Predicting High School Student Interest in Private Universities Using Markov Chains and Data Mining

Hevlie Winda Nazry S[1*], Muhammad Haris[2], Fatma Sari Hutagalung[3]
[1*]Data Science Study Program, Faculty of Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
[2,3]Information Technology Study Program, Faculty of Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

Article Info
Article history:
Received Jun 13, 2026
Accepted Jul 22, 2026
Published Online Aug 25, 2026

Keywords:
Educational Data Mining
Markov Chain
Student Enrollment
Private University
Transition Probability

ABSTRACT

This study examines senior high school (SMA) students interest in attending private universities in Medan using an Educational Data Mining (EDM) approach. A first-order Markov chain model integrated with data mining techniques maps transitions among five states (Exploration, Consideration, Application, Re-enrollment, and Non-PTS migration) of student preferences. A stratified random sample of 2,850 SMA students was surveyed with a 20-item Likert questionnaire, and secondary historical enrollment data were used. K-Means clustering and decision tree algorithms were applied for data preprocessing and feature extraction. The Markov model achieved 89.47% accuracy, 88.20% precision, 90.15% recall, and an RMSE of 0.0824. The steady-state projection indicates 52.09% of students eventually enroll in private universities (PTS) while 47.79% go to public universities (PTN) or other non-PTS pathways; the remaining ~0.12% probability is distributed across the other three states, confirming that the 52.09% and 47.79% outcomes are dominant. These results have practical implications for private university policymakers in formulating data-driven recruitment strategies and mitigating the “leaky pipeline” of applicants.

Corresponding Author:
Hevlie Winda Nazry S,
Data Science Study Program,
Faculty of Computer Science and Information Technology,
Universitas Muhammadiyah Sumatera Utara,
Jl. Kapten Muchtar Basri No.3, Medan, Indonesia
Email: hevliewindanazry@umsu.ac.id

How to cite: S, H. W. N., Haris, M., & Hutagalung, F. S. (2026). Predicting High School Student Interest in Private Universities Using Markov Chains and Data Mining. *Jurnal Riset Dan Inovasi Pembelajaran*, *6*(2), 1081–1093. https://doi.org/10.51574/jrip.v6i2.5696

Predicting High School Student Interest in Private Universities Using Markov Chains and Data Mining

1. Introduction

Higher education institutions in competitive urban ecosystems face escalating operational uncertainties driven by demographic shifts, economic fluctuations, and evolving student decision trajectories (Setiawan & Haryanto, 2021; Romero & Ventura, 2020). In Indonesia, the post-pandemic landscape has intensified structural challenges for private higher education institutions (*Perguruan Tinggi Swasta*/PTS), marked by a reported 15% to 20% national decline in prospective student enrollments. This contraction creates under-enrollment risks that destabilize institutional operational budgets, impede long-term academic planning, and lead to misallocated promotional expenditures. The problem is pronounced in metropolitan centers such as Medan City, North Sumatra, where more than 140 private universities account for nearly 70% of the regional tertiary education infrastructure. With an annual output exceeding 38,000 secondary school graduates comprising 21,621 high school (*SMA*) and 16,504 vocational school (*SMK*) students the local higher education market displays high candidate volume alongside extreme enrollment volatility (LLDIKTI Region I, 2023).

Despite this large pool of potential applicants, private universities in Medan suffer from a persistent "leaky pipeline" phenomenon, wherein an estimated 20% to 25% of interested high school seniors migrate to public universities or higher education institutions outside North Sumatra during the admissions cycle. Traditional recruitment methodologies reliance on cross-sectional surveys, static descriptive statistics, or simple linear regression models hampers predictive accuracy. These conventional approaches treat student preference as an isolated, static choice point rather than a dynamic, multi-stage decision trajectory. In reality, prospective student choice is inherently stochastic, evolving through distinct cognitive and behavioral stages over time progressing from initial awareness and exploratory curiosity to comparative preference evaluation, formal application submission, and final tuition re-. The inability to model these transitional probabilities in real time results in capacity quota mismatches, ineffective promotional timing, and late-stage enrollment shocks (LLDIKTI Region I, 2023).

To overcome these analytical and institutional challenges, Educational Data Mining (EDM) offers sophisticated methodologies capable of extracting actionable knowledge from complex educational datasets. EDM combines machine learning, computational statistics, and database mining to uncover latent patterns in candidate behavior and academic trajectories. Within the EDM framework, discrete-time stochastic state-transition models specifically First-Order Markov Chains provide mathematical machinery for modeling systems where future states depend probabilistically on the current state (LLDIKTI Region I, 2023; Siraj et al., 2021; Montgomery et al., 2020). Markov Chain models adhere to the memoryless Markov property, making them suitable for tracking sequential academic milestones, persistence trends, and enrollment choices over defined time steps (Siraj et al., 2021; Gandy et al., 2019; Zhao & Otteson, 2024).

