




Predicting Student Academic Success with Hidden Markov Models

Veljko Lončarević^{1*} ^[0009-0007-4296-2709], Vučelja Lekić¹ ^[0000-0003-3848-0604] and
Nada Damljanović¹ ^[0000-0001-7133-2852]
¹ University of Kragujevac, Faculty of Technical Sciences, Čačak, Serbia
* veljkoloncnevicharry@gmail.com

Abstract: This research paper presents an approach for predicting student academic success using Hidden Markov Models (HMMs). Leveraging a comprehensive dataset encompassing students' demographics, academic performance, attendance records, and course engagement, the study employs an HMM framework to model levels of student academic success. Observable emissions derived from the data, such as grades and interaction patterns, are utilized to train the HMM and infer the most likely sequence of hidden states for new students. Evaluation of the proposed model demonstrates promising predictive accuracy. Through rigorous assessment using standard metrics including state prediction accuracy and state transition accuracy, the effectiveness of the HMM in capturing diverse student trajectories is demonstrated, underscoring the potential of HMMs as a powerful tool for understanding and predicting student outcomes, offering valuable insights for educational interventions and support systems.

Keywords: *Hidden Markov Models; academic success prediction; student trajectories; predictive modeling; educational data analysis*

1. INTRODUCTION

Predicting student academic success is a critical challenge in educational research, with significant implications for targeted interventions and resource allocation. This research focuses on using Hidden Markov Models (HMMs) to analyze and predict student academic trajectories by incorporating various dimensions of student data. Hidden Markov Models, well-regarded for their ability to model time series data and capture latent state transitions, offer a robust framework for understanding the dynamic nature of student performance over time.

The study integrates a diverse dataset, including demographic information, academic performance metrics, attendance records, and patterns of course engagement. These features are utilized to develop a model that can identify and infer latent states of overall academic achievement. Observable emissions, such as grades and interaction frequencies, are employed to train the HMM, enabling it to predict the sequence of hidden states that most likely represent student behaviors and outcomes.

By applying HMMs to educational data, this research aims to uncover insights into the factors driving academic success and challenges. This approach facilitates the identification of at-risk students and the development of tailored support strategies, enhancing educational interventions

and contributing to improved student retention and success rates.

2. HIDDEN MARKOV MODELS

Hidden Markov Models (HMMs) are statistical models used to describe systems that are assumed to be Markov processes with hidden states. They are particularly useful for modeling time series data where the system being modeled is not directly observable (hidden) but can be inferred through observable sequences. An HMM is characterized by the following components [1, 2]:

- **States (S):** A finite set of hidden states $S = \{S_1, S_2, \dots, S_N\}$. The actual state at time t is denoted as S_t , which is not directly observable.
- **Observations (O):** A finite set of possible observations $O = \{O_1, O_2, \dots, O_M\}$. At any time t , an observation O_t is made, which is dependent on the current hidden state S_t .
- **Transition Probabilities (A):** A matrix $A = [a_{ij}]$ representing the probabilities of transitioning from one state to another. Specifically, a_{ij} is the probability of transitioning from state S_i to state S_j :

$$a_{ij} = P(S_{t+1} = S_j | S_t = S_i) \quad (1)$$

The rows of A must sum to 1:

$$\sum_{j=1}^N a_{ij} = 1 \quad \forall i \quad (2)$$

- **Emission Probabilities (B):** A matrix $B = [b_j(o)]$ representing the probability of observing o given state S_j . $b_j(o)$ is the probability of observation o being emitted from

state S_j :

$$b_j(o) = P(O_t = o | S_t = S_j) \quad (3)$$

For discrete observations, each row of B must sum to 1:

$$\sum_{o \in O} b_j(o) = 1 \quad \forall j \quad (4)$$

- **Initial State Probabilities (π):** A vector $\pi = [\pi_i]$ representing the probability distribution over the initial states. π_i is the probability that the system starts in state S_i :

$$\pi_i = P(S_1 = S_i) \quad (5)$$

The probabilities must sum to 1:

$$\sum_{i=1}^N \pi_i = 1 \quad (6)$$

An HMM is often denoted by the triple $\lambda = (\pi, A, B)$. The **Markov property** of an HMM implies that the probability of transitioning to the next state depends only on the current state and not on the sequence of states that preceded it:

$$P(S_{t+1} | S_1, S_2, \dots, S_t) = P(S_{t+1} | S_t) \quad (7)$$

Given the current state, the probability of an observation depends only on that state and is independent of previous observations:

