IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 005, LIR327
DOI: https://doi.org/10.52731/lir.v005.327

A Model for Understanding Student Status Using Attendance Data in the First Semester of University

Naruhiko Shiratori *

Abstract

This study developed a Hidden Markov Model (HMM) to analyze attendance behaviors of first-year university students during their spring semester, aiming to identify distinct behavioral patterns and examine their impacts. Weekly attendance data was used to estimate latent states, and clustering revealed four representative attendance patterns, including stable attendance and increased absenteeism. The results highlight the potential impact of specific behaviors on academic outcomes, underscoring the importance of preventive interventions in student support and its applicability to future academic guidance.

Keywords: Hidden Markov Model, Attendance Patterns, Student State Analysis, Behavior Analysis in University Students

1 Introduction

1.1 University Dropout and Academic Performance in Japan

The university dropout rate in Japan remains a pressing issue for educational institutions, impacting not only students and their families but also the reputation and educational quality of universities. According to the Ministry of Education, Culture, Sports, Science and Technology, the dropout rate for Japanese universities and junior colleges was 1.94% in the 2023 academic year, equating to 52,495 students [1]. Rates reported by the Yomiuri Shimbun vary, with dropout percentages at 2.9% for national universities, 4.2% for public universities, and 8.0% for private universities [2]. While these rates appear low compared to other OECD countries, they average out considerable variation across institutions and disciplines, with some universities experiencing dropout rates as high as 20% over a four-year period. Shimizu et al. further highlight that dropout rates correlate with entrance exam scores, with some fields, such as social sciences, showing dropout rates near 12% for programs with average entrance scores around 45 [3].

Academic performance, particularly the GPA of the first-year spring semester, has been identified as a critical factor in dropout prediction. Takahashi et al. found that students with lower motivation and performance upon enrollment at the University of the Ryukyus were more likely to be at risk of dropout [4]. Ortiz Lozano et al., studying dropout among engineering students in Spain, demonstrated that academic performance was a predictive dropout factor and showed that using first-semester performance data improves prediction accuracy [5]. This body of research underscores the importance of monitoring early academic performance as an effective strategy for dropout prevention.

* Tokyo City University, Tokyo, Japan

1.2 Relationship Between Attendance and Academic Performance

Enhancing first-year students' academic performance in the spring semester is vital for dropout prevention. However, it is also crucial to intervene before final semester grades are available. Ortiz Lozano et al. reported a dropout prediction accuracy of 76% when data from spring semester grades became available. Likewise, Shiratori et al. showed a prediction accuracy of 0.799, with a recall of 0.702, using spring semester grades [6]. However, these findings suggest that waiting until final grades are released could delay interventions.

To enable timely intervention, evidence-based approaches using in-semester indicators are necessary. Shiratori et al. used weekly attendance and pre-enrollment data to predict first-year spring semester grades, finding that student behavior tends to stabilize by the fifth or sixth session [7]. Similarly, Kondo et al. classified at-risk students based on attendance and learning management data, achieving prediction accuracies ranging from 0.5 by the fourth week to between 0.6 and 0.7 by the tenth week [8]. These studies highlight the predictive potential of class attendance data as a reliable indicator of both academic performance and dropout risk, which underscores its importance for early intervention.

1.3 Purpose of the Study

As noted above, attendance data is valuable in predicting academic performance and university dropout risk. However, this study explores the additional aspect of how attendance patterns across the semester impact student state transitions. Attendance data directly correlates with academic performance and dropout rates, yet specific patterns within the semester, their influence on state transitions, and their impact on academic outcomes remain underexplored. This study leverages weekly attendance data to gain insights into students' latent state transitions over time, providing information that can enhance dropout prediction models and frameworks for real-time student state monitoring, ultimately supporting proactive dropout prevention strategies.

2 Related Studies

2.1 Understanding and Monitoring Student Status

Monitoring students' academic status throughout the semester is challenging. Regular surveys or tests could provide insights into learning states, but conducting them weekly is impractical. Therefore, methods that leverage LMS data and attendance records are essential. Shiratori et al., as mentioned previously, predicted first-year spring semester GPAs using a random forest model, incorporating pre-enrollment data and weekly attendance variations to cluster and analyze student states [7]. Kondo et al. built a Bayesian network model to estimate academic status, predicting future enrollment and dropout risk using attributes like gender, GPA, and attendance rates [9]. While these studies have used regression and probabilistic models to assess student status, few have focused on weekly attendance data to track latent, unobservable state transitions.

