad tecnologica de Bolivar, Cartagena, Colombia
memoreno@utb.edu.co, epuerta@utb.edu.co and jcmartinezs@utb.edu.co

Abstract – This study proposes a deterministic finite automaton to represent academic progression through institutionally defined states and rules. A total of 83,419 observations from 9,972 students enrolled in 18 STEM programs at the Universidad Tecnologica de Bolivar between 2016 and 2025 were analysed. The derived variables enabled the identification of academic stability, risk, recovery, relapse, and pathways, complemented by first-order Markov transition matrices. Regular academic standing accounted for 55.71% of the observations and showed a persistence probability of 0.826. Women exhibited lower rates of entry into and progression toward more severe academic risk. The largest difference was observed in the transition from PAP to PAT, which was 8.1 percentage points higher among men. Effect sizes were small. This approach revealed variations in academic progression that are not captured by final-outcome indicators.

Keywords—Academic trajectories, deterministic finite automaton, STEM education, student progression, gender, academic risk, Markov chains.

I. INTRODUCTION

In Colombia, strengthening educational trajectories is a public policy priority. The National Government recognizes that ensuring complete educational trajectories is essential to expand young people's development opportunities, reduce persistence gaps in higher education, and support the development of comprehensive life projects [1][2].

STEM programs pose challenges because of their academic demands and the persistence difficulties experienced by part of the student population. Understanding how students move through academic conditions, rather than examining only their final outcomes, is therefore relevant for designing institutional support strategies[3], [4],[5] .

Accordingly, higher education institutions have adopted educational analytics to understand student behavior, improve retention, and support decision-making aimed at strengthening academic trajectories [6]. Longitudinal approaches also make it possible to examine academic progression and identify trajectory patterns that can inform strategies for student persistence and academic success [7] .

Although recent studies have made significant contributions to the analysis of academic trajectories through Learning Analytics and Process Mining techniques [8], [9], standardized methodologies for formally representing academic progression remain limited. This restricts the comparison of trajectories across students and institutions [10]. This gap motivates the development of structured models that explicitly describe the states and transitions involved in academic pathways.

This study proposes a formal representation of academic trajectories in STEM programs through a deterministic finite automaton based on institutional regulatory states and explicit transition rules. The contribution lies in transforming longitudinal student records into structured trajectories from which indicators of stability, risk, recovery, relapse, and progression can be derived. These indicators are used to characterize the observed trajectories and examine differences between women and men.

II. RELATED WORK

Educational data analysis has evolved considerably to provide a better understanding of student behavior and support decision-making in higher education. Learning Analytics and Educational Data Mining have emerged as two key approaches for extracting knowledge from academic records. These approaches focus on interpreting data to improve educational processes and on developing analytical methods and computational models [11].

These strategies have led to the development of early warning systems capable of identifying students at risk of dropout using academic, demographic, and relational data, thereby enabling timely interventions to improve student retention [12][13].

The availability of longitudinal data has encouraged the application of techniques for analyzing processes and sequences. Process mining facilitates the discovery of temporal patterns, the identification of frequent pathways, and the analysis of how student behavior evolves based on event logs. Sequence analysis, in turn, makes it possible to characterize dynamic educational processes by examining the order and repetition of student activities[14].

Although these approaches provide a temporal perspective on learning, their use has been largely limited to events recorded on educational platforms and behaviors observed within specific courses. At the same time, several studies have examined academic trajectories from a longitudinal perspective. Some indicate that university progression is not a linear path, as it may involve program changes, delays, and varying times to graduation[15].

Other studies represent academic trajectories through temporal graphs in which nodes correspond to courses and the relationships among them preserve the chronological sequence of the curriculum, thereby improving academic performance prediction[16]. These studies highlight the importance of capturing the temporal dimension of academic trajectories, although their models are constructed from courses, events, or observed variables.

Another line of research represents student progression through probabilistic models. Markov chains estimate the probabilities of persistence, dropout, and graduation [17], whereas Hidden Markov Models use latent states to identify curricular patterns[18]. In both cases, the states are defined according to probabilistic criteria rather than explicit institutional academic rules.

The literature addresses academic trajectories through predictive models, process mining, sequence analysis, graphs, and probabilistic models [16][19] [20] However, the formal representation of academic progression based on regulatory academic states and explicit transition rules remains underexplored. To address this limitation, this study proposes a deterministic finite automaton (DFA) for modeling academic trajectories and deriving structural variables to compare progression patterns between women and men enrolled in STEM programs.

A. Data Preparation

The data were provided by Universidad Tecnologica de Bolivar following institutional authorization and submission of a request to the Ethics Committee. The data set consisted of anonymized academic records of undergraduate students from 2016 to 2025. Each student was represented by a coded identifier, which enabled the reconstruction of their academic progression without using personally identifiable information. The unit of observation was the student-period record, while the unit of analysis was each student's complete academic trajectory.

The initial dataset contained information on 12,000 students. To define the analytical population, students enrolled in 18 programs considered within the STEM scope adopted in this study were selected. After applying the inclusion and exclusion criteria, 9,972 students were retained. The programs analyzed covered engineering, architecture, data science, economics, finance, and other related disciplines included in the classification used in this study. Selection was based on the academic program institutionally recorded for each student.