$$P(O_t | S_1, S_2, \dots, S_t, O_1, O_2, \dots, O_t) = P(O_t | S_t) \quad (8)$$

3. METHODOLOGY

3.1. Data Preparation

The dataset utilized in this study was meticulously compiled from multiple sources, ensuring a comprehensive and diverse representation of student data [3-6]. These sources included institutional academic records, online learning platforms, student information systems, and educational surveys. The integrated dataset encapsulated a wide range of student characteristics and behaviors, crucial for modeling academic success. Key features extracted from these sources, as detailed in Table 1, encompass various aspects of academic performance, socio-demographic attributes, and behavioral indicators. The resulting dataset comprises time-series data of student characteristics tracked and recorded for each semester, providing a dynamic view of their academic progression and behavior over time.

Once the dataset was compiled, the preprocessing phase began with encoding categorical features. Categorical features such as Gender, Parental Education Level and Degree Type were transformed using One-Hot Encoding. This method converts categorical variables into a binary matrix, where each unique category is represented by a separate column, and the presence of a category is indicated by a '1' while its absence is indicated by a '0'. For example, the Degree Type feature, which could take values such as "B.Sc.", "B.A.", or "M.Sc.", was expanded into multiple binary columns, each representing one of these categories. For numerical features, such as Current GPA, Current Semester, and Days Since Enrollment, scaling was performed using the MinMax scaler. This approach scales each numerical feature to a range between 0 and 1,

based on the minimum and maximum values of that feature. This normalization ensures that all numerical features contribute equally to the model and prevents features with larger ranges from disproportionately influencing the model's predictions.

Handling missing data was also a critical aspect of data preparation. Features with a small percentage of missing values were imputed using statistical technique of mean imputation, ensuring that these gaps did not affect the model's performance. However, features with a high percentage of missing values were removed from the dataset to maintain data integrity and model accuracy. This systematic approach to data cleaning and preparation ensured that the dataset was robust, reliable, and ready for subsequent modeling processes.

Table 1. Features in the gathered dataset

Feature	Description
Age	The student's age.
Gender	The student's gender.
Semester	The academic term or semester the student is currently enrolled in.
Degree Type	The specific academic degree program or major the student is enrolled in, such as Bachelor of Science (B.Sc.), Master of Science (M.Sc.), etc.
Days Since Enrollment	The number of days that have elapsed since the student's initial enrollment date.
Current GPA	Provides a continuous measure of the student's academic performance averaged across all courses for the current term.
Class Attendance Rate	Percentage of classes attended out of the total scheduled classes.
Weekly E-Learning Platform Logins	Number of logins to the online course platform per week.
Parental Education Level	Highest educational attainment of the student's parents or guardians (e.g., high school, college).
Assignment Submission Rate	Percentage of assignments submitted on time in each course.
Number of Courses Enrolled	Total number of courses the student is enrolled in during the current term or semester.
Library Visits per Month	Number of visits to the library per month for academic purposes.

For this research paper's data preprocessing, the Python programming language was utilized alongside its robust libraries: pandas, numpy, and scikit-learn. Pandas facilitated the efficient handling and manipulation of the dataset, allowing for seamless integration and transformation of data from multiple sources. Numpy provided essential support for numerical operations and array management, critical for statistical calculations.

Scikit-learn offered powerful tools for encoding categorical features using One-Hot Encoding, scaling numerical data with the MinMax scaler, and managing missing values through the mean imputation technique.

3.2. Defining Hidden States

In the process of creating a Hidden Markov Model (HMM) for predicting student academic success, the definition of hidden states was undertaken to capture the latent conditions that influence observable academic outcomes. These hidden states represent unobservable factors that significantly affect students' academic trajectories but are not directly measurable through the dataset. The definition of hidden states was informed by a combination of domain knowledge, research objectives, and the nature of the available data.

The identification of latent variables was a crucial step in defining the hidden states. It was essential to represent underlying factors that influence observable features such as grades, attendance, and engagement. Hidden states defined in this study are presented in Table 2.

Table 2. Hidden States

Hidden State	Description
At Risk of Drop-Out	This state characterizes students who exhibit poor academic performance, low attendance, and high stress levels. Such students are identified as being at a high risk of discontinuing their studies. This state captures patterns of disengagement and underperformance.
On Track	Students in this state are those performing satisfactorily, maintaining average grades, and consistent attendance. This state indicates a normal progression through their academic program without significant issues.
Excellent	This state represents students who excel academically, demonstrating high grades, strong attendance, and active participation in academic activities. These students are identified as high achievers likely to succeed.