2.2 Applications of Hidden Markov Models

Hidden Markov Models (HMM) are valuable for capturing unobservable states through observable data. For instance, Balakrishnan et al. used HMM to model student persistence and dropout in MOOCs, leveraging behaviors like video views and forum posts [10]. Tadayon and Pottie applied HMM in the educational game "Save Patch" to predict grades and track learning

progress [11]. Gupta et al. developed an HMM with four assessment levels to identify at-risk students early [12]. These studies illustrate HMM's usefulness in education for prediction and for gaining insights into learning processes.

This study also employs HMM to estimate and track student learning states throughout the semester using weekly attendance data. This approach aims to provide foundational insights for enhanced learning support.

3 Methodology

3.1 Research Questions

This study evaluates and assesses student states weekly using attendance data, tracking how individual students' states transition throughout the first-year spring semester and identifying characteristic transition patterns. By deriving these state transition patterns, we aim to determine the optimal timing for intervention and identify shifts in states within attendance patterns.

3.2 Data and Variables

The dataset comprises records from 181 students who entered a Tokyo-based liberal arts university (University A) in 2018. University A is a relatively small humanities-focused institution. Pre-enrollment data includes high school GPA (initial_gpa), total absences (total_absences), and high school type (school_type), detailed in Table 1. Descriptive statistics for initial_gpa and total_absences are shown in Table 2. School type is classified as "full-time" (166 students) or "other" (15 students).

Post-enrollment variables include weekly attendance data, divided into foundational seminars (attendance_seminar) and other courses (attendance_ot_weekly). The foundational seminar, designed to aid university adaptation, records attendance weekly (1 = attendance, 0 = absence) over 15 sessions, with a mean absence of 1.72. For other courses, weekly absences are summed, resulting in attendance data from week 1 to week 15 (see Table 3). Figure 1 shows weekly trends in absences, indicating both an increase in average absences and variance over the semester.

Table 1: Variables Used in This Study

Variable Name	Type	Description
initial_gpa	Numeric	GPA score at the time of enrollment for each student
total_absences	Numeric	Total number of absences during high school (integer)
school_type	Categorical	Type of high school attended by each student (0 = public, 1 = private, etc.)
attendance_seminar	Numeric	Weekly attendance status in the Basic Seminar for each student (0 = present, 1 = absent)
attendance_ot_weekly	Numeric	Weekly count of absences in OT courses for each student (integer representing the number of absences each week)

Table 2: Descriptive Statistics of Variables Available at Admission

	mean	median	std
initial_gpa	3.32	3.2	0.54
total_absences	11.13	4	17.72

Table 3: Descriptive Statistics of Weekly Absences in Other Courses

	week_1	week_2	week_3	week_4	week_5	week_6	week_7	week_8	week_9	week_10	week_11	week_12	week_13	week_14	week_15
mean	0.38	0.26	0.44	0.66	0.75	0.72	0.80	0.91	0.94	1.01	1.50	1.61	1.81	1.98	1.86
median	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
std	0.80	0.64	0.93	1.12	1.22	1.20	1.36	1.43	1.51	1.65	1.76	1.81	1.84	1.89	1.82

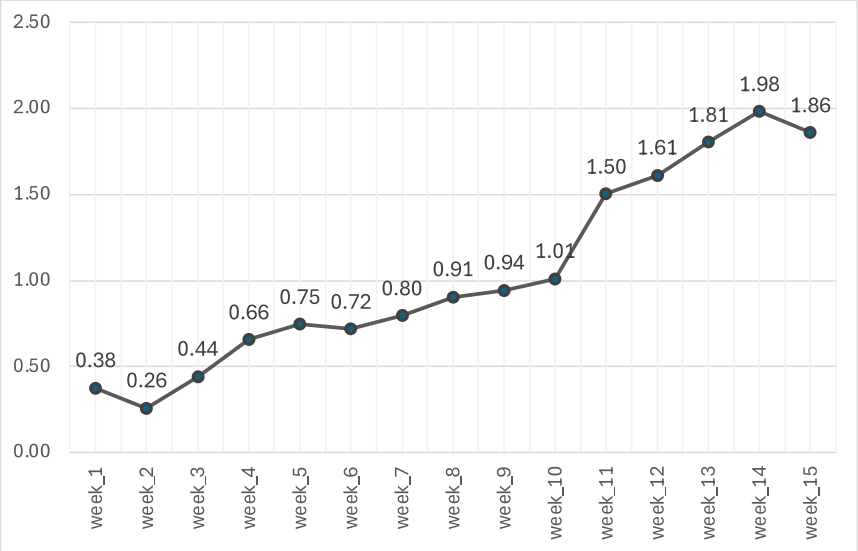

Figure 1: Weekly Trends in Average Absences in Other Courses

3.3 Model Development for Student State Identification

Based on the described data and variables, we developed a model to estimate weekly student states throughout the semester using a Hidden Markov Model (HMM). Each student's state is represented as time-series data, connected to weekly seminar and course attendance, with final states assumed to correlate with end-of-semester academic performance. Figure 2 overviews the model structure for estimating student states.

