Journal of Contemporary Educational Research, 2025, Volume 9, Issue 12
http://ojs.bbwpublisher.com/index.php/JCER
ISSN Online: 2208-8474
ISSN Print: 2208-8466

Predicting the Ranking of Engineering Mechanics Students Using the Bayesian Model

Kuahai Yu[1,2], Xiangqian Sheng[2]*, Sibo Dang[2], Ruhuan Yu[2]

[1]Institute of Science and Technology Development, Henan University of Science and Technology, Luoyang 471000, Henan, China
[2]School of Civil Engineering and Architecture, Henan University of Science and Technology, Luoyang 471000, Henan, China

Corresponding author: Xiangqian Sheng, 2008shengxq@163.com

Copyright: © 2025 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: GPA plays an important role in the entire learning process of students. The value of GPA not only reflects students' current grades but also affects their future progress, motivation, and opportunities. It is worth noting that the grades of specific courses also have an impact on GPA. Therefore, it is necessary to predict students' performance in future courses based on their current grades. In this paper, a Bayesian model is employed to classify course grades and estimate the probability of these grades being affected by other factors in the first semester, enabling the prediction of subsequent performance. The Bayesian approach integrates prior knowledge of grade distributions through four key steps: establishing a prior probability distribution, using a likelihood function to relate grades to academic ability, combining prior and new evidence to compute posterior probabilities, and forecasting next-semester results. These predictions support timely academic interventions and adjustments to teaching strategies. By utilizing data such as assignment and exam scores, a Bayesian classification model can analyze and predict outcomes. The actual grades of students in the second semester are used to validate the predictive accuracy of the model.

Keywords: Bayesian model; Predictive data; Academic performance forecasting; Data mining; Model evaluation

Online publication: December 31, 2025

1. Introduction to the principles

1.1. Principles of Bayesian model

The Bayesian principle is a foundational concept in probability theory that outlines how to update probability distributions when new evidence becomes available. Specifically, it describes how to compute posterior probabilities by combining prior probabilities with observed data. The core idea of this principle is that we can revise our beliefs about events (posterior probability) by incorporating both our existing knowledge (prior probability) and newly acquired information.

1.2. Mathematical formulation of Bayesian theory

1.2.1. Prior probability

The prior probability refers to our initial estimate or belief about the probability of an event occurring before observing any data. In this paper, the prior probability is defined as $P(c_i)$, where represents the class of sample data. It is typically approximated by the ratio of the number of samples belonging to class c_i ($|c_i|$) to the total number of samples $|D|$, namely

$$P(c_i) = \frac{|c_i|}{|D|}. \tag{1}$$

1.2.2. Likelihood probability

The likelihood probability refers to the conditional probability of observing a specific outcome given a set of parameters. It represents the probability of the results under known parameters [1]. Mathematically, it is expressed as $P(X|c_i)$, which denotes the probability of the target variable X occurring when the random variable is known to be c_i. If $X=\{a_1,a_2,a_3\cdots\cdots\}$, the mathematical expression for the likelihood probability is [2]

$$P(X|c_i) = P(a_1, a_2, ..., a_m|c_i), i \in (1, |c_i|). \tag{2}$$

It is worth noting that each random variable is treated as an independent entity in Bayesian probability computations due to the relatively complex calculation of Eq. (2). Thus, the likelihood probability (class-conditional probability) in the above expression is equivalent to the product of the probabilities corresponding to each independent variable, namely

$$P(X|c_i) = P(a_1, a_2, ..., a_m|c_i) = \prod_{i=1}^{m} P(a_i|c_i), i \in (1, |c_i|). \tag{3}$$

1.2.3. Posterior probability

Posterior probability refers to the probability of an event occurring after taking into account observed data or evidence. In Bayesian inference, posterior probability is calculated using Bayes' theorem. The mathematical expression is given as $P(c_i)$, representing the probability that the target variable X belongs to a random variable c_i,

$$P(c_i|X) = \frac{P(X|c_i)P(c_i)}{P(X)}, i \in (1, |c_i|). \tag{4}$$

Since $P(X)$ remains constant across all variables, in practical applications, we only need to compute the numerator of Bayes' theorem and determine its maximum value [3]. The target variable X is then classified into the category c_{cap} corresponding to this maximum value.

$$P(c_{cap}|X) = \max\{P(X|c_i)P(c_i)\}. \tag{5}$$

2. Experimental procedure

(1) Data collection

Collect historical academic performance data for the course, including previous semester grades, attendance records, and regular assignment scores.