The hidden states were defined with the dual objectives of predictive accuracy and providing actionable insights for educational interventions. States were chosen to meaningfully capture variations in student trajectories relevant to predictive goals. It was imperative that the defined states facilitate the identification of students at risk and those excelling, thereby allowing for targeted interventions to support struggling students or enhance the performance of high achievers. The state transition diagram is shown on Fig. 1.

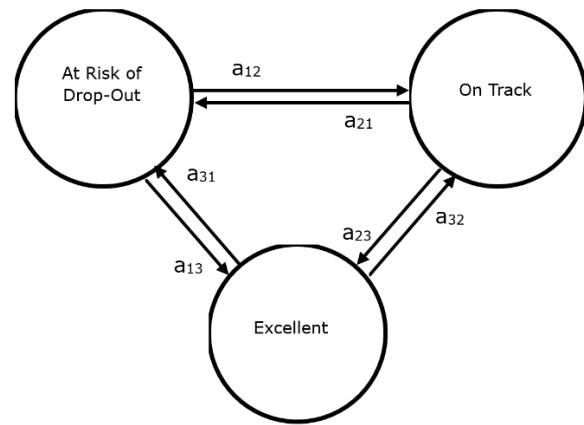


Figure 1. State Transition Diagram

3.3. Clustering and Labeling the Dataset

The initial dataset compiled for this study did not include pre-assigned labels indicating the academic success or risk levels of students. Consequently, an unsupervised learning approach was employed to categorize the data into distinct clusters. The K-Means clustering algorithm was selected for this task due to its effectiveness in partitioning data based on inherent similarities. K-Means was applied to the dataset, aiming to group student data points into three clusters, each representing different levels of academic engagement and performance.

The clustering process involved analyzing a multidimensional feature space comprising the student. The K-Means algorithm iteratively adjusted the cluster centroids to minimize within-cluster variance, effectively grouping students with similar academic profiles. After convergence, three distinct clusters were identified, each capturing unique patterns in the student data. These clusters were then subjected to further analysis to interpret their academic implications.

Through detailed examination of the clusters' characteristics, descriptive labels were assigned to each cluster based on the observed data patterns. One cluster, characterized by low grades, poor attendance, and high stress levels, was labeled as "At Risk of Drop-Out", reflecting students who are likely to struggle academically and potentially discontinue their studies. A second cluster, exhibiting average academic performance and consistent attendance, was labeled as "On Track", indicating students who are progressing normally without significant issues. The third cluster, marked by high grades, strong attendance, and active engagement, was labeled as "Excellent", representing students who are thriving academically. This labeling process transformed the unsupervised clusters into meaningful categories, enabling the subsequent training of the Hidden Markov Model with these inferred state labels. Table 3 presents the summary of key statistics (Average Current GPA, Average Class Attendance Rate (%), Average Weekly E-Learning Platform

Logins, Average Assignment Submission Rate) for each cluster. This information highlights distinct patterns and potential intervention points for each group. Additionally, this table aids in identifying characteristic profiles for each cluster, offering clear benchmarks to assess student progress and tailor support strategies effectively.

Table 3. Key Cluster Statistics

Cluster	GPA	CAR	EPL	ASR
At Risk of Drop-Out	2.1	58.4	2.1	54.2
On Track	3.0	76.0	7.3	86.2
Excellent	3.6	92.7	11.2	97.7

In Table 3 GPA represents Average Current GPA, CAR represents Average Class Attendance Rate (%), EPL represents Average Weekly E-Learning Platform Logins, and ASR represents Average Assignment Submission Rate.

3.4. Calculating HMM Input Values

Training a Hidden Markov Model (HMM) involves estimating the model parameters, namely the initial state probabilities, transition probabilities, and emission probabilities, from a given dataset.

To calculate **initial probabilities** for an HMM, it is needed to determine the probabilities of starting in each of the hidden states. These initial probabilities represent the likelihood of being in a particular state at the beginning of the sequence. It is necessary to count the occurrences of each hidden state at the start of the sequences. The dataset is examined and it has been counted how many sequences start in each hidden state.

Emission probabilities represent the likelihood of observing a particular feature or set of features given a specific hidden state at time t . The dataset was grouped based on the hidden state labels for each time step. The estimation process for emission probabilities differs based on whether data is discrete or continuous.

- **Discrete Observations:** If the observations are discrete, the frequency of each observation in each state is calculated.
- **Continuous Observations:** If the observations are continuous (e.g., attendance rates), they are modeled using probability distributions. In this research paper, we utilized a Gaussian probability distribution. For each feature, such as attendance rate, a multivariate Gaussian distribution is employed to model the joint probability distribution of all observations given the hidden state. The parameters of the Gaussian distribution, including mean vector and covariance matrix, are estimated based on the observations associated with each hidden state.