The dataset contained temporal, academic, and institutional information. However, the reconstruction of academic trajectories primarily relied on the student identifier, admission and observation periods, academic program, institutional

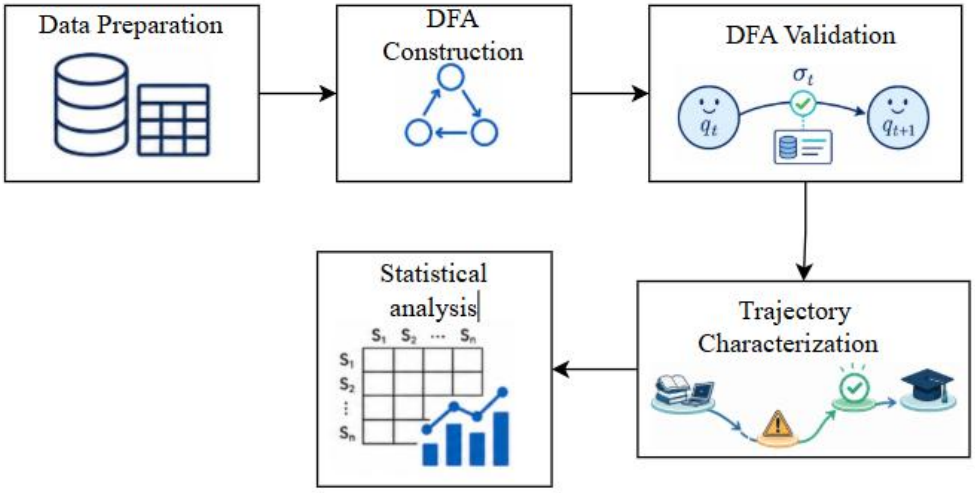


Fig. 1 Methodology.

III. METHODOLOGY

The methodology was organized into five consecutive phases. The process began with the preparation of longitudinal academic records and continued with the formalization of academic trajectories using a DFA based on institutional states and rules. Subsequently, the consistency of the generated sequences was validated, structural variables were derived to characterize student progression, and statistical tests were applied to describe and compare the trajectories of women and men enrolled in STEM programs. Fig. 1 summarizes this methodological pipeline.

academic standing, grade point average, course grades, courses enrolled in and passed, and the number of credits attempted, earned, and required by each program.

The variables gender, age, marital status, study modality, admission test scores, and scholarship status were retained to characterize the study population and conduct subsequent comparative analyses. Regular academic terms were considered to represent semester by semester progression.

The data-leaning process included the standardization of student identifiers, program names, gender categories, and academic term codes. The records were then arranged in

ascending order by student and academic term. Thus, the academic history of student $i$was represented as shown in (1):

$$\mathcal{T}_i = \left(r_{i1}, r_{i2}, \ldots, r_{iT_i}\right) \qquad (1)$$

where r_{it}denotes the academic record of student $i$during term t, and $T_i$represents the total number of terms observed for that student.

B. Automaton Formalization and Determinization

In the second phase, the institutional logic governing academic progression was formalized. Initially, this logic was represented using a NFA, since some conditions could be associated with more than one institutional state. Subsequently, the subset construction procedure was applied to obtain an equivalent DFA. This transformation ensured that, for each state and input symbol, there was only one possible transition, a necessary condition for reconstructing student trajectories.

TABLE I

CORRESPONDENCE OF STATES

DFA Symbol	NFA Symbol	NFA ID	States
q0	qunit, q0, q4	5, 0, 4	Init, Enrolled Applicant, External Transfer
q1	q0	0	Enrolled Applicant
q2	q1	1	First-Time Student in a Degree Program
q3	q6	11	PAP
q4	q7	7	PFU
q5	q2	2	Regular Continuing Student
q6	q9	9	Internal Transfer
q7	q12	12	PAT
q8	q5	15	Restart
q9	q8	8	Readmission
q10	q3	3	Graduation
q11	q12	13	Academic Recovery
q12	q9, q2	9, 2	Internal Transfer, Regular Continuing Student
q13	q10	14	Academic Exclusion

The deterministic automaton was formally defined as:

$$M = (Q, \Sigma, \delta, q_0, F) \qquad (2)$$

where $Q$is the finite set of academic and administrative states; $\Sigma$is the alphabet of institutional events and conditions; $\delta$is the transition function; $q_0$is the initial state; and Fcontains the terminal states established for the analysis. The determinization process produced a set of 14 states, labeled $q_0$through q_{13}.

Table I presents the correspondence between the NFA states and the resulting DFA states. Most deterministic states retained direct correspondence with a single institutional status. In contrast, $q_0$and q_{12}resulted from grouping multiple NFA states.

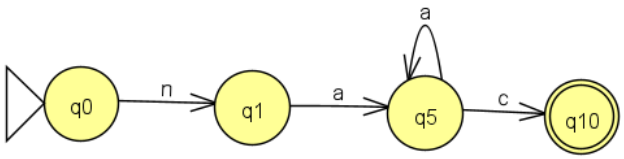


Fig. 2 Ideal trajectory: Where the transition function is fulfilled δ= naaaaaaaaac.

Specifically, q_0 combined the initial configuration, enrolled applicant status, and external transfer status, whereas q_{12}represented the combination of internal transfer and regular continuing status. These composite states retained the information required to resolve nondeterminism without altering the institutional logic represented by the automaton.