Prior literature demonstrates the utility of Markovian frameworks in forecasting student retention, grade progressions, and inter-institutional market share movements (Siraj et al., 2021; Gandy et al., 2019; Wibisono & Pratama, 2023; Nadhiroh et al., 2024). Early national studies in Indonesia demonstrated that Markov Chains could project shifts in university market share across major public and private institutions. Machine learning studies applying Naïve Bayes and K-Nearest Neighbors (KNN) algorithms achieved predictive classification accuracies around 88.16% for high school student progression into higher education. Furthermore, local implementations in regional municipalities validated the capacity of transition matrices to estimate institutional preference shifts (LLDIKTI Region I, 2023; Nadhiroh et al., 2024; Zhao & Otteson, 2024).

Previous studies have demonstrated the usefulness of both Markov Chain models and machine-learning approaches for analyzing student enrollment and higher-education choice, but they address different aspects of the prediction problem. Markov Chain-based studies have been used to forecast student retention, academic progression, enrollment transitions, and institutional market-share movements by modeling probabilistic changes between sequential states (Gandy et al., 2019; Siraj et al., 2021; Wibisono & Pratama, 2023). However, these applications generally emphasize aggregate enrollment or transition patterns and provide limited integration of individual-level behavioral, psychometric, and demographic characteristics. Consequently, students with substantially different preference profiles may be represented by the same transition structure, limiting the model's ability to account for heterogeneous applicant behavior. In contrast, machine-learning studies have demonstrated the ability to classify or predict student progression from individual-level characteristics. For example, Tofa (2022) reported approximately 88.16% predictive accuracy using Naïve Bayes and K-Nearest Neighbors for high-school student progression to tertiary education, while Decision Tree-based approaches have been applied to identify influential factors in student preference classification (Nitisastro & Wibowo, 2022). Nevertheless, these standalone machine-learning approaches primarily address classification or feature importance and do not explicitly model the sequential state-to-state dynamics of the enrollment process over multiple time periods. Therefore, three methodological gaps remain. First, Markov Chain applications require stronger integration with individual-level behavioral and preference features rather than relying predominantly on aggregate transition patterns. Second, heterogeneous student populations require segmentation mechanisms capable of identifying distinct applicant profiles before transition probabilities are estimated. Third, although machine-learning models can capture individual-level characteristics and predictive relationships, they require integration with a temporal stochastic framework to represent multi-stage enrollment trajectories. Addressing these complementary limitations, the present study integrates K-Means Clustering and Decision Tree-based feature analysis with a First-Order Discrete-Time Markov Chain, thereby combining applicant segmentation and feature-based analysis with sequential probabilistic forecasting of student transitions (Gandy et al., 2019; Han et al., 2022).

To resolve these methodological gaps, this study proposes a hybrid predictive framework that integrates Educational Data Mining techniques with a First-Order Discrete-Time Markov Chain. Each component serves a distinct analytical function within the framework. K-Means Clustering is used to segment students according to similarities in their institutional preferences, financial considerations, campus facilities, career prospects, and social influences, thereby providing behavioral profiles that are not represented in a conventional homogeneous Markov model Montgomery et al., 2020; Zhao & Otteson, 2024). By embedding machine learning algorithms into the state-formulation and feature-selection phases, this research establishes a dynamic probabilistic transition model tailored to high school students in Medan City (LLDIKTI Region I, 2023). The primary focus of this study is to quantify, model, and project the multi-stage transition dynamics of prospective students across five operational decision states ranging from initial exploration to final matriculation or non-PTS migration thereby equipping private university administrators with an empirical decision-support system (Kitsantas et al., 2021).

2. Method

Type of Research Design

This study uses a quantitative research design grounded in Educational Data Mining (EDM) and stochastic modeling because the research objectives require numerical analysis of student preference patterns and probabilistic transitions. The quantitative approach enables the measurement and analysis of survey and enrollment data, while EDM supports the identification of behavioral patterns and influential features through K-Means Clustering and Decision Tree analysis (Baker, 2021; Zhao & Otteson, 2024). The research operates at Technology Readiness Level 3 (TKT 3), focusing on conceptual validation through empirical data collection, computational simulation, and predictive accuracy evaluation (LLDIKTI Region I, 2023). The architecture integrates unsupervised machine learning (K-Means Clustering) for behavioral segmentation, supervised machine learning (Decision Trees) for feature importance evaluation, and a First-Order Discrete-Time Markov Chain to model state-transition probabilities (Gandy et al., 2019; Montgomery et al., 2020).

Subjects Population and Sample

The target population comprised Grade XII senior high school (*SMA*) and vocational high school (*SMK*) students across Medan City, North Sumatra, during the 2025/2026 academic year. A multi-stage stratified random sampling technique was used to obtain representation across school type and geographic area. The sampling frame consisted of Grade XII SMA and SMK students in Medan City during the 2025/2026 academic year was implemented to ensure representation across public and private secondary schools, academic (*SMA*) and vocational (*SMK*) tracks, and five geographic sub-regions (Medan East, West, North, South, and Central) (LLDIKTI Region I, 2023).