In this model, latent student states are estimated weekly with a categorical distribution governed by initial state probabilities (initial_probs) and a state transition matrix (transition_matrix). Observed variables, including foundational seminar attendance (obs_seminar) and absences in other classes (obs_ot), follow beta (p_seminar) and exponential (lambda_ot) distributions, respectively. Expected performance (gpa_mu) depends on the final week's state, allowing prediction of actual semester grades (GPA_1_1) through a normal distribution (gpa). This model thus connects hidden states with observed data, state transitions, and semester performance.

.

The latent student state (state) is a time-series categorical distribution, determined by initial

state probabilities and the state transition matrix, collectively defining transitions between states. Observed data—weekly absences in seminars (attendance_seminar_data) and other classes (attendance_ot_data)—are state-dependent.

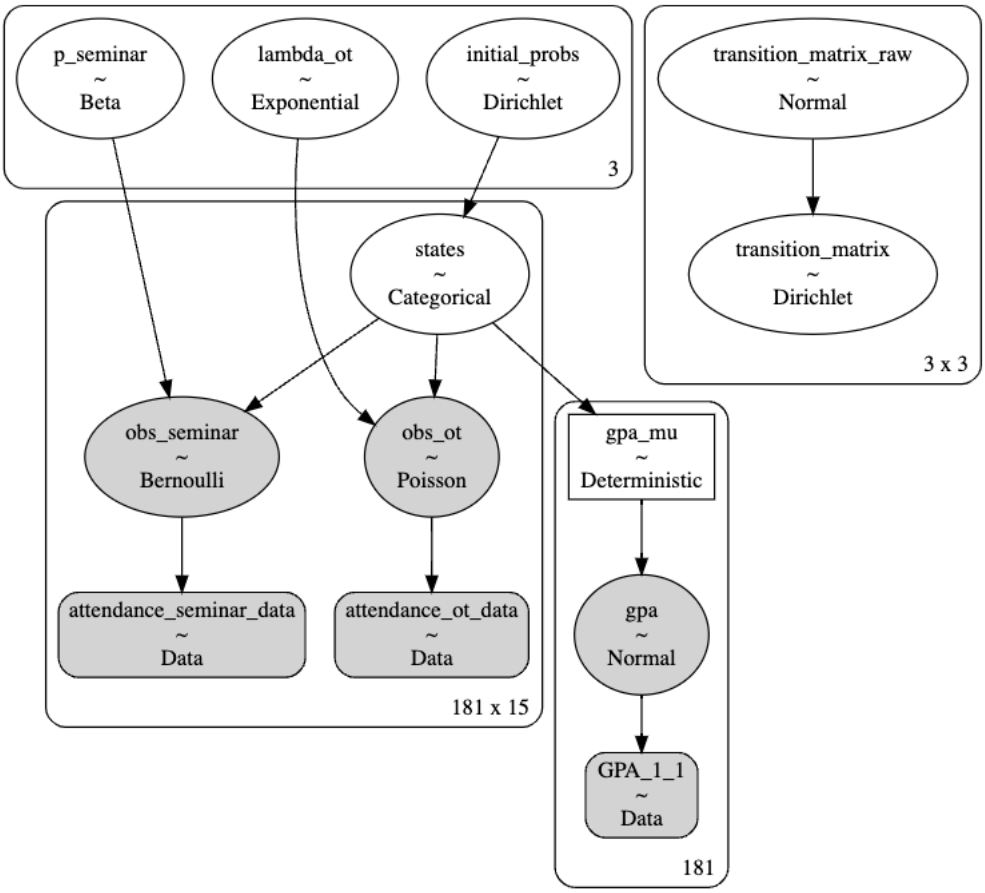

Figure 2: Student State Identification Model

The initial state probability (initial_prob) reflects each student's starting state likelihood, following a Dirichlet distribution with equal probability across three states. Pre-enrollment data, including high school GPA and absences, is incorporated to align initial states with prior academic backgrounds. The transition matrix is a 3×3 matrix representing state transitions, derived from a raw matrix (transition_matrix_raw) with each element following a normal distribution (mean = 0, SD = 0.5). Softmax transformation is applied to normalize values within [0, 1], resulting in row-specific transition probabilities.