Gather other potentially influential factors, such as students' study habits and family background.

Calculate the correlation between predictor variables and the target variable (e.g., final grade). Select variables with strong positive correlations and discard those with weak or negative correlations.

(2) Data preprocessing

Clean and process the collected data, handling missing values and outliers.

Perform feature engineering, such as encoding categorical variables into numerical values, standardization, or normalization.

(3) Building the Bayesian model

Select an appropriate Bayesian model (Naive Bayes, Bayesian Linear Regression) based on the problem complexity. This experiment uses the Naive Bayes model.

Determine the prior probability distribution, which may rely on prior knowledge or domain expert recommendations.

(4) Model training

Train the Bayesian model using historical data to estimate prior probabilities, class-conditional probabilities, and posterior probabilities.

(5) GPA prediction

Use the trained model to predict the next semester's course performance for new students.

Depending on requirements, output either a probability distribution or a deterministic prediction.

(6) Model evaluation

Assess the model's predictions by comparing them with actual results.

Measure accuracy, precision, recall, or other relevant metrics.

(7) Feedback and improvement

Refine the model based on evaluation results, adjusting model structure or feature selection as needed.

Continuously collect new data and iterate through the steps to enhance model performance.

Compare predictions with those from other models to validate the Bayesian model's accuracy and effectiveness.

3. Data processing

3.1. Calculating correlation between variables and the target variable

When constructing a Bayesian predictive model, it is crucial to understand the correlation between each predictor variable and the target variable. In particular, numerous factors can influence a student's academic performance in a given subject. Without proper consideration, the experiment may become overly complex, leading to reduced prediction accuracy. Features with higher correlations are more likely to impact the target variable. Therefore, incorporating these features into the model can enhance both accuracy and efficiency. By analyzing the correlation between predictor variables and the target variable, we can select the most influential features as model inputs, thereby more precisely capturing variations in student performance. This careful feature selection method not only reduces model complexity and improves computational efficiency but also enhances predictive performance, making the model more aligned with real-world scenarios. Thus, considering correlation when building a Bayesian predictive model is a critical step in improving model performance and ensuring both accuracy and efficiency [1,4].

$$\mathbf{R} = \frac{\sum\left(\mathbf{x}_i - \mathbf{x}\right)\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)}{\sqrt{\sum\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^2 \sum\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)^2}} . \quad (6)$$

The range of R values is from -1 to 1, with specific meanings as follows:

R = -1 indicates a perfect negative correlation between the random variable and the target variable

R = 1 indicates a perfect positive correlation between the random variable and the target variable

R = 0 indicates no linear relationship between the random variable and the target variable.

Thus, the closer R is to 1, the stronger the correlation between the random variable and the target variable. In this experiment, random variables with R values greater than 0.50 were selected.

Through our analysis, as shown in **Table 1**, factors such as family background, gender, and age exhibited correlation coefficients (R-values) with final grades that were below

0.50 indicates a significantly lower correlation with academic performance compared to other factors. Consequently, when calculating the probabilistic influence of variables on the target outcome, we opted to exclude these variables. This indicates that these factors have minimal impact on students' course grades and can be reasonably disregarded in the model.

Table 1. Correlation between certain factors and final grades

Factor	Final grade
Interest in the course	0.528698
Final exam	0.971053
Regular grades	0.662559
Family background	-0.16481
Gender	0.262736
Age	-0.0728

Therefore, in constructing the Bayesian predictive model, we selected variables demonstrating strong correlation with the target variable while excluding those with weaker associations. This approach serves to streamline the model while enhancing both predictive accuracy and reliability.

Through such analytical refinement, we gain deeper insights into students' academic performance, enabling more effective educational guidance and support. This Bayesian prediction framework not only assists students in optimizing their learning pathways and improving academic outcomes but also provides valuable decision-making references for educational institutions. By informing curriculum design and instructional planning, the model contributes to comprehensive improvements in educational quality and student development.