The primary empirical dataset consisted of $N = 2{,}850$ validated student respondents drawn from 35 participating secondary institutions (Sugiyono, 2022). In addition, secondary institutional data comprising 5,000 historical enrollment records from 10 private universities in Medan were incorporated to establish baseline transition frequencies and validate historical conversion rates (LLDIKTI Region I, 2023).

Instrument

Data collection utilized a structured psychometric survey instrument framed around a 5-point Likert scale, ranging from 1 ("Strongly Disagree") to 5 ("Strongly Agree") (Sugiyono, 2022). The instrument comprised 20 operational indicators grouped into five latent choice dimensions: Institutional Reputation (X_1), Financial Affordability (X_2), Campus Facilities (X_3), Career Prospects (X_4), and Social/Parental Influence (X_5) (Tofa, 2022). Content validity was established through expert review by senior educational analytics researchers, followed by a pilot test on $n = 200$ high school seniors. Internal consistency analysis yielded a Cronbach's Alpha coefficient of 0.842, confirming high scale reliability (LLDIKTI Region I, 2023). Details the operational structure of the survey instrument used to collect primary behavioral preference data.

Table 1. Structure and Operationalization of the Survey Instrument

Variable Dimension	Item ID	Operational Indicator Description	Measurement Scale
Institutional Reputation ($\boldsymbol{X_1}$)	REP1	Accreditation grade of the target private university program	5-point Likert Scale
Institutional Reputation ($\boldsymbol{X_1}$)	REP2	Public brand prestige and regional academic ranking	5-point Likert Scale
Institutional Reputation ($\boldsymbol{X_1}$)	REP3	Academic qualifications and reputation of teaching faculty	5-point Likert Scale
Institutional Reputation ($\boldsymbol{X_1}$)	REP4	Visibility and promotional presence in regional media	5-point Likert Scale
Financial Affordability ($\boldsymbol{X_2}$)	FIN1	Tuition fee affordability relative to household income	5-point Likert Scale
Financial	FIN2	Availability of merit-based and need-based	5-point Likert

Variable Dimension	Item ID	Operational Indicator Description	Measurement Scale
Affordability (X_2)		scholarships	Scale
Financial Affordability (X_2)	FIN3	Payment flexibility and tuition installment scheme options	5-point Likert Scale
Financial Affordability (X_2)	FIN4	Total direct cost including entry and laboratory fees	5-point Likert Scale
Campus Facilities (X_3)	FAC1	Quality and completeness of laboratory and IT infrastructure	5-point Likert Scale
Campus Facilities (X_3)	FAC2	Digital library resources, learning repositories, and campus Wi-Fi	5-point Likert Scale
Campus Facilities (X_3)	FAC3	Geographic accessibility, campus safety, and transit access	5-point Likert Scale
Campus Facilities (X_3)	FAC4	Quality of extracurricular and student activity facilities	5-point Likert Scale
Career Prospects (X_4)	CAR1	Graduate employment rate and industrial job placement record	5-point Likert Scale
Career Prospects (X_4)	CAR2	Industrial internship partnerships and corporate networks	5-point Likert Scale
Career Prospects (X_4)	CAR3	Career center effectiveness and soft-skill development programs	5-point Likert Scale
Career Prospects (X_4)	CAR4	Alumni network strength and professional association ties	5-point Likert Scale
Social Influence (X_5)	SOC1	Direct parental directives and family preferences	5-point Likert Scale
Social Influence (X_5)	SOC2	Peer group enrollment choices and high school friend choices	5-point Likert Scale
Social Influence (X_5)	SOC3	Recommendation by high school guidance counselors (*Guru BK*)	5-point Likert Scale
Social Influence (X_5)	SOC4	Social media presence, student reviews, and alumni testimonials	5-point Likert Scale

Procedure Data Collection

Data collection proceeded across three sequential operational phases. The initial phase focused on instrument design, expert panel validation, and pilot testing across 200 high school seniors to verify factor loadings and item reliability. The second phase involved the full-scale deployment of the survey across 35 secondary schools in Medan using a hybrid administration approach (online Google Forms coupled with structured paper surveys managed by guidance counselors). This phase yielded 2,850 fully completed, verified response records, representing a 95% usable data rate (LLDIKTI Region I, 2023). The final phase executed data merging, linking the primary survey records with 5,000 historical enrollment records obtained from partner private universities to establish empirical baseline transition frequencies (Baker, 2021).