The attendance probability for the seminar (p_seminar) reflects the likelihood of attendance based on state, generating observed seminar attendance (obs_seminar) through a Bernoulli distribution. This probability varies by state, following a beta distribution ($\alpha = 0.5$, $\beta = 5$), thus accommodating state-specific attendance tendencies. For other courses, the mean weekly absence (lambda_ot) is modeled via an exponential distribution, serving as a parameter in the Poisson distribution that defines observed attendance data (obs_ot). Observed absences depend on the lambda_ot value of each student's current state.

The variable gpa_mu represents the expected GPA for each latent state. In this model, states are assigned different expected GPA values ([4.0, 3.0, 2.0]), with final-week latent states deter-

mining expected GPA. The variable gpa is the observed GPA from the semester, modeled as a normal distribution centered on gpa_mu (μ = gpa_mu, σ = 0.5), constrained within [0, 4], capturing natural GPA variation around each latent state's expected value..

3.4 Clustering of Student State Patterns Using K-Means

Following the weekly state estimations generated by the model, we performed a typological analysis of state transitions for 181 students over 15 weeks (State_1, State_2, ..., State_15). Using the K-means clustering algorithm, we categorized these patterns and identified 4 as the optimal number of clusters, based on the elbow method.

To capture the variations in student state transitions, we analyzed and visualized state patterns over the 15 weeks of the first-year spring semester. Each student's weekly state data was aggregated into a 15-dimensional state vector. Using this data, we applied the K-means clustering method to classify students' state patterns. During clustering, we determined the optimal cluster number using the elbow method, which involves calculating the Sum of Squared Errors (SSE) for each cluster count and observing changes in SSE. We identified an "elbow" where SSE reduction slows around 4 clusters. Thus, we adopted 4 clusters as the optimal number, classifying students' state patterns into four distinct clusters. This clustering enables the identification of student groups with similar state transition patterns, providing a foundation for subsequent analysis of differing behaviors and impacts on academic performance across these groups.

4 Experiment Results and Discussion

4.1 Experimental Environment

In this study, we implemented the Hidden Markov Model (HMM) using Python and the PyMC library. PyMC is a probabilistic programming language (PPL) that enables sampling through Markov Chain Monte Carlo (MCMC) methods in Python. Specifically, we used the No-U-Turn Sampler (NUTS) algorithm, a variant of Hamiltonian Monte Carlo (HMC) with an adaptive step count feature.

For NUTS sampling, we set the target acceptance rate (target_accept) to 0.99 and the tree depth (max_treedepth) to 20, ensuring stable sampling even in complex parameter spaces. Additionally, we set 2000 tuning steps to optimize the exploration of parameter space in the initial phase, thereby enhancing convergence. These settings allowed us to improve the estimation accuracy and efficiency of the HMM, while accommodating variability and heterogeneity in the observed data.

4.2 Training Results of the Student State Identification Model

We performed 4000 sampling iterations, with a burn-in period of 2000, using two Markov chains. The Gelman-Rubin statistic (R-hat) for all parameters was below 1.1, confirming convergence.

Figure 3 presents trace plots, which visually depict the convergence status and posterior distributions of each parameter. The left side shows histograms of the posterior distributions, while the right side displays the sampling trajectories, confirming favorable convergence. The trace for initial state probabilities (initial_probs) is stable, indicating appropriate convergence for the initial state probabilities. Similarly, the transition probabilities (transition_matrix and transition_matrix_raw) are estimated without significant variance, showing stable convergence.