The resulting states represent different stages of the academic trajectory. Enrolled applicants, transfer, readmission, and restart describe administrative procedures that modify the student's relationship with the institution or an academic program. First-time and regular continuing statuses represent academic progression conditions. Partial Academic Probation (PAP) and Total Academic Probation (PAT) identify successive levels of regulatory academic risk, whereas academic recovery represents an improvement following a critical condition. Outside the University (PFU) indicates that the student was not enrolled during a regular academic term. Finally, graduation and exclusion represent two forms of trajectory completion: the former results from fulfilling institutional requirements, whereas the latter results from the loss of student status due to academic performance.

The alphabet Σ was constructed from the academic conditions and administrative events that trigger changes between states. Therefore, its symbols represent not only actions performed by students but also institutional decisions and conditions resulting from their academic performance. The averages used in these rules were the term weighted grade point average (PPP) and cumulative weighted grade point average (PPA), for which the regulatory threshold of 3.2 was adopted. The complete definition of the alphabet is presented in Table II.

Based on these elements, the function δ linked the current state to the observed event to determine a single subsequent state. Fig.2 presents the reduced and expanded representations of an ideal trajectory ending in graduation, in which the student completes the credits required for graduation within the expected number of semesters. Fig. 3 provides an example: initial enrollment leads from enrolled applicant to first-time student through the corresponding symbol. If the student subsequently obtains a PPP equal to or greater than 3.2, the corresponding transition leads to regular continuing status. In contrast, a PPP below 3.2 activates the transition to PAP.

Repeated low academic performance may lead from PAP to PAT, whereas an improvement in the PPP without having yet recovered the PPA enables the transition from PAT to academic recovery. All possible representations are shown in Fig. 3, and Table III presents all possible transitions.

Thus, the DFA was not limited to representing isolated states; it formalized the conditions under which a student remains in the same status, progresses, enters academic risk, recovers, interrupts their studies, or completes their trajectory. The output of this phase was a deterministic automaton with explicitly defined states and transitions, which was used in the following phase to transform longitudinal records into formal trajectory sequences.

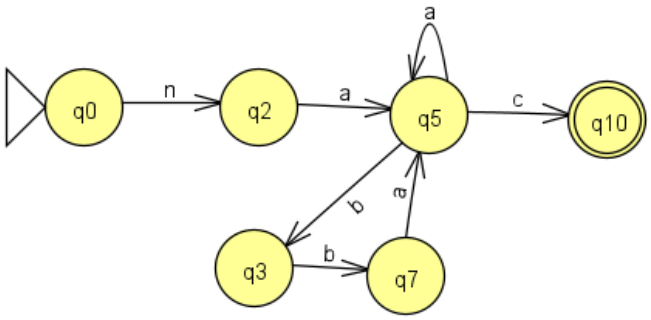


Fig. 3 Trajectory: Where the transition function is fulfilled

δ= naaaabbaaaac

TABLE II

ALPHABET OF THE AUTOMATON

Symbol	Transition Event or Condition
(a)	The student maintains academic and financial enrollment and obtains a PPP ≥ 3.2.
(b)	The student maintains enrollment but obtains a PPP < 3.2.
(c)	The student fulfills the academic and administrative requirements established for graduation.
(d)	The student simultaneously records a PPP < 3.2 and a PPA < 3.2.
(e)	The student obtains a PPP ≥ 3.2 but maintains a PPA < 3.2.
(f)	The internal transfer request is approved, and the student formally enrolls in the new program.
(g)	The student submits a formal application for readmission to the institution.
(h)	The readmission application is approved, and the student formally completes academic and financial enrollment again.
(i)	The student submits a request to restart their academic trajectory.
(j)	The restart request is approved, and the corresponding enrollment is formally completed.
(k)	The student has no academic or financial enrollment during a regular academic term.
(m)	The dual-degree program application is approved, and the student formally enrolls.
(n)	The student is admitted and formally completes their initial academic and financial enrollment.
(r)	The student formally completes an internal transfer between programs and continues their academic trajectory.
(s)	An external transfer request or process from or to another institution is recorded.

C. Automaton Validation

Validation was conducted at the formal, computational, and manual levels. Each observed transition was compared with the DFA transition function, as defined in (3):

$$\delta(q_t, \sigma_t) = q_{t+1} \qquad (3)$$

Transitions were classified as valid, undefined, ambiguous, or inconsistent. For computational validation, the records were ordered by students and academic terms. The final state of each trajectory was excluded from the evaluation because it had no subsequent state. In cases involving absence of enrollment, readmission, transfer, or restart, institutional rules were applied to reconstruct only the required administrative states, without imputing nonexistent academic information.

Manual validation was conducted on a sample of trajectories selected to represent frequent states and transitions, academic risk cases, and more complex situations, including missing academic terms, re-entry, program changes, and estimated graduation. For each case, the automatically generated sequence was compared with the original academic records and the institutional rules used to define the automaton's states and events.

Discrepancies were documented and resolved by reviewing the source records and the corresponding institutional rule. Agreement was calculated for state and transition assignments, both overall and by category.