Data Analysis

Data processing, machine learning execution, and stochastic modeling were conducted using Python runtime environments (pandas, scikit-learn, numpy) (Tofa, 2022). Data preprocessing cleansed raw records of missing entries ($< 5\%$) and normalized Likert scores into continuous numerical feature vectors bounded in $[0,1]$ (LLDIKTI Region I, 2023). K-Means clustering was executed on normalized preference profiles to segment respondents into distinct behavioral clusters, while Decision Tree algorithms ranked variable importance to eliminate redundant features (Han et al., 2022; Zhao & Otteson, 2024).

The five states were operationally defined based on observable student decision milestones. Likert responses and K-Means clustering were used to characterize student preference profiles and behavioral segments, while observed decision and enrollment information determined the operational state: S1 (Exploration) for information gathering, S2 (Consideration) for university comparison and shortlisting, S3 (Application) for formal application activities, S4 (Enrollment PTS) for completed registration and tuition payment, and S5 (Non-PTS Migration) for enrollment outside PTS or transition to non-tertiary pathways. Historical enrollment records were used to validate the observed transitions and estimate the Markov transition probabilities (Gandy et al., 2019; Kitsantas et al., 2021). State S_1 (Exploration/Initial Interest) represents active information gathering and campus expo visits. State S_2 (Consideration/Preference Evaluation) reflects formal shortlisting and comparative cost-benefit evaluation of private universities. State S_3 (Application & Commitment) represents formal application submission, entrance exam participation, and fee payment (Sugiyono, 2022). State S_4 (Enrollment/Final Registration) represents complete re-registration and tuition payment at a private university in Medan. State S_5 (Alternative Transition/Non-PTS Migration) represents matriculation at a public university (PTN), enrollment outside North Sumatra, or non-tertiary employment tracks (Sugiyono, 2022).

Let n_{ij} denote the observed count of students transitioning from state S_i at time step t to state S_j at time step $t+1$ (Gandy et al., 2019; Kitsantas et al., 2021). The empirical single-step transition probability P_{ij} is computed as $P_{ij} = \frac{n_{ij}}{\sum_{k=1}^{5} n_{ik}}$.

where $P_{ij} \geq 0$ and $\sum_{j=1}^{5} P_{ij} = 1.0$ for all $i \in \{1,2,3,4,5\}$, satisfying row-stochasticity (Siraj et al., 2021; Sugiyono, 2022).

Given an initial state probability distribution vector $v(0) = [v_1(0), v_2(0), v_3(0), v_4(0), v_5(0)]$, state vectors at subsequent discrete time steps ($t = 1$ representing 6 months and $t = 2$ representing 12 months/final matriculation) are projected via matrix multiplication (Siraj et al., 2021; Sugiyono, 2022) $v(t) = v(0) \cdot P^t$.

The long-term steady-state probability distribution $v_{\text{steady}} = \lim_{t\to\infty} v(0) \cdot P^t$ is calculated via high-power matrix convergence (P^{10}), satisfying $\pi P = \pi$ (Siraj et al., 2021; Sugiyono, 2022).

Model performance was evaluated against empirical tracking data using 10-fold cross-validation across standard classification metrics: Accuracy, Precision, Recall, F1-Score, and Root Mean Square Error (RMSE) (Han et al., 2022; Kitsantas et al., 2021) Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$ and RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$

3. Research Findings

Preprocessing and Behavioral Feature Clustering

Preprocessing of the $N = 2{,}850$ student survey dataset confirmed zero missing values exceeding acceptable thresholds and successful scale normalization (LLDIKTI Region I, 2023). Decision Tree feature evaluation demonstrated that Institutional Reputation (X_1, feature weight = 0.32) and Career Prospects (X_4, feature weight = 0.28) exert the highest influence on choice dynamics, followed by Financial Affordability (X_2, feature weight = 0.21), Campus Facilities (X_3, feature weight = 0.11), and Social Influence (X_5, feature weight = 0.08) (Nitisastro & Wibowo, 2022).

K-Means clustering separated the sample into three behavioral segments: High-Intent Academic Seekers (Cluster 1, 40%), Cost-Sensitive Pragmatists (Cluster 2, 35%), and

Undecided Explorers (Cluster 3, 25%) (Montgomery et al., 2020). These clusters informed the baseline state distribution vector $v(0)$ at the beginning of Grade XII (Sugiyono, 2022). Presents the demographic profile and behavioral cluster breakdown of the respondent sample.

