Docens Series in Education
2583-1054

Yunosheva & Dukhanov, 2026

Volume 10, pp. 43-51

DOI- https://doi.org/10.20319/dv.10.4351

Received: 05th May 2026

Revised: 11th May 2026

Accepted: 08th June 2026

Date of Publication: 30th July 2026

This paper can be cited as: Yunosheva, V.V. & Dukhanov, A.V. (2026). Bayesian Modeling of Student Academic Performance: Evaluating Motivational Factors. Docens Series in Education, 10, 43-51

BAYESIAN MODELING OF STUDENT ACADEMIC PERFORMANCE: EVALUATING MOTIVATIONAL FACTORS

Victoria Vadimovna Yunosheva
Department of Education and Math Studies, ITMO University, Moscow
vvyunosheva@edu.hse.ru

Alexey Valentinovich Dukhanov
Scientific supervisor – Doctor of Science, Associate Professor, ITMO University, Moscow
dukhanov@itmo.ru

Abstract

Over the past 10 years, according to source [1], most of the supplementary education in the CIS countries and around the world has been moving to online platforms. Many private schools and public institutions are interested in creating a tool that not only tracks student satisfaction and progress but also predicts their academic performance and success in completing courses. In this regard, they collect information about students, their academic and non-academic activities, assessment results, teacher activity, and the performance of other staff. This paper proposes a method based on working with students of various ages studying Chinese, using Bayesian modeling to analyze their productivity and success during the learning process. The main attempt of this paper is to analyze and predict learners' performance through online conferences to better *understand their marks and predict students' success. The results of this study are going to be presented for international language school and adopted through the usage of methodology.*

Keywords:
Bayesian Modeling, Educational Data Mining, Student Success Prediction, Age-Stratified Analysis, Chinese Language Learning, Attendance Tracking, Predictive Analytics, Online Education

Introduction

This study focuses on the development of a method for assessing and predicting the academic performance of both students and research teams, grounded in Bayesian modeling and cluster analysis. The primary objective is to create an analytical framework applicable to a wide range of educational institutions, including schools, universities, and supplementary education providers (e.g., language courses, tutoring centers). The proposed method enables a systematic examination of behavioral patterns and educational outcomes, empowering institutions to engage in data-driven decision-making based on information collected during teacher-learner interactions, whether in formal lessons or informal communication settings. By integrating factor analysis, video lesson data, and student self-reflection results, the approach offers a holistic and multidimensional understanding of the key factors that shape student academic performance.

Literature Review

The choice of Bayesian modeling is not accidental. According to the classification proposed in [6], Bayesian models of student academic performance are divided into three categories, depending on the source of structure and probability formation. In the first case, all parameters are set by experts. In the second, efficiency is achieved by artificially constraining the network structure. In the third, the model is fully trained on historical data. The Bayesian approach is attractive due to two properties: high performance and interpretability of results through graph-based representation [6].

A crucial role in developing a forecasting and modeling solution is played by the final assessment and the predicted variable [2]. After reviewing the literature, it was decided not to limit the solution solely to Bayesian graphical models of student knowledge levels. These models are among the most frequently used and provide a high level of flexibility [4]. However, they assume a graph-based construction, which imposes a linear logic on the learning process within a discipline

— an assumption that is not always valid for a student whose curriculum may not follow a linear progression.

Methodology

During the research, more than 300 video lessons were downloaded and processed. These lessons involved students of various ages and different group types, ranging from individual lessons to groups of up to 5 people, with varying language proficiency levels, professional fields and activities, as well as different learning goals. Several limitations of this study should be acknowledged. First, the adult sample size (n = 21) was considerably smaller than the child sample (n = 210), which may affect the precision of posterior estimates for the adult group. Future research should collect additional adult learner data to validate and refine the models. Second, the study was conducted within a single international language school, which may limit generalizability to other educational contexts (e.g., public schools, university language programs, or online-only platforms).

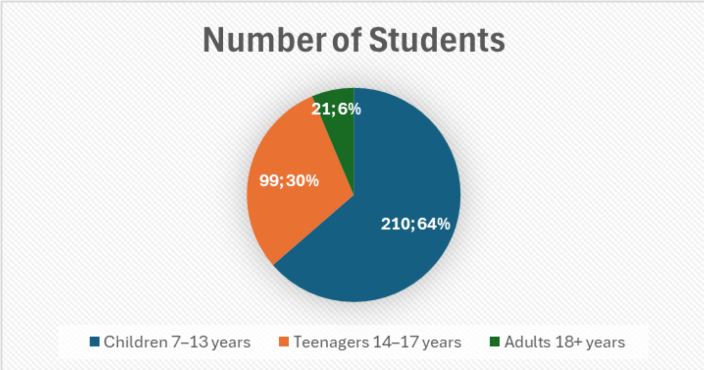
In addition to video data, metrics were collected to assess the following:

- Time spent learning during the lesson
- Time spent on homework completion
- Grades
- Personal data including age, gender, and the presence of siblings
- Presence of father and mother
- Duration of study
- Learning goals
- Level of education

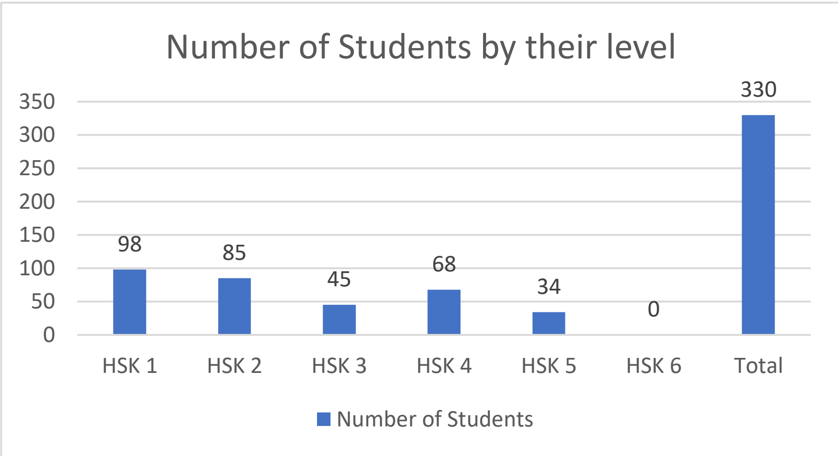
Teacher data including nationality,

- presence of specialized higher education
- work experience
- language proficiency (if the teacher is not a native speaker)
- Teacher workload

In total, more than 80 metrics directly influencing the educational process were collected. Based on the results of the data collection, the following statistics were identified. Undergone are presented picture 1 and 2 with the collected data.



Picture 1. *Overview Statistics by Age*



Picture 1. *Overview Statistics by Language Level*

To determine whether the observed age distribution in your sample (children/teenagers/adults) significantly differs from a hypothesized population distribution was used Chi-Square Goodness-of-Fit Test for Age Distribution Representativeness. The formula is presented with the first link.

Chi-Square Calculation:

$$\chi 2 = \sum \frac{E(O - E)^2}{E} \quad (1)$$

Were

- $\chi 2$-chi-square test statistic
- O_i — **observed frequency** in the i-th category
- E_i — **expected frequency** in the i-th category under the null hypothesis
- i — index of the category (from 1 to k)
- k — total number of categories

Null hypothesis (H_0): The sample age distribution matches the reference population distribution.

Alternative hypothesis (H_1): The sample age distribution is significantly different.

The results are presented in Table 1 and 2.

Degrees of freedom: df = 3 – 1 = 2	Data
Critical value ($\alpha = 0.05$, df = 2)	5.991
Calculated χ^2	**31.04**
Since 31.04 > 5.991, we **reject the null hypothesis**.	

Table 1. *The result of Chi-Square Calculation*

Since 31.04 > 5.991, we reject the null hypothesis.

Thus, the sample is not homogeneous: the age distribution differs significantly from that of the general population. Therefore, to accurately predict successful course completion, it is necessary to model the academic trajectory separately for children and adults.

Based on Bayesian modeling, predictive models were constructed to estimate the probability of successful course completion. **Successful completion** is defined as passing a local test on the learned material over a study period of **1 year (or 12 months)** .

For children (aged 7–13 years) and adults (aged 18+ years), various Bayesian models were applied to estimate the probability of success depending on the following predictors:

- Initial HSK level
- Homework time (minutes/week)
- Class attendance rate (%)
- Parental support (for children) / self-organization level (for adults)
- Teacher workload (number of students per teacher)

Results

Based on the collected data, there were created Bayes models for identifying key factors and their influence on success. Here are the results for the two groups by age to identify what type of factors are the key while talking about learning process and success estimation. In Table 1 is observed that the probability of success increases with higher initial HSK levels, rising from 52% for HSK 1 to 81% for HSK 3+, with regular attendance being the strongest predictor. Table 3 observes the adult learners show lower success probabilities at HSK 1 (44%) compared to children, with success depending more on self-discipline and purposeful study than external control. Table 4 gives the results that children have a higher mean success probability (67% vs. 62%) and require less homework time (>90 min) and a lower attendance threshold (70%) than adults (>120 min, 75%), while adults show a slightly stronger influence of help/support (β = 0.49 vs. 0.43).