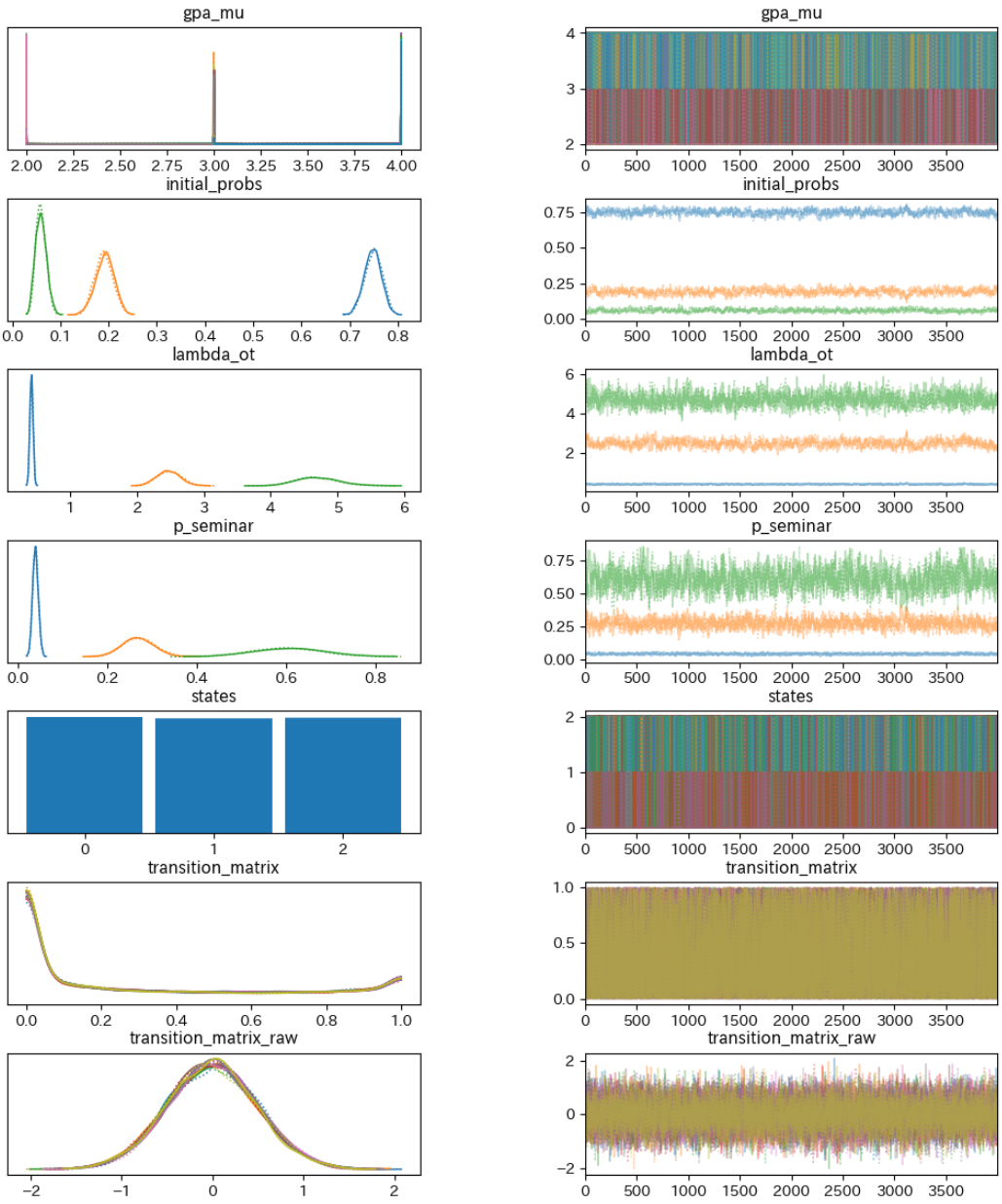

Figure 3: Trace Plot of the Student State Identification Model

The traces for the mean absences in other classes (lambda_ot) and seminar attendance probability (p_seminar) also demonstrate proper convergence, adjusted according to the respective states. The trace of hidden states (states) reveals that students' weekly states are consistently classified, and state transitions are appropriately modeled. Additionally, the expected GPA (gpa_mu) and actual GPA (gpa) are both sampled stably, supporting the validity of the semester-end GPA predictions.

Overall, as each parameter trace is smooth and R-hat values remain below 1.1, the sampling can be considered well-converged.

4.3 Analysis of Latent Student State Patterns

In the model constructed for this study, student latent states are categorized into three distinct patterns. These states are characterized by attendance rates in foundational seminars, absence frequency in other courses, and final grades (GPA). Figure 4 illustrates the absence rate for foundational seminars, the number of absences in other courses, and first-year spring semester GPAs for each state.

State 1 represents students with high attendance rates and few absences in other courses. The absence rate for foundational seminars is the lowest, and the absence frequency in other courses is also minimal. Students in this state have an expected GPA of 4.0, indicating a high level of academic performance. Such students are likely to manage their studies efficiently while minimizing absences, actively engaging in learning outside of classes, and exhibiting strong self-regulatory learning behaviors.

State 2 reflects students with a slightly higher absence rate in foundational seminars compared to State 1, along with a tendency toward increased absences in other courses. The absence rate for foundational seminars is higher than that in State 1, and the frequency of absences in other courses also increases. The expected GPA for students in this state is approximately 3.0, indicating average academic performance. These students demonstrate learning motivation but may benefit from revising their study plans, given their absences in certain courses, to improve their academic outcomes.