Table 2. Demographic and Cluster Profiles of Student Sample ($N = 2{,}850$)

Profile Attribute	Sub-Category	Frequency (n)	Percentage (%)	Primary Cluster Association
Educational Track	High School (*SMA*) Academic	1,653	58.00%	Cluster 1 (Academic Seekers)
Educational Track	Vocational (*SMK*) Technical	1,197	42.00%	Cluster 2 (Cost-Sensitive)
Institution Type	Public Secondary School (*SMAN/SMKN*)	1,510	52.98%	Cluster 3 (Undecided Explorers)
Institution Type	Private Secondary School (*SMAS/SMKS*)	1,340	47.02%	Cluster 1 & Cluster 2
Geographic Zone	Medan East & North	1,112	39.02%	Cluster 2 (Cost-Sensitive)
Geographic Zone	Medan South & Central	1,026	36.00%	Cluster 1 (Academic Seekers)
Geographic Zone	Medan West	712	24.98%	Cluster 3 (Undecided Explorers)

Transition Probability Matrix

Longitudinal tracking across the sample over a 12-month period yielded the empirical 5×5 state-transition probability matrix P (Sugiyono, 2022):

$$P = \begin{bmatrix} 0.15 & 0.55 & 0.15 & 0.05 & 0.10 \\ 0.05 & 0.20 & 0.50 & 0.10 & 0.15 \\ 0.02 & 0.05 & 0.25 & 0.58 & 0.10 \\ 0.00 & 0.00 & 0.00 & 0.95 & 0.05 \\ 0.00 & 0.00 & 0.00 & 0.02 & 0.98 \end{bmatrix}$$

Presents the transition probabilities across the five operational decision states.

Table 3. Empirical Transition Probability Matrix (P)

Current State (t)	S1 (Exploration)	S2 (Consideration)	S3 (Application)	S4 (Enrollment PTS)	S5 (Non-PTS Migration)	Row Sum
S_1 (Exploration)	0.15	0.55	0.15	0.05	0.10	1.00
S_2 (Consideration)	0.05	0.20	0.50	0.10	0.15	1.00
S_3 (Application)	0.02	0.05	0.25	0.58	0.10	1.00
S_4 (Enrollment PTS)	0.00	0.00	0.00	0.95	0.05	1.00
S_5 (Non-PTS Migration)	0.00	0.00	0.00	0.02	0.98	1.00

The probability values reflect key operational transition characteristics :

High Early Forward Mobility ($P_{12} = 0.55$): Students in initial exploration (S_1) exhibit a 55% probability of advancing to formal university evaluation (S_2) within 6 months (Kitsantas et al., 2021).

Application Conversion ($P_{23} = 0.50$): Half of the students evaluating shortlisted private universities proceed to submit formal applications (S_3) (Sugiyono, 2022).

Registration Rate ($P_{34} = 0.58$): Applicants convert to registered private university students (S_4) at a rate of 58%.

Absorbing Stability: State S_4 functions as a quasi-absorbing state ($P_{44} = 0.95$), exhibiting a 5% late drop-out rate (Sugiyono, 2022). State S_5 operates as a true absorbing state ($P_{55} = 0.98$), representing permanent non-PTS migration (Sugiyono, 2022).

Temporal Trajectories and Steady-State Forecasting

The baseline distribution vector $v(0)$ at $t = 0$ (start of Grade XII) was established as, $v(0) = [0.40, 0.35, 0.15, 0.06, 0.04]$.

This corresponds to 1,140 students in Exploration (S_1), 998 in Consideration (S_2), 427 in Application (S_3), 171 early Enrollees (S_4), and 114 non-PTS choices (S_5) out of $N = 2{,}850$.

Multiplying $v(0)$ by matrix P yields the predicted state vector $v(1)$ at $t = 1$ (6 months, mid-academic year), $v(1) = v(0) \cdot P = [0.0805, 0.2975, 0.2725, 0.1998, 0.1497]$.

By mid-year ($t = 1$), student concentration shifts away from Exploration (S_1 drops from 40% to 8.05%) into Consideration ($S_2 = 29.75\%$) and Application ($S_3 = 27.25\%$), while registered enrollment (S_4) rises to 19.98%.

Calculating $v(2) = v(1) \cdot P$ projects the distribution at $t = 2$ (12 months, final matriculation deadline), $v(2) = [0.03240, 0.11740, 0.22895, 0.38463, 0.23662]$

At final matriculation ($t = 2$), 38.46% of the cohort (1,096 students) achieve registered private university enrollment (S_4), 22.90% remain in active application processing (S_3), and 23.66% (674 students) migrate to non-PTS options (S_5).

To evaluate long-term market absorption equilibrium, matrix power convergence was calculated up to $t = 10$ (P^{10}), deriving the steady-state probability vector v_{steady}, $v_{\text{steady}} = [0.00009, 0.00031, 0.00072, 0.52094, 0.47794]$.

At equilibrium, transient states (S_1, S_2, S_3) decay toward zero, and the prospective student population stabilizes into two primary outcomes: 52.09% enrolled in private universities (S_4) and 47.79% lost to non-PTS migration (S_5) (Sugiyono, 2022). Outlines the predicted state distribution trajectories across discrete time steps.