TABLE III

BEHAVIOR OF THE AUTOMATON

Antecedent State (symbol DFA)	transition	Consequent State (symbol DFA)
q12	c	q10
q0	s	q1
q5	b	q3
q3	a	q5
q4	i	q8

q12	f	q2
q12	r	q6
q5	i	q8
q12	b	q3
q2	a	q5
q7	a	q5
q1	n	q2
q2	k	q4
q5	c	q10
q11	a	q5
q12	i	q8
q5	r	q6
q5	a	q5
q0	n	q2
q5	k	q4
q13	i	q8
q2	r	q6
q4	k	q4
q7	e	q11
q9	h	q12
q2	b	q3
q3	b	q7
q11	i	q8
q12	a	q5
q12	k	q4
q11	d	q13
q8	j	q2
q6	f	q2
q4	g	q9

D. Characterization of Academic Trajectories

Once the DFA had been validated, each trajectory was represented as an ordered sequence of states and transitions, as shown in (4):

$$T_i = \left(q_{i1}, \sigma_{i1}, q_{i2}, \ldots, \sigma_{i,n_i-1}, q_{i,n_i}\right) \qquad (4)$$

where q_{it}represents the academic or administrative state of student $i$during term t, and σ_{it}denotes the event that determines the transition to the subsequent state. This representation made it possible to analyze academic progression as a longitudinal process rather than solely through the final outcome achieved.

The characterization was first conducted for the entire population and then by gender. The states visited, trajectory duration, final outcome, and most frequent pathways were

analyzed. In addition, structural variables were derived, including entries into academic risk, recovery, relapses, positive and negative transitions, transition balance, and stability and risk indices.

Frequent pathways were identified by grouping trajectories with the same sequence of states. Additionally, an empirical transition matrix was constructed to represent the probabilities with which transitions occurred in data Fig.4.

E. Statistical Evaluation

The statistical evaluation determined whether the variables derived from the DFA could identify differences in academic progression between women and men. First, the overall study population was described, followed by gender-based comparisons. Categorical variables were analyzed using the chi-square test and Cramér's V, whereas quantitative variables were compared using the Mann–Whitney U test and rank-biserial correlation.

Additionally, a first-order Markov chain was applied to estimate transition probabilities between academic states. The matrices obtained for women and men were compared to identify differences in persistence, academic risk, recovery, interruption, and graduation. A significance level of 0.05 was adopted, while effect sizes and the academic relevance of the results were also considered.

IV. RESULTS AUTOMATION

This section presents the results of the trajectory characterization and first order Markov analysis. It also examines gender differences in structural indicators and transition probabilities,

A. Overall Characterization of Academic Trajectories

Of the 9,972 students analyzed, 5,895 (59.1%) had an active trajectory at the end of the observation period, 2,899 (29.1%) were classified as estimated graduates, 976 (9.8%) ended in an academic risk condition, and 202 (2.0%) had an interrupted trajectory. Regarding risk-related events, 3,609 students (36.2%) experienced at least one episode of academic risk, 2,938 (29.5%) experienced at least one recovery, and 1,700 (17.0%) experienced at least one relapse.

Regular standing (RS) was the predominant academic state, accounting for 46,476 records, equivalent to 55.71% of the 83,419 longitudinal observations. It was followed by first enrollment (FE; 13.22%), applicant (11.95%), PAP (5.42%),

PFU (3.75%), and graduation (GR; 3.48%). The least frequent states were PAT (1.95%), academic recovery (AR; 0.86%), and academic exclusion (AE; 0.72%).

These results show that RS predominated among the

After reaching AR, only 11.6% of the transitions returned to RS, whereas 88.4% progressed to AE. This finding identifies AR as a particularly critical point in the institutional progression process.