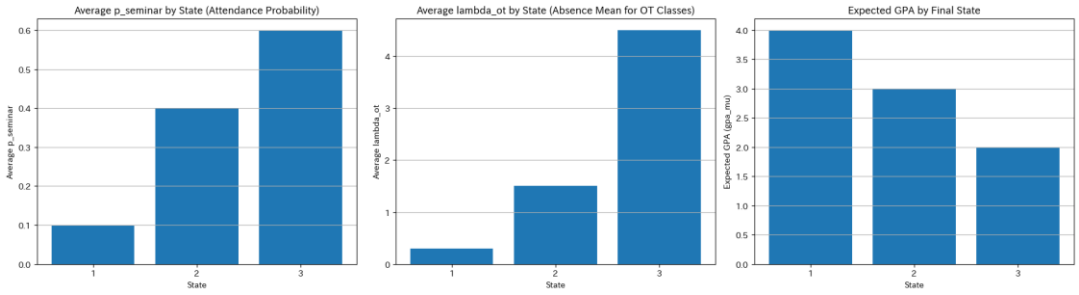

Figure 4: Absence Rate for Foundational Seminars, Absences in Other Courses, and GPA by State

State 3 represents students with the highest absence rate in foundational seminars and a high absence rate in other courses, along with lower academic performance. The absence rate for foundational seminars is the highest, and the frequency of absences in other courses is also maximal. The expected GPA for students in this state is 2.0, the lowest among the states. Students in this state appear to face challenges with attendance in both foundational seminars and other courses, suggesting that their overall academic performance may be at risk. Consequently, they may require additional support to achieve balanced academic engagement.

Through these three state patterns, student behavioral characteristics can be clearly understood, providing valuable indicators for developing learning support and intervention strategies tailored to each state.

4.4 Analysis of Latent Student State Patterns

Figure 5 visualizes the frequency of each hidden state by week, illustrating the predominant states students occupy over the course of the semester. For each of the three hidden states—State 1, State 2, and State 3—the frequencies were calculated on a weekly basis.

First, State 1 (depicted in dark blue) is observed most frequently, especially in the early stages of the semester. The frequency of State 1 is high from Week 1 to Week 10, then gradually decreases as the semester progresses. This pattern suggests that many students demonstrate a stable attendance trend during the early part of the semester.

In contrast, the frequency of State 2 (represented in green) tends to increase over time. The occurrence of State 2 rises notably from Week 11 onward, reaching its peak in the final week. This trend may indicate that some students begin to exhibit changes in their attendance behavior during the latter half of the semester, potentially due to factors such as fatigue or other external influences leading to increased absences.

Lastly, although the frequency of State 3 (shown in yellow) remains low overall, it notably increases after Week 10. This suggests that a small number of students begin to show unstable attendance patterns in the latter part of the semester. Students in State 3 are likely at higher risk of low attendance rates and, consequently, may face adverse effects on their academic performance.

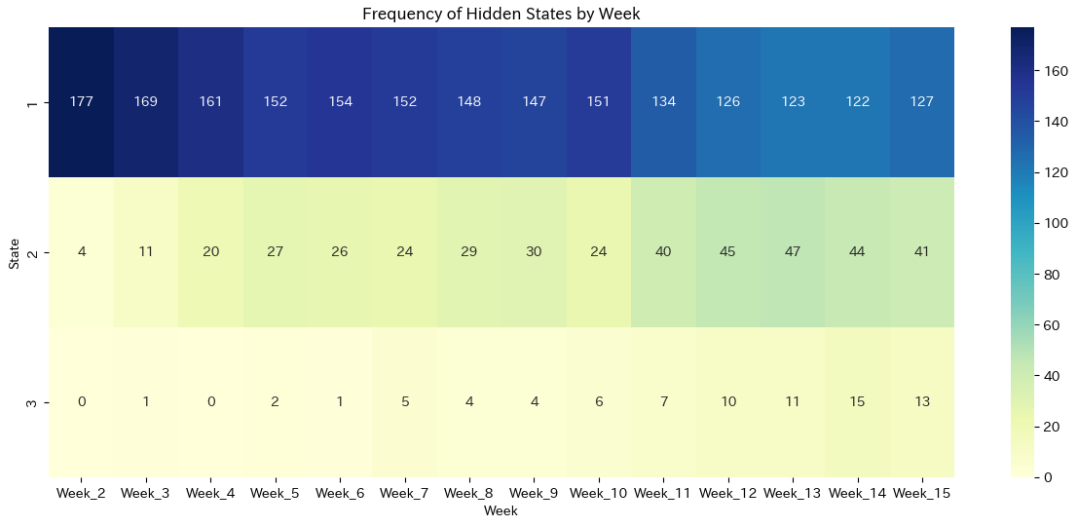

Figure 5: Weekly Counts for Each State

4.5 Clustering of State Transitions

Clustering was conducted based on the mode of weekly hidden states for each student, with the results shown in Figure 6. The number of clusters was set to four, and distinct behavioral patterns were observed for each cluster. Below, the characteristics of each cluster are described.