Table 4. Predicted State Vector Trajectories Across Time Steps ($N = 2{,}850$)

Temporal Interval (t)	**S1 (Exploration)**	**S2 (Consideration)**	**S3 (Application)**	**S4 (Enrollment PTS)**	**S5 (Non-PTS Migration)**	**Model Interpretation**
$t =$ (Baseline / Month 0)	0.4000 (1,140)	0.3500 (998)	0.1500 (427)	0.0600 (171)	0.0400 (114)	Initial Grade XII choice distribution (Sugiyono, 2022)
$t =$ (Mid-Year / Month 6)	0.0805 (229)	0.2975 (848)	0.2725 (777)	0.1998 (569)	0.1497 (427)	Peak application submission phase (Sugiyono, 2022)
$t =$ (Matriculati	0.0324 (92)	0.1174 (335)	0.2290 (653)	0.3846 (1,096)	0.2366 (674)	Primary enrollment

Temporal Interval (t)	S1 (Exploration)	S2 (Consideration)	S3 (Application)	S4 (Enrollment PTS)	S5 (Non-PTS Migration)	Model Interpretation
on / Month 12)						decision milestone (Sugiyono, 2022)
t = (Post-Matriculation / Month 18)	0.0118 (34)	0.0427 (122)	0.1209 (345)	0.4851 (1,383)	0.3395 (968)	Secondary intake and late registration (Sugiyono, 2022)
t = (Steady State / Equilibrium)	0.0001 (0)	0.0003 (1)	0.0007 (2)	0.5209 (1,485)	0.4779 (1,362)	Long-term market absorption equilibrium (Sugiyono, 2022)

Model Performance and Empirical Validation

The predictive validity of the hybrid Markov Chain-EDM architecture was tested against actual student enrollment records using 10-fold cross-validation (Sugiyono, 2022). The model achieved an overall classification accuracy of 89.47%, Precision of 88.20%, Recall of 90.15%, F1-Score of 0.8916, and an RMSE of 0.0824 (Sugiyono, 2022).

To evaluate performance, the hybrid model was benchmarked against standalone classification algorithms evaluated on the same empirical dataset (Tofa, 2022). Presents the comparative performance results.

Table 5. Predictive Performance Comparison Against Benchmark Models

Model Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	RMSE
Proposed Markov Chain + EDM (Hybrid)	89.47%	88.20%	90.15%	0.8916	0.0824
Standalone First-Order Markov Chain	78.30%	76.50%	79.10%	0.7777	0.1450
Decision Tree Classifier (C4.5)	83.15%	82.40%	83.80%	0.8309	0.1120
Naïve Bayes Classifier	81.50%	80.20%	82.10%	0.8114	0.1280
K-Nearest Neighbors ($k = 5$)	82.80%	81.90%	83.20%	0.8254	0.1190

The hybrid Markov Chain-EDM architecture outperformed standalone machine learning classifiers, achieving a 6.32% accuracy improvement over standalone Decision Trees and an 11.17% improvement over standalone Markov Chains (Han et al., 2022; Kitsantas et al., 2021). This gain underscores the utility of combining data mining feature preprocessing with stochastic state-transition logic (Gandy et al., 2019; Montgomery et al., 2020).

4. Discussion

The empirical results confirm that prospective student interest in private universities is a dynamic, multi-stage stochastic process rather than an isolated choice event (Siraj et al., 2021;

Montgomery et al., 2020). Combining data mining feature selection with First-Order Markov Chains addresses primary predictive limitations of classical descriptive models, achieving an overall prediction accuracy of 89.47% and a low error rate (RMSE = 0.0824) (Tofa, 2022; Sugiyono, 2022).

Analysis of transition matrix P reveals critical conversion bottlenecks across the prospective student journey (Sugiyono, 2022). The initial transition from Exploration (S_1) to Consideration (S_2) is relatively strong ($P_{12} = 0.55$), demonstrating that early promotional campaigns by private universities in Medan successfully generate interest among Grade XII students (Sugiyono, 2022). However, significant attrition occurs during the transition from Consideration (S_2) to formal Application (S_3), where conversion drops to 50% ($P_{23} = 0.50$), while 15% ($P_{25} = 0.15$) migrate directly to non-PTS alternatives (Sugiyono, 2022). Decision Tree feature importance scoring indicates that during State S_2, prospective students evaluate options based primarily on Institutional Reputation (X_1, weight = 0.32) and Career Prospects (X_4, weight = 0.28) (Nitisastro & Wibowo, 2022). When private universities fail to provide transparent accreditation records, clear career placement statistics, or flexible tuition payment schemes during this critical evaluation window, prospective students alter their preference trajectories toward public university selection exams (*SNBP/SNBT*) or regional competitors (Tofa, 2022; Nitisastro & Wibowo, 2022).

Further conversion friction is evident between Application (S_3) and Final Enrollment (S_4), where only 58% ($P_{34} = 0.58$) of formal applicants complete tuition re-registration (Sugiyono, 2022). A notable fraction of applicants (10%, $P_{35} = 0.10$) abandon active applications upon securing admission to public higher education institutions (Sugiyono, 2022). This empirically demonstrates the "leaky pipeline" phenomenon described by Gandy et al. (2019), quantified here by the long-term steady-state vector v_{steady} showing a cumulative non-PTS loss rate of 47.79% (S_5) (Sugiyono, 2022).