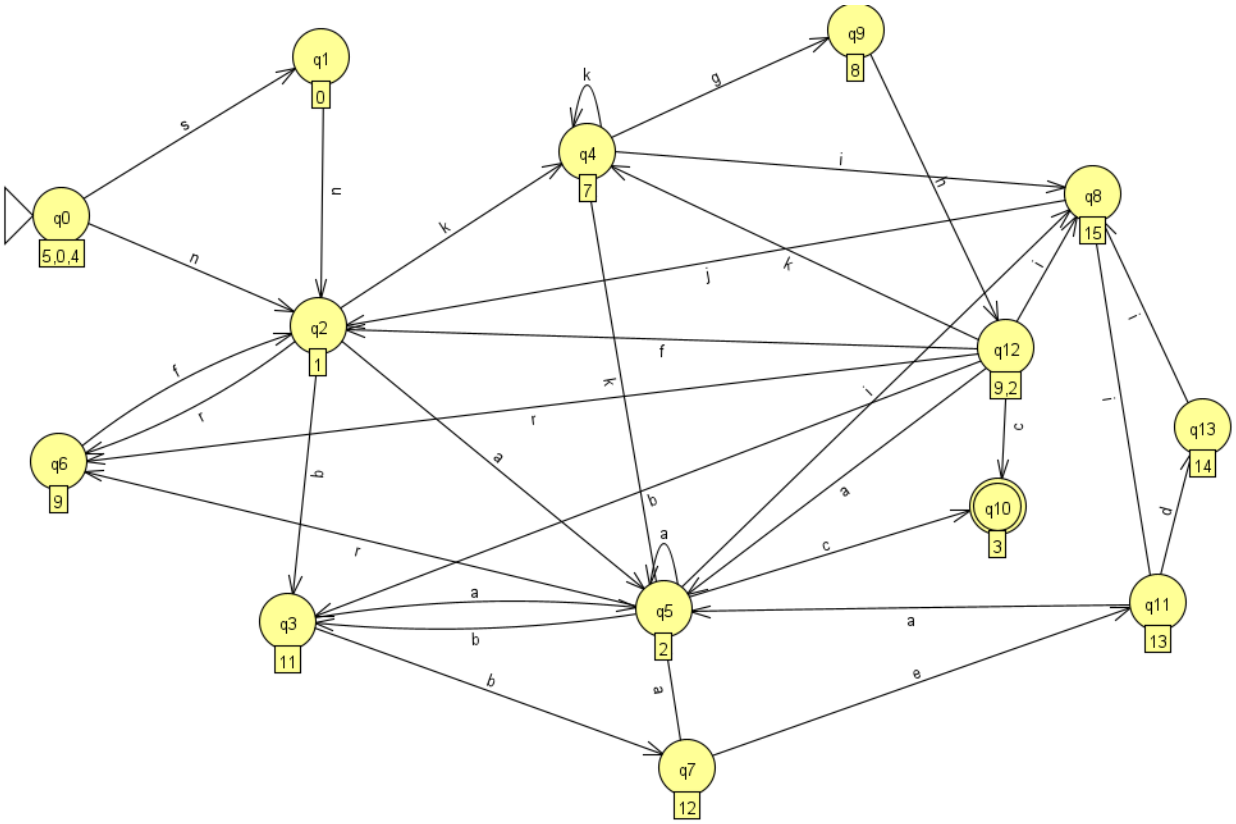


Fig.4 Representation with all DFA transitions.

student period records. However, the 55.71% should not be Fid.5

interpreted as the percentage of students with stable trajectories because this value was calculated from longitudinal observations rather than individual students.

The first order Markov analysis included 58,773 observed transitions among the eight academic states selected for the transition matrix, as shown in Fig. 5. The most frequent transition was from RS to RS, with 35,010 occurrences and a conditional probability of 0.826. Therefore, after reaching RS, students had an 82.6% probability of remaining in the same state during the subsequent observed term.

From FE, the probability of progressing to RS was 0.852, whereas the probabilities of entering PAP and PFU were 0.112 and 0.033, respectively. From RS, the probability of entering PAP was 0.084, the probability of moving to PFU was 0.023, and the probability of reaching GR was 0.068.

Progression from PAP followed two main pathways. Of the transitions originating from PAP, 56.7% returned to RS, whereas 43.3% progressed to PAT. From PAT, 46.3% returned directly to RS, whereas 53.7% moved to AR.

GR and AE were represented as absorbing states in the Markov matrix, although their observed outgoing transition counts were zero. Thus, the probabilistic representation recognizes that no subsequent states were observed after these terminal conditions without creating nonexistent empirical transitions.

The PFU TOPFU transition probability of 1.00 was calculated from 1,853 transitions associated with the reconstruction of missing academic terms. Therefore, this value should be interpreted because of the trajectory reconstruction rules rather than as direct evidence of actual academic persistence in PFU.

At the student level, the predominant pattern was *Stable active trajectory*, identified in 4,537 students (45.50%). This was followed by *Stable graduate*, with 1,826 students (18.31%); *Resilient active trajectory with recurrent risk*, with 673 (6.75%); *Resilient active trajectory*, with 648 (6.50%); *Resilient graduate*, with 605 (6.07%); and *Persistent risk*, with 604 students (6.06%).

Among the 2,899 estimated graduates, 1,826 (63.0%) had stable trajectories without risk episodes. In contrast, 605 (20.9%) reached graduation after experiencing a nonrecurrent

risk episode, 454 (15.7%) graduated after experiencing recurrent risk episodes, and 14 (0.5%) reached graduation after passing through critical academic states. Therefore, graduation was not limited to linear or consistently stable trajectories.

The most frequent functional pathway was *Entry–Stability*, observed in 2,586 students (25.93%). This was followed by *Entry*, with 1,947 students (19.52%), and *Entry–Stability–Graduation*, with 1,799 students (18.04%).

Pathways that included episodes of academic risk exhibited greater heterogeneity. For example, *Entry–Stability–Risk* accounted for 4.56% of the trajectories, whereas *Entry–Stability–Risk–Stability–Graduation* accounted for 4.21%.

Nevertheless, the predominance of the *Stable active trajectory* pattern should be interpreted cautiously because of the incomplete follow-up of recent cohorts. Students in this category had a mean of 3.24 observed semesters and a median of two semesters. Furthermore, 1,949 students in this pattern (42.96%) had only one observed semester, and 55.30% had no more than two semesters of follow-up.

Therefore, this category primarily represents active or right-censored trajectories and does not necessarily provide evidence of consolidated long-term academic stability.

B. Comparison of Academic Trajectories by Gender

Women and men exhibited statistically significant differences in the categorical trajectory indicators. The proportion of students who experienced academic risk was lower among women than among men: 28.3% versus 40.3%. Women also had a higher estimated graduation rate, at 34.3%, compared with 26.5% among men. In addition, women had lower proportions of trajectories ending in PAP, PAT, and AE.

The experience of academic risk was significantly associated with gender, $\chi^2(1) = 137.06$, $p < 0.001$, although the association was small ($V = 0.117$). Estimated graduation was also significantly associated with gender, $\chi^2(1) = 65.89$, $p < 0.001$, but the effect was negligible ($V = 0.081$).