Group 1: Children (7–13 years, n = 210)

Parameter	Posterior probability of success (95% HDI)	Key factors
Initial HSK 1	0.52 [0.47 – 0.57]	Low self-organization requires parental control
Initial HSK 2	0.68 [0.63 – 0.73]	Moderate influence of homework >90 min/week
Initial HSK 3+	0.81 [0.77 – 0.85]	High probability of regular attendance

Table 2. *The result for Children (7–13 years, n = 210)*

Group 2: Adults (18+ years, n = 21)

Parameter	Posterior probability of success (95% HDI)	Key factors
Initial HSK 1	0.44 [0.35 – 0.53]	Influence of workload, low regularity
Initial HSK 2	0.62 [0.54 – 0.70]	Moderate prediction, strong dependence on self-discipline
Initial HSK 3+	0.79 [0.72 – 0.86]	High probability with purposeful study

Table 3. *Adults (18+ years, n = 21)*

Metric	Children (7–13 years)	Adults (18+ years)
Mean predicted success probability (all levels)	0.67 [0.62 – 0.72]	0.62 [0.54 – 0.70]
Influence of help/support (β)	0.43	0.49
Critical attendance	70%	75%
Required homework time (min/week)	>90	>120

Table 4. *Comparison Between Groups (Posterior Difference)*

Compared to earlier studies [2, 4, 6], which focused primarily on university students or homogeneous age groups, our research addresses a significant gap by analyzing children and adults within the same language learning context, using a unified Bayesian modeling methodology. First, it provides empirical evidence for age-stratified predictive modeling, demonstrating that a onesize-fits-all approach to predicting academic success is inadequate when working with heterogeneous learner populations. Second, it quantifies the differential impact of key predictors (attendance, homework time, initial proficiency, and external support) across age groups, offering actionable thresholds for practitioners. Third, it integrates real-world video lesson data with student performance metrics, creating a comprehensive analytical framework that goes beyond traditional test-score-based predictions.

Discussion

Additionally, the study was conducted within a single international language school context, and results may not fully extend to other educational environments such as public schools, university programs, or fully online platforms.

Future research should focus on three main directions:

(1) collecting larger and more balanced datasets across multiple institutions to validate and refine the proposed Bayesian models;

(2) incorporating additional predictors such as socioeconomic status, prior language learning experience, and learning style preferences; and

(3) developing a real-time early warning dashboard that operationalizes the Bayesian networks for classroom use by teachers and administrators.

Conclusion

This study set out to develop a method for assessing and predicting the academic performance of students learning Chinese as a foreign language, using Bayesian modeling and cluster analysis. Based on the analysis of more than 300 video lessons and over 80 collected metrics across two age groups (children aged 7–13 years and adults aged 18+ years), several key conclusions can be drawn regarding the factors influencing successful course completion and the practical implications for educational institutions. For children, parental involvement and attendance tracking are critically important. For adults, goal-setting workshops and managing the student-to-teacher ratio are more effective. The first-year threshold for passing a local test is realistic for students with an initial HSK level of 2 or higher. Students starting at HSK level 1 require additional support, especially children (predicted probability of success is only 0.23 with low attendance).

References

Priyanka, Gupta, A., & Kohli, S. (2021). *Supplementary education: A global perspective*. Central University of Jammu.

Saeedi, S., Božanić, D., & Safa, R. (2024). Strategic analytics for predicting students' academic performance using cluster analysis and Bayesian networks. *Education Science and Management, 2*(4), 197–214. https://doi.org/10.56578/esm020402

Deviatiarova, E., Fadeev, S., & Dukhanov, A. (2023). Analyzing proficiency patterns in test results using clustering and augmentation algorithms. *Procedia Computer Science, 229*, 254–264. https://doi.org/10.1016/j.procs.2023.12.027

Sundar, P. V. P. (2013). A comparative study for predicting student's academic performance using Bayesian network classifiers. *IOSR Journal of Engineering, 3*(2), 37–42. https://doi.org/10.9790/3021-03213742

Nudelman, Z., Moodley, D., & Berman, S. (2019). Using Bayesian networks and machine learning to predict computer science success. In *International Conference on Artificial Intelligence in Education* (pp. 14-22). Springer. https://doi.org/10.1007/978-3-030-05813-5_14

Osmanbegović, E., & Suljić, M. (2012). Data mining approach for predicting student performance. *Journal of Economics and Business, 10*(1), 3-12.

Liang, F. (2005). Bayesian neural networks for non-linear time series forecasting. *Statistics and Computing, 15*, 13–29. https://doi.org/10.1007/s11222-005-4786-8

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, 47*, 168–181. https://doi.org/10.1016/j.chb.2014.09.034

Heckerman, D. (2008). A tutorial on learning with Bayesian networks. In *Innovations in Bayesian networks* (pp. 33-82). Springer. https://doi.org/10.1007/978-3-540-78666-7_2

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.