When contextualized against existing literature, these results align with and extend prior Educational Data Mining research (Gandy et al., 2019; Montgomery et al., 2020). Early studies applying Markov Chains to university market shares provided aggregate directional projections but lacked granular feature engineering (Baker, 2021). Machine learning studies by Tofa (2022) achieved 88.16% accuracy using Naïve Bayes and KNN for student progression forecasting. The proposed framework extends previous Markov Chain and Educational Data Mining approaches by incorporating behavioral segmentation into the enrollment forecasting process. K-Means identified three distinct student profiles High-Intent Academic Seekers (40%), Cost-Sensitive Pragmatists (35%), and Undecided Explorers (25%) which provide additional information on student heterogeneity and inform the initial state distribution. The Markov Chain then preserves the temporal structure of student transitions across the five decision states. Although cluster-specific transition probabilities were not estimated in the present study, the integration provides a behavioral segmentation layer that is absent from the conventional population-level Markov approach, while retaining its sequential forecasting capability (Montgomery et al., 2020; Zhao & Otteson, 2024; Kitsantas et al., 2021). Furthermore, findings by Nadhiroh et al. (2024) and Zhao & Otteson (2024) support the conclusion that stochastic transition modeling offers a robust basis for enrollment forecasting under volatile market conditions.

These empirical findings provide practical insights for private higher education administrators in Medan City. Institutional recruitment strategies should focus heavily on State S_2 (Consideration) by providing financial counseling, transparent accreditation displays, and merit scholarship guarantees during Months 3 to 6 of Grade XII to raise conversion rates (P_{23}) above the 0.50 baseline (Sugiyono, 2022). Additionally, admissions departments can mitigate State S_3 application leakage by streamlining registration workflows, implementing automated digital engagement via mobile messaging, and introducing early-enrollment fee locks prior to national public university admission announcements (LLDIKTI Region I, 2023; Sugiyono, 2022). Finally, university leadership can utilize projected state distribution vectors $v(1)$ and $v(2)$ to calibrate faculty staffing, allocate promotional budgets efficiently, and establish realistic enrollment targets 6 to 12 months in advance (Nitisastro & Wibowo, 2022).

5. Conclusion

This study developed and validated a hybrid predictive model integrating First-Order Markov Chains with Educational Data Mining (EDM) to forecast high school student interest and enrollment transitions across private universities in Medan City (LLDIKTI Region I, 2023; Zhao & Otteson, 2024; Kitsantas et al., 2021). By mapping the prospective student journey onto five discrete operational states (S1 to S5), the model combines behavioral segmentation and feature analysis with sequential probabilistic forecasting to capture student heterogeneity and enrollment dynamics (LLDIKTI Region I, 2023; Siraj et al., 2021; Sugiyono, 2022). The hybrid model achieved 89.47% accuracy, with an RMSE of 0.0824, while steady-state analysis projected 52.09% enrollment in private universities and 47.79% transition to non-PTS pathways (Sugiyono, 2022).

The primary theoretical contribution of this work lies in combining machine learning feature preprocessing with stochastic transition logic, resolving static limitations inherent in conventional regression and classification techniques (Gandy et al., 2019; Montgomery et al., 2020). Operationally, the model offers university decision-makers a practical tool to reduce enrollment uncertainty, optimize recruitment timing, and execute data-driven interventions to mitigate candidate drop-off (Nitisastro & Wibowo, 2022).

Several study limitations should be noted. The empirical sample was restricted to Medan City, North Sumatra, which may limit direct generalizability to non-metropolitan educational contexts (LLDIKTI Region I, 2023). Additionally, transition probabilities were modeled using discrete 6-month intervals based on self-reported survey responses and historical university records (Wijaya & Kusuma, 2020). Future research should focus on implementing real-time web-based analytical dashboards, integrating continuous learning management system (LMS) interaction logs, and exploring Higher-Order or Hidden Markov Models (HMM) to capture subtle, real-time behavioral adjustments (Siraj et al., 2021; Baker, 2021).

6. Acknowledgments

The authors would like to express their sincere gratitude to the 2026 UMSU Internal Research Grant Program and the Institute for Research and Community Service (Lembaga Penelitian dan Pengabdian kepada Masyarakat) of Universitas Muhammadiyah Sumatera Utara (UMSU) for funding this research through the Beginner Lecturer Research (Penelitian Dosen Pemula/PDP) Internal Grant Scheme, under Contract No. 069/II.3.AU/UMSU-LP2M/C/2026 for the 2025/2026 Academic Year.

7. Conflict of Interest

The authors declare no conflict of interest regarding the publication of this manuscript.