The strongest categorical association was observed for the overall trajectory pattern, $\chi^2(12) = 231.09$, $p < 0.001$, with a small effect size ($V = 0.152$). Overall, Cramér's V values ranged from 0.081 to 0.152. This indicates that the differences were systematic but limited in practical magnitude.

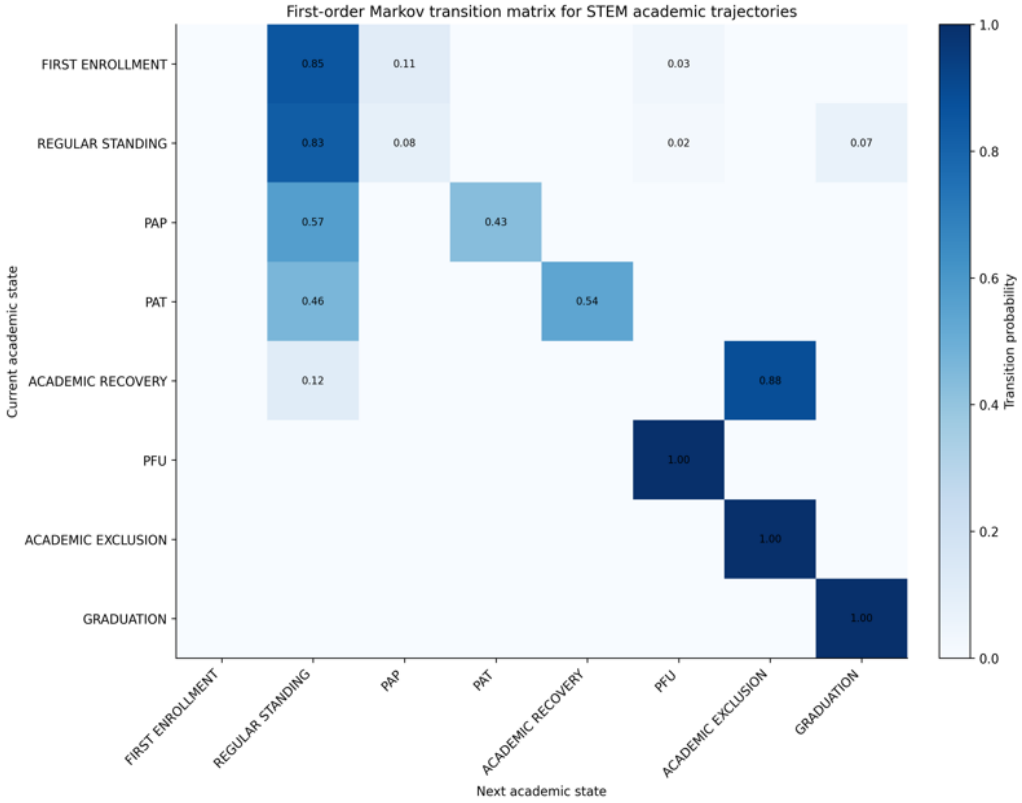


Fig. 5 First order Markov transition matrix for STEM academic

The quantitative indicators showed a similar trend. Women had a higher mean stability index than men (0.918 versus 0.862; $p < 0.001$, $r_{rb} = 0.138$) and a more favorable transition balance (0.168 versus −0.059; $p < 0.001$, $r_{rb} = 0.142$).

In contrast, men had a higher mean number of entries into risk states (0.734 versus 0.464; $p < 0.001$, $r_{rb} = -0.131$) and negative transitions (0.927 versus 0.566; $p < 0.001$, $r_{rb} = -0.134$). Although these differences were statistically significant, their effect sizes remained negligible to small.

The unadjusted proportion of students who experienced at least one recovery was higher among men than among women: 32.5% versus 23.6%. However, this result was influenced by men's greater exposure to academic risk.

Among students who experienced academic risk, 793 of the 952 women recovered at least once (83.3%), compared with 2,141 of the 2,653 men (80.7%). Furthermore, among those who recovered, relapse occurred in 48.7% of women and 61.3% of men.

Therefore, the higher absolute number of recoveries among men should not be interpreted as evidence of more favorable academic trajectories. Men experienced more recoveries because they also entered risk states more frequently. When the analysis was conditioned on having experienced academic risk, women showed a slightly higher recovery rate and a lower rate of subsequent relapse.

C. Gender Differences in Transition Probabilities

The gender-specific Markov matrices showed that the main differences were concentrated in entry into and progression toward more severe academic risk. Fig. 6 presents the differences in transition probabilities, calculated as the probability for women minus the probability for men. Positive values indicate transitions that were more likely among women, whereas negative values indicate transitions that were more likely among men.