4.5.1 Cluster 1 (n=119, 65.7%)

Cluster 1 comprises 65.7% of the students, with a nearly constant weekly average state remaining at a low value (close to State 1). This group of students consistently exhibits a "high attendance tendency" or "low absence tendency," suggesting stability in learning attitudes and attendance behaviors. Such students are likely to maintain a proactive attitude toward academics.

4.5.2 Cluster 2 (n=30, 16.6%)

Cluster 2 includes 16.6% of the students, with an average state slightly lower than that of other clusters but displaying weekly fluctuations. Notably, a temporary increase is observed around the middle of the semester (Weeks 5 to 7), suggesting that many students in this group experi-

ence changes in attendance or absences during specific times in the semester. This fluctuation may be influenced by external factors, such as exams or assignment submission periods, indicating that mid-semester support might be effective for these students.

4.5.3 Cluster 3 (n=16, 8.8%)

Cluster 3 comprises 8.8% of the students, with the average state gradually increasing as the semester progresses. A notable rise occurs after Week 11, eventually reaching close to State 3. Students in this cluster tend to shift toward "increased absences" or "reduced attendance" in the latter half of the semester, possibly due to academic burden or decreased motivation. Support aimed at maintaining motivation and reducing academic load in the latter half of the semester would be beneficial for students in this cluster.

4.5.4 Cluster 4 (n=16, 8.8%)

Cluster 4 also comprises 8.8% of the students. Like Cluster 3, the average state rises in the latter half of the semester, but this increase occurs earlier than in Cluster 3. A marked increase is observed between Weeks 5 and 7, indicating a tendency toward "increased absences" from the early to mid-semester. This pattern suggests that students in this group may struggle with attendance from the beginning of the semester, and early-stage support could be effective.

Through this analysis, four distinct attendance behavior patterns based on students' attendance actions were identified. Notably, there is a wide variation in attendance behaviors, ranging from students with stable attendance patterns, like those in Cluster 1, to students with increasing absences as the semester progresses, such as those in Clusters 3 and 4. These results indicate the potential for tailored support based on the characteristics of each group. For instance, early intervention may be effective for students whose absences increase in the latter half of the semester.

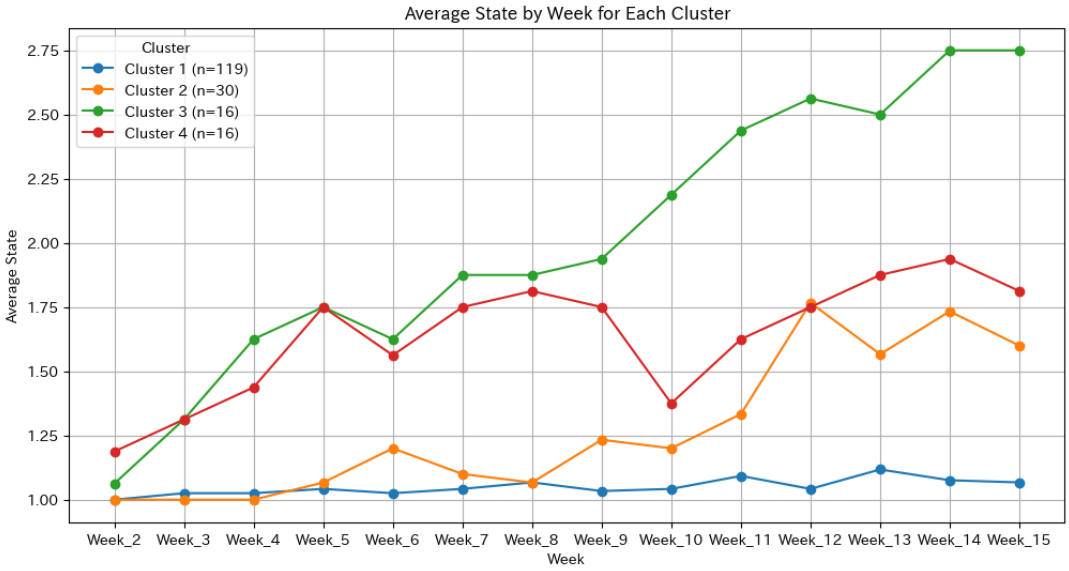

Figure 6: Average State by Week for Each Cluster

5 Conclusion and Future Issues

5.1 Summary of the Study

In this study, we constructed a Hidden Markov Model (HMM) for first-year university students during the spring semester to capture each student's attendance behavior as a state transition, with the aim of identifying latent behavioral patterns and understanding their impacts over the semester. Specifically, we estimated latent states based on weekly attendance data and clustered the 15-week state patterns, identifying four representative patterns of student attendance behavior. Through cluster-based analysis, we confirmed the existence of different behavioral trends, including students who consistently exhibit stable attendance, those who experience an increase in absences midway, those with periodic absences, and those who show increased absences in the latter part of the semester. Additionally, we examined how the latent states in the final week were associated with academic performance, discussing the potential impact of specific behavioral patterns on grades. This study provides important insights for considering preventive interventions in student support.