3.2. Data preprocessing

For privacy protection purposes, original student names have been replaced with numerical identifiers, and personal information such as gender and student ID numbers has been hidden. Only essential academic data, including grades and course names, has been retained.

Prior to constructing the Bayesian network model, the data underwent discretization preprocessing since this model operates on discrete data. In this study, academic grades were categorized into four levels:

Grade “A” for scores above 90

Grade “B” for scores above 70

Grade “C” for scores above 60

Grade “D” for scores below 60

This discretization process converts continuous grade data into discrete categories, facilitating the development of the Bayesian network model. Such an approach effectively captures variations in student performance across different courses while simplifying the modeling process. Moreover, it preserves the predictive capability for student GPA. The results of the data preprocessing are presented in **Table 2**.

Table 2. Preprocessing of random variable data

Number	Name	Final grade	Regular assessment scores	Course interest level
1	Student 1	96A	94A	91A
2	Student 2	96A	87B	87B
3	Student 3	93A	91A	93A
4	Student 4	96A	95A	89B
5	Student 5	94A	96A	90A
6	Student 6	89B	93A	90A

3.3. Building the Bayesian model

3.3.1. Sample size determination

When constructing the Bayesian probability model, determining an appropriate sample size is critical. Particularly when the target variable is known, the sample size of random variables directly impacts the model’s accuracy and reliability. By collecting a sufficient number of samples, we can more precisely estimate the probability distributions between variables, thereby enhancing the model’s predictive performance.

Moreover, increasing the sample size helps reduce estimation errors and improve model stability. Therefore, it is essential to carefully consider these factors and select an adequate sample size to ensure the validity and reliability of the Bayesian probability model [5]. The results are shown in **Table 3**.

Table 3. The sample size of the random variable under the premise of the target variable

Course interest level				Regular assessment scores				Final exam grade				Final grade	
A	B	C	D	A	B	C	D	A	B	C	D		
6	6	0	0	11	1	0	0	12	0	0	0	A	12
6	18	3	0	13	14	0	0	1	21	5	0	B	27
0	2	1	1	1	3	0	0	0	0	3	1	C	4
0	2	0	0	0	1	0	1	0	0	0	2	D	2

3.3.2. Calculating relevant probabilities

Using the relevant Bayesian formula, we can compute the prior and posterior probabilities. The prior probability is the probability before observing any evidence, while the posterior probability is the updated probability after considering the observed evidence. These probabilities play a key role in Bayesian inference, helping us better understand and predict the occurrence of events, as shown in **Table 4**.

Table 4. Random variable's prior probability and posterior probability

Course interest level				Regular assessment scores				Final exam grade				Final grade	
A	B	C	D	A	B	C	D	A	B	C	D		
0.5	0.5	0	0	0.92	0.08	0	0	1	0	0	0	A	0.27
0.22	0.67	0.11	0	0.48	0.52	0	0	0.04	0.78	0.19	0	B	0.6
0	0.5	0.25	0.25	0.25	0.75	0	0	0	0	0.75	0.25	C	0.09
0	1	0	0	0	0	0	0.5	0	0	0	1	D	0.04

4. Application of Bayesian models

A student sample is randomly selected, and three random variables are statistically analyzed with specific values X. Based on Eq. (1) and Eq. (3) in the Bayesian model, the classification of the student's final grade in this course is predicted.

First, the probability that the grade classification is A is calculated as:

$$P(\mathrm{C_A}|\ X) = P(X|\ \mathrm{C_A})\mathrm{P}(\mathrm{C_A}) = P(X|\ \mathrm{C_A})\cdot P(\mathrm{A}|\ \mathrm{C_A})\cdot P(\mathrm{B}|\ \mathrm{C_A})\cdot P(\mathrm{C_A}). \quad (7)$$

According to the calculation procedure described above, if a student's final grade is classified as a while the probability of their final exam score being B is 0, then the resulting probability computed using the relevant Bayesian formula would also be 0. However, this outcome is unreasonable because we failed to account for students who missed the exam or submitted blank answer sheets. Therefore, the model requires modification to incorporate such cases and ensure reasonable results. This adjustment may involve revising the prior probabilities or re-evaluating the model's variables and conditions.