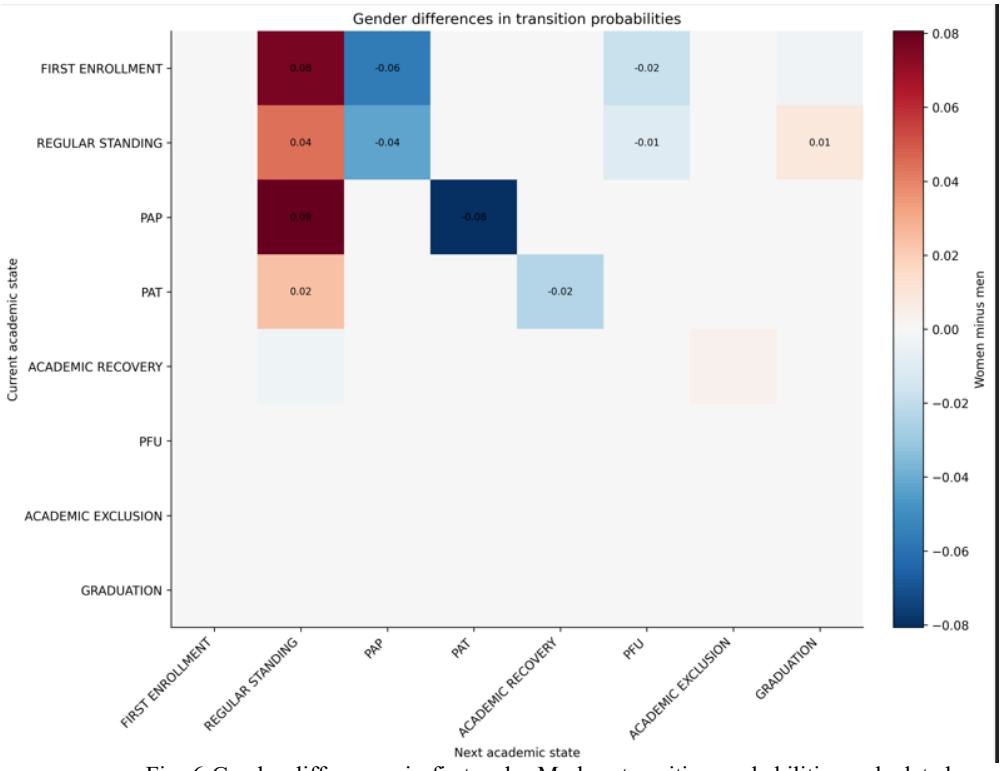


Fig. 6 Gender differences in first order Markov transition probabilities, calculated as women minus men.

The probability of moving from FE to RS was 0.903 among women and 0.826 among men. Conversely, the probability of moving from FE to PAP was lower among women (0.074) than among men (0.131). The largest difference between the groups was observed in the transition from PAP to PAT. Its probability was 0.372 for women and 0.453 for men. The difference, calculated as women minus men, was −0.081, indicating that this transition was 8.1 percentage points more likely among men.

Similarly, the probability of returning from PAP to RS was higher among women than among men: 0.628 versus 0.547. Overall, these results indicate that the most relevant gender differences occurred during entry into and the initial progression toward more severe academic risk.

The differences decreased once students reached more critical academic conditions. From PAT, the probabilities of returning directly to RS were 0.483 for women and 0.459 for men, corresponding to a difference of 2.4 percentage points.

From AR, the probability of progressing to AE was 0.887 for women and 0.883 for men. The difference was only 0.003. This indicates that once students reached the most critical institutional states, their subsequent transition probabilities were virtually identical across both groups.

V. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

The deterministic finite automaton provided a structured representation of academic progression based on institutional states and explicit transition rules. Unlike analyses limited to graduation or dropout, this representation captured risk, recovery, relapse, and the routes followed by students throughout their academic trajectories.

The derived variables revealed differences between women and men, mainly in the entry into and progression toward academic risk. The largest descriptive difference occurred in the transition from PAP to PAT, which was 8.1 percentage points higher among men. Most effect sizes were small, and the differences decreased in the most critical states. The findings therefore do not support the existence of completely different trajectories by gender, but rather moderate variations within a shared institutional progression system.

The study was conducted at a single institution, and recent cohorts had limited follow-up time. Graduation was estimated, some PFU periods were reconstructed, and the first-order Markov model considered only the current state when estimating the next transition. The gender comparisons were not adjusted for program, cohort, or observation time. These conditions limit causal interpretation and the direct generalization of the findings to other institutions.

Future work should externally validate the automaton using complete cohorts and data from other universities. Direct comparisons between women and men within academic programs and across cohorts should also be conducted through stratified analyses that account for follow-up time and the number of students and transitions available in each subgroup. Higher-order and duration-sensitive models could incorporate trajectory history and time spent in each academic state. Before operational use in early warning systems, predictive performance and error rates should be evaluated across demographic and academic groups.

ACKNOWLEDGMENTS

The authors express their gratitude to the Call 933 ``Training in National Doctorates with a Territorial, Ethnic and Gender Focus in the Framework of the Mission Policy — 2023" of the Ministry of Science, Technology and Innovation (Minciencia). In addition, we thank the team of the Artificial Intelligence Laboratory VerbaNex \https://github.com/VerbaNexAI, affiliated with the UTB, for their contributions to this project.

REFERENCES

[1] Consejo Nacional de Política Económica y Social, "Pacto Colombia con las Juventudes: Estrategia para fortalecer el desarrollo integral de la juventud, Documento CONPES 4040.," Bogota,Colombia, 2021.

[2] I. Katsantonis, "Tracking adolescent students' educational pathways to university through school engagement, parental expectations, and student aspirations," *European Journal of Psychology of Education*, vol. 40, no. 1, p. 4, Mar. 2025, doi: 10.1007/s10212-024-00917-3.