Transition probabilities represent the likelihood of moving from one hidden state to another between consecutive time steps. The data needs to be organized as sequences over time (e.g., each

student's progression through terms) — which was completed during the data preparation phase. The number of transitions from each state to every other state in the dataset needs to be counted. These counts are in turn normalized to get transition probabilities (by dividing by the total number of transitions out of each state).

3.5. Training and Evaluating the Hidden Markov Model

In this study, Python programming language was utilized along with the `hmmlearn` library for HMM implementation. The `hmmlearn` library provides an efficient and easy-to-use interface for training HMM models and estimating their parameters. The `hmmlearn` library handles parameter estimation by internally, when calling the `fit()` function. Example code for training an HMM model is shown on Fig. 2.

```

1 from hmmlearn import hmm
2 hidden_states = ["At Risk of Drop-Out", "On Pace", "Excellent"]
3 model = hmm.MultinomialHMM(
4     n_components=len(hidden_states),
5     n_iter=100
6 )
7
8 # Train the HMM model
9 model.fit(X)
10
11 # Display the parameters
12 print("Initial state probabilities:", model.startprob_)
13 print("Transition probabilities:", model.transmat_)
14 print("Emission probabilities:", model.emissionprob_)

```

Figure 2. Training the HMM model

A Multivariate Hidden Markov Model (HMM) was used here to model the joint probability distribution of multiple observed features, allowing for a more accurate representation of the complex relationships and dependencies among the observed variables.

For this research paper, the dataset was split into train and test parts in an 80-20 ratio, allowing for model training on the larger portion of the data while reserving a smaller portion for evaluation purposes. State prediction accuracy and state transition accuracy were subsequently calculated using standard evaluation metrics, enabling the assessment of the model's performance on the test set.

State Prediction Accuracy evaluates the accuracy of predicting the correct state sequence for new sequences of observations. It measures how well the HMM model predicts the latent student trajectories. **State Transition Accuracy** assesses how accurately the model predicts transitions between different states over time. It evaluates whether the model captures the expected transitions in student trajectories.

4. RESULTS AND DISCUSSION

4.1. Results and Discussion

After training and evaluating the Multivariate Hidden Markov Model, the state prediction accuracy was determined to be 91.12%, indicating that the model accurately predicted the latent student

trajectories, including "At Risk of Drop-Out", "On Track", and "Excellent", for a significant portion of the dataset. Additionally, the state transition accuracy was found to be 86.70%, demonstrating the model's ability to effectively capture the transitions between different student states over time. These results highlight the promising performance of the HMM in predicting student academic success trajectories. The detailed evaluation metrics, including state prediction accuracy and state transition accuracy, are presented in Table 4, providing a comprehensive overview of the model's performance in capturing the underlying dynamics of student outcomes. These metrics are crucial for understanding and predicting student outcomes as they provide insights into the model's ability to discern and anticipate changes in students' academic trajectories.

Table 4. HMM Performance Evaluation Results

Metric	Test Result
State Prediction Accuracy	91.12%
State Transition Accuracy	86.70%

The high state prediction accuracy and state transition accuracy obtained in this study demonstrate the potential of HMMs as a powerful tool for understanding and predicting student outcomes. By accurately capturing the dynamics of student progress and identifying patterns in their academic trajectories, HMMs offer valuable insights for educational interventions and support systems. For instance, based on the predicted trajectories, educators and policymakers can tailor interventions to provide timely support to students who are deemed at risk of drop-out, thereby improving retention rates and fostering academic success. Additionally, insights derived from HMMs can inform the development of personalized learning pathways and intervention strategies, ultimately enhancing student engagement, performance, and overall educational outcomes. Thus, the demonstrated effectiveness of HMMs in predicting student outcomes underscores their potential as a valuable tool for educational research, policy-making, and practice.

4.2. Comparison With Related Research

In a similar manner to [7] this research paper also utilized clustering techniques to label its dataset. By employing the k-means algorithm, the dataset in [7] was clustered based on 12 engagement metrics, categorized into interaction-related and effort-related aspects. This approach enabled the identification of distinct groups of students with varying levels of engagement, thereby facilitating the assessment of student involvement and potential areas for intervention. The clustering process allowed for the categorization of students into different engagement levels, which is crucial

for personalized e-learning experiences and effective educational interventions. By leveraging machine learning techniques like clustering, both research papers aimed to address challenges in e-learning platforms, such as personalization and student engagement, ultimately contributing to the improvement of learning outcomes and experiences in online education settings.