5.2 Future Research Challenges

However, this study has certain limitations. First, the HMM was constructed using only limited indicators such as attendance and academic performance data, which did not account for other factors that may influence student behavior and performance (e.g., extracurricular activities or individual motivation to learn). The absence of this information imposes some constraints on the accuracy of state transition predictions and the interpretability of the model. Additionally, subjective factors were involved in the choice of the number of clusters and model parameter settings, meaning that different parameter configurations or numbers of clusters could yield different results. Furthermore, settings for sampling iterations and burn-in periods also affect model convergence, necessitating thorough examination.

Acknowledgement

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research C 23K02668.

References

[1] Ministry of Education, Culture, Sports, Science and Technology. [Survey on the status of student enrollment (drop-outs and leaves of absence) (as of the end of the 2022 academic year)] “Gakusei no shugakujokyo (Chutai, Kyugaku) tou ni kansuru Chosa no Kekka ni Tsuite (in Japanese)”. 2023.

[2] Yomiuri Shimbun Kyoiku Network, [University Competencies 2019] ”Daigaku no Jitsuryoku 2019 (in Japanese)”,Chuo Koron Shinsho, 2018.

[3] Hajime Shimizu, [An Examination of the Factors that Lead to University Students Dropping Out: The Case of Social Science Departments] “”Daigakusei no Taigakuyoin no Kosatsu: Shakaikagakukei Gakubu no Case (in Japanese)”, Osakakeidai Ronshu, vol. 71, no. 5, pp.

1-13, 2021.

[4] Nozomi Takahashi, Yusuke Fujimoto, and Hiroki Nishimoto, [Investigation into the early detection of withdrawal, leave of absence, expulsion, and repeating the year in the undergraduate program at the University of the Ryukyus] ”Ryukyudaigaku Gakushikatei ni Okeru Taigaku Kyugaku Joseki Ryunen no Soukihakken ni Muketa Kento,” Ryukyu University Education Center Bulltein, no. 21, pp. 89–100, 2019.

[5] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa, “University student retention: Best time and data to identify undergraduate students at risk of dropout,” Innov. Educ. Teach. Int., vol. 57, no. 1, pp. 74–85, Jan. 2020.

[6] Naruhiko Shiratori, Tetsuya Oishi, Shintaro Tajiri, Masao Mori, and Masao Murota, [Making Dropout Patterns Using Transition of Dropout Probability] ”Chutaikakuritsu no Seni wo motiita chutaigakusei no ruikeika (in Japanese),” Nihon Kyoikukougakkai Ronbunshi (Japan journal of educational technology), vol 44, no.1, pp. 11-22, 2020.

[7] Naruhiko Shiratori, Tetsuya Oishi, Shintaro Tajiri, Masao Mori, and Masao Murota, [Clustering of Student Status in the Spring Semester of the First Year Using Predicted GPA Trends] “Yosoku GPA no Suii wo Motiita 1nenjiharugakki Gakushujotai no Ruikeika (in Japanese),” Kyoiku Shisutemu Johogakkaishi (Transactions of Japanese Society for Information and Systems in Education), vol. 39, no. 4, pp. 440–451, 2022/

[8] Nobuhiko Kondo, M. Okubo, and T. Hatanaka, “Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data,” in 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 198–201.2017.

[9] Nobuhiko Kondo, and Toshiharu Hatanaka, [Modeling of Learning Process based on Bayesian Network] “Bayesian Network niyoru Shugakujotai Suii Model no Kotiku (in Japanese),” Nihon Kyoikukougakkai Ronbunshi (Japan journal of educational technology), vol. 41, no. 3, pp. 271–281, 2018.

[10] G. Balakrishnan and D. Coetzee, “Predicting student retention in massive open online courses using hidden markov models,” Electrical Engineering and Computer Sciences University of California at Berkeley, vol. 53, pp. 57–58, 2013.

[11] M. Tadayon and G. Pottie, “Predicting student performance in an educational game using a hidden Markov model,” IEEE Trans. Educ., vol. 63, pp. 299–304, Apr. 2019.

[12] A. Gupta, D. Garg, and P. Kumar, “Mining Sequential Learning Trajectories With Hidden Markov Models For Early Prediction of At-Risk Students in E-Learning Environments,” IEEE Trans. Learn. Technol., vol. 15, no. 6, pp. 783–797, Dec. 2022.