[3] R. Sweeder, M. N. Kursav, and S. Valles, "Mechanisms of STEM Retention: Lessons from Interview Data from Students in STEM Cohort Program."

[4] C. Li, N. Herbert, S. Yeom, and J. Montgomery, "Retention Factors in STEM Education Identified Using Learning Analytics: A Systematic Review," *Educ. Sci. (Basel).*, vol. 12, no. 11, p. 781, Nov. 2022, doi: 10.3390/educsci12110781.

[5] M. Moreno-Novoa, E. Puertas, and J. C. Martinez-Santos, "Enhancing Vocational Guidance with Machine Learning: Predicting STEM Career Viability for High School Students," 2024, pp. 150–162. doi: 10.1007/978-3-031-75236-0_12.

[6] J. P. Salazar-Fernandez, J. Munoz-Gama, J. Maldonado-Mahauad, D. Bustamante, and M. Sepúlveda, "Backpack process model (Bppm): A process mining approach for curricular analytics," *Applied Sciences (Switzerland)*, vol. 11, no. 9, May 2021, doi: 10.3390/app11094265.

[7] H. Lee *et al.*, "To stay, switch, or leave: A four-year longitudinal study of the situated and stable social influences on women's STEM major choices," *Contemp. Educ. Psychol.*, vol. 79, p. 102324, Dec. 2024, doi: 10.1016/j.cedpsych.2024.102324.

[8] R. F. Semler, J. R. Semler, M. A. Wehrmeister, L. F. P. Southier, D. Casanova, and M. Teixeira, "Educational process mining: literature classification, gaps, and emerging opportunities," *Front. Educ. (Lausanne).*, vol. 10, Jun. 2025, doi: 10.3389/feduc.2025.1543761.

[9] I. W. AlHmoud, S. Poudel, S. Deshmukh, C. S. Booth, G. Monty, and M. Bikdash, "Importance of alternative pathways to STEM through analysis of longitudinal student dataset," *Discover Education*, vol. 3, no. 1, p. 134, Aug. 2024, doi: 10.1007/s44217-024-00220-9.

[10] N. Ismail and U. K. Yusof, "A systematic literature review: Recent techniques of predicting STEM stream students," Jan. 01, 2023, *Elsevier B.V.* doi: 10.1016/j.caeai.2023.100141.

[11] R. Cerezo, J. A. Lara, R. Azevedo, and C. Romero, "Reviewing the differences between learning analytics and educational data mining: Towards educational data science," *Comput. Human Behav.*, vol. 154, May 2024, doi: 10.1016/j.chb.2024.108155.

[12] A. N. de Vasconcelos *et al.*, "Advancing school dropout early warning systems: the IAFREE relational model for identifying at-risk students," *Front. Psychol.*, vol. 14, Jul. 2023, doi: 10.3389/fpsyg.2023.1189283.

[13] M. M. Novoa *et al.*, "EarlySTEM: Machine Learning Platform for Early Identification of Dropout Risk in STEM Programs," in *2025 IEEE Colombian Caribbean Conference (C3)*, IEEE, Sep. 2025, pp. 1–6. doi: 10.1109/C366505.2025.11339990.

[14] S. He, C. Demmans Epp, F. Chen, and Y. Cui, "Examining change in students' self-regulated learning patterns after a formative assessment using process mining techniques," *Comput. Human Behav.*, vol. 152, Mar. 2024, doi: 10.1016/j.chb.2023.108061.

[15] J. P. Pêgo, V. L. Miguéis, and A. Soeiro, "Students' complex trajectories: exploring degree change and time to degree," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, p. 8, Jan. 2024, doi: 10.1186/s41239-024-00438-5.

[16] G. A. Ahmed, J. L. Ávila-Jiménez, M. I. Al-Twijri, and S. Ventura, "Academic Trajectory Graphs for Temporal Modelling of Student Academic Progression," *Applied Sciences (Switzerland)*, vol. 16, no. 11, Jun. 2026, doi: 10.3390/app16115642.

[17] M. N. Tedeschi, T. M. Hose, E. K. Mehlman, S. Franklin, and T. E. Wong, "Improving models for student retention and graduation using Markov chains," *PLoS One*, vol. 18, no. 6 JUNE, Jun. 2023, doi: 10.1371/journal.pone.0287775.

[18] M. Moreno Novoa, Y. Henriquez Nuñez, J. C. Martinez-Santos, and E. Puertas, "Predictive Model for STEM Vocational Guidance through Profile Analysis and Information Adaptation with a Gender Perspective," in *Proceedings of the 5th LACCEI International Multiconference on Entrepreneurship, Innovation and Regional Development (LEIRD 2025): "Entrepreneurship with Purpose: Social and Technological Innovation in the Age of AI,"* Latin American and Caribbean Consortium of Engineering Institutions, 2025. doi: 10.18687/LEIRD2025.1.1.1122.

[19] S. López-Pernas, K. Misiejuk, and M. Saqr, "A Systematic Literature Review of Sequence Analysis in Learning Analytics."

[20] S. Bacci and B. Bertaccini, "A Mixture Hidden Markov Model to Mine Students' University Curricula," *Data (Basel).*, vol. 7, no. 2, p. 25, Feb. 2022, doi: 10.3390/data7020025.