Both [8] and this research paper utilize Hidden Markov Models (HMMs) to analyze student behavior in online educational environments, albeit for slightly different purposes. Ref. [8] focuses on predicting student retention in Massive Open Online Courses (MOOCs) by leveraging HMMs to understand student behavior over time. It addresses the challenge of student dropout rates in MOOCs by modeling latent characteristics of students that influence their perseverance using observable interactions with the course. The HMM framework allows for the prediction of a student's behavior in the next time step based on previous states and observable actions.

Ref. [9] focuses on a classification problem, attempting to predict student success or failure based on similar data points — demographic information, studying routines, attendance behaviors, and epistemological beliefs. It compares the prediction accuracy of various supervised classification algorithms, with the Neural Network algorithm achieving the highest accuracy.

5. CONCLUSION

This research paper demonstrates an approach for predicting student academic success using Hidden Markov Models. By integrating a comprehensive dataset comprising students' demographics, academic performance, attendance records, and course engagement, the study effectively employs an HMM framework to model varying levels of student academic success. The model utilizes observable emissions, such as grades and interaction patterns, to infer the most likely sequence of hidden states for new students. The evaluation results reveal a state prediction accuracy of 91.12% and a state transition accuracy of 86.70%, highlighting the HMM's robust capability to predict latent student trajectories. These results showcase the model's efficacy in capturing the underlying dynamics of student outcomes.

The high accuracy rates underscore the potential of HMMs as a powerful tool for understanding and predicting student outcomes, offering valuable insights into educational interventions and support systems. By accurately modeling the dynamics of student progress and identifying critical patterns in their academic trajectories, HMMs can enable educators and policymakers to tailor interventions to students' needs, thereby enhancing retention rates and promoting academic success. This approach can inform the development of

personalized learning pathways and timely intervention strategies, ultimately leading to improved student engagement and performance.

Future research could explore several extensions of this study to further enhance the application and effectiveness of HMMs in educational settings. First, incorporating additional features such as social interactions, extracurricular activities, and psychological factors could provide a more holistic understanding of student behavior and success. Additionally, integrating temporal factors more dynamically into the HMM framework could improve the model's responsiveness to changes in student behavior over time. Another avenue for future work involves exploring the application of HMMs to different educational contexts, such as vocational training or adult education, to assess their generalizability and adaptability. Furthermore, comparative studies with other machine learning models, such as neural networks or ensemble methods, could provide insights into the relative advantages and limitations of HMMs. Finally, the development of interactive tools and dashboards based on HMM predictions could facilitate real-time monitoring and intervention by educators, enhancing the practical utility of the model in educational practice.

ACKNOWLEDGEMENT

This study was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, and these results are parts of the Grant No. 451-03-66/ 2024-03/200132 with University of Kragujevac – Faculty of Technical Sciences Čačak.

REFERENCES

- [1] Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson Education.
- [2] Awad, M., & Khanna, R. (2015). Hidden Markov Model. In *Efficient Learning Machines*. doi:10.1007/978-1-4302-5990-9_5.
- [3] OC2 Lab. (Accessed on 12th April 2024). Student-Performance-and-Engagement-Prediction-eLearning-datasets. Retrieved from <https://github.com/Western-OC2-Lab/Student-Performance-and-Engagement-Prediction-eLearning-datasets>.
- [4] Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *Proceedings of 5th Future Business Technology Conference (FUBUTECH 2008)* Porto, Portugal: EUROSIS. ISBN 978-9077381-39-7.
- [5] Zhu, Y. *Student Interaction and Performance in Online Courses*. Retrieved from <https://data.world/yifanzhu>. Accessed on 17th April 2024.
- [6] Suzan, M.M.H., Samrin, N.A., Biswas, A.A., & Pramanik, M.A. (2021). Students' Adaptability Level Prediction in Online Education using Machine Learning Approaches. In *Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IIT - Kharagpur, Kharagpur, India. doi:10.1109/ICCCNT51525.2021.9579741.
- [7] Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2020). Student Engagement Level in e-Learning Environment: Clustering Using K-means. *American Journal of Distance Education*. doi:10.1080/08923647.2020.1696140.
- [8] Balakrishnan, G. (2013). Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models. Technical Report No. UCB/EECS-2013-109, Electrical Engineering and Computer Sciences, University of California at Berkeley.
- [9] Yıldız, M. B., & Börekçi, C. (2020). Predicting Academic Achievement with Machine Learning Algorithms. *Journal of Educational Technology & Online Learning*, 3(3), 372-392. doi:10.31681/jetol.773206.