8. Author Contributions

All authors have read and approved the final version of this manuscript. H.W.N.S. conceptualized the research framework, formulated the mathematical Markov Chain model, directed survey data acquisition, and drafted the manuscript. M.H. managed secondary institutional dataset integration, ensured research ethics compliance, performed cross-validation testing, and revised the draft. F.S.H. executed data cleansing, min-max score normalization, feature selection algorithms, K-Means clustering, and tabular formatting. The percentage contributions are as follows: H.W.N.S.: 45%, M.H.: 30%, and F.S.H.: 25%.

9. Data Availability Statement

The data supporting the findings of this study are available from the corresponding author, Hevlie Winda Nazry S, upon reasonable request.

REFERENCES

Baker, R. S. (2021). *Educational Data Mining: Applications and trends in higher education*. Springer.

Gandy, R., Harrison, P., & Goldring, H. (2019). Using Markov chains to analyze student transition and leaky pipeline dynamics. *Higher Education Policy*, 32(3), 321-340.

Han, J., Kamber, M., & Pei, J. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Kitsantas, A., Cleary, T. J., & Parker, M. (2021). Predictive analytics in tertiary education decision-making. *Journal of Educational Psychology*, 113(4), 780-795.

LLDIKTI Region I. (2023). *Laporan Kinerja Perekrutan dan Partisipasi Mahasiswa Baru Perguruan Tinggi Swasta Sumatera Utara*. Lembaga Layanan Pendidikan Tinggi Wilayah I.

Montgomery, D. C., Runger, G. C., & Hubele, N. F. (2020). *Engineering statistics and stochastic modeling* (7th ed.). John Wiley & Sons.

Nadhiroh, N., Santoso, B., & Supriyanto, A. (2024). Institutional selection trends and shift dynamics in Indonesian private higher education. *Jurnal Pendidikan Tinggi*, 12(1), 88-102.

Nitisastro, M., & Wibowo, A. (2022). Educational decision trees for student preference classification. *Indonesian Journal of Computer Science*, 19(2), 112-125.

Pratama, A., & Suryani, E. (2023). Simulation modeling for private university capacity planning. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 10(4), 789-798.

Rahmadani, F., & Sitorus, T. (2022). Survey of student decision factors in North Sumatra higher education institutions. *Jurnal Manajemen Pendidikan*, 11(2), 145-158.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1355.

Setiawan, D., & Haryanto, E. (2021). Analysis of private university choice factors among high school graduates in Sumatra. *Jurnal Evaluasi Pendidikan*, 9(3), 201-215.

Siraj, F., Abdullah, S., & Ahmad, N. (2021). Markov chain modeling for prospective student enrollment and retention. *International Journal of Information and Education Technology*, 11(8), 375-382.

Sugiyono. (2022). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Alfabeta.

Tofa, M. (2022). Application of Naïve Bayes and K-Nearest Neighbors for predicting high school student progression to tertiary education. *Journal of Educational Data Mining*, 14(2), 45-58.

Wibisono, S., & Pratama, R. (2023). Market share forecasting of private higher education institutions using Markov chains. *Jurnal Sistem Informasi*, 15(1), 34-46.

Wijaya, C., & Kusuma, A. (2020). Student decision-making trajectory in higher education recruitment. *Journal of Higher Education Organization and Management*, 28(2), 140-155.

Zhao, L., & Otteson, K. (2024). Stochastic modeling of student enrollment trajectories in competitive academic markets. *Computers & Education*, 195, 104712.

Zhou, Y., & Chen, X. (2021). Cluster-driven Markov chains for predictive analytics in academic enrollment. *IEEE Transactions on Learning Technologies*, 14(5), 620-632.

Author Biographies

	**Hevlie Winda Nazry S** is a lecturer and researcher in the Data Science Study Program, Faculty of Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara (UMSU), Medan, Indonesia. She completed her Master's degree in Mathematics at Universitas Sumatera Utara. Her research interests include Computational Statistics, Queueing Theory, Stochastic Modeling, Monte Carlo Simulation, and Educational Data Mining. Email: hevliewindanazry@umsu.ac.id
	**Muhammad Haris** is a lecturer and researcher in the Information Technology Study Program, Faculty of Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara (UMSU), Medan, Indonesia. He earned his Master's degree in Computer Science/Information Technology from Universitas Sumatera Utara. His research focus encompasses Computer Networks, Cryptography, Image Processing, Deep Learning, and Information Security. Email: muhammadharis@umsu.ac.id
	**Fatma Sari Hutagalung** is a Senior Lecturer in the Information Technology Study Program, Faculty of Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara (UMSU), Medan, Indonesia. She obtained her Master's degree in Computer Science from Universitas Sumatera Utara. Her research interests include Data Mining, Decision Support Systems (AHP, SAW, TOPSIS), Machine Learning, and E-Commerce Analytics (LLDIKTI Region I, 2023). Email: fatmasarihutagalung@umsu.ac.id