The calculation rules of Bayesian probability models indicate that when a random variable has zero observed samples, an improved Bayesian model can be employed to compute conditional probabilities more accurately. This approach expands the actual observations (n_i) by adding m virtual samples following the P_i distribution, where m is a constant representing the equivalent sample size. $P_i = 1/A_i$ .Simply put, the numerator becomes 0 plus the total sample count multiplied by 1/3, while the denominator becomes the sample count of C_i plus the total sample count [6–8]

$$P = \frac{n_{ij} + mp_i}{n_j + m}. \quad (8)$$

Using the aforementioned formula, we calculated the probabilities for the student's final grade classification as follows:

Probability of classification B: 0.0494,

Probability of classification C: 0.0021,

Probability of classification D: 0.00325

The maximum probability among these values corresponds to classification B. Therefore, the student's final grade for the second semester of freshman year is classified as B, with a GPA range of [70,89]. Comparison with the actual grade classifications in the student sample dataset for that semester confirmed that the prediction matched the actual result.

5. Evaluation of Bayesian models

5.1. Research methods for model validation

The fact that a single student's predicted grade category matches the actual grade category is insufficient to prove the validity and accuracy of the Bayesian model. Evaluating a Bayesian model requires deeper analysis and assessment, including a careful consideration of its predictive capability, generalization ability, robustness, and comparison with other models. Therefore, multiple evaluation methods must be employed to verify the performance and reliability of the Bayesian model. The methods for model validation include, but are not limited to, the following aspects.

(1) Model selection and hypothesis validation: Bayesian models rely on certain assumptions, such as the choice of prior distributions. These assumptions may not always hold in real-world problems, so the model's assumptions must be validated and tested.

(2) Model evaluation: Appropriate evaluation metrics, such as predictive accuracy and generalization ability, should be used to assess the Bayesian model's performance. Techniques like cross-validation and holdout methods can be applied for evaluation.

(3) Comparison with other models: Bayesian models should be compared with other commonly used predictive models, such as linear regression, decision trees, and support vector machines. By comparing the performance of different models, it can be determined whether the Bayesian model has advantages for specific problems.

(4) Domain knowledge and practical application: The effectiveness of a Bayesian model also depends on its performance in real-world applications and its recognition by domain experts. In some cases, Bayesian models may better capture the characteristics and patterns of the data.

5.2. Model assessment

In this experiment, we used a Bayesian classification model to predict and categorize the sampled data, comparing the predicted grade categories with the actual grades to verify its accuracy and effectiveness. Based on 45 training samples, the model performed well on the test dataset, achieving an accuracy rate of 91%. Furthermore, when tested on a random set of 12 students from the same course, the accuracy remained above 80%, further demonstrating the model's validity.

The Bayesian classification model has clear interpretability, is easy to understand, and features low computational complexity, making it suitable for large-scale databases. Additionally, it exhibits strong incremental learning capabilities, allowing continuous optimization as new data is incorporated. Therefore, the Bayesian classification model is an effective and reliable method for predicting and categorizing student grades, providing schools with valuable academic guidance and decision-making support [2].

6. Conclusion

This paper has detailed the fundamental theory of Bayesian models and applied a Bayesian classifier to analyze student performance. A first-year course grade was selected as an attribute, with prior probabilities and class-conditional probabilities calculated using training samples before processing test data. The analysis revealed that final exam scores were the primary factor influencing student performance in this course, while interest in the subject and regular grades also had some impact. Testing the model with sample data yielded an accuracy rate of 83%. However, limitations include the small sample size and the consideration of only a few influencing factors. Additionally, Bayesian classification models require known prior probabilities and assume attribute independence,

making them unsuitable for scenarios with numerous attributes or high inter-attribute correlations.

Funding

Research and Practice Project on Teaching Reform in Higher Education in Henan Province (2024SJGLX0094)

Disclosure statement

The authors declare no conflict of interest.

Author contributions

Data curation: Ruhuan Yu
Formal analysis: Xiangqian Sheng
Funding acquisition: Kuahai Yu
Investigation: Ruhuan Yu
Methodology: Xiangqian Sheng, Sibo Dang
Resources: Kuahai Yu
Software: Ruhuan Yu
Supervision: Kuahai Yu
Writing – original draft: Xiangqian Sheng
Writing – review & editing: Kuahai Yu

References

[1] Mu CD, Dai JB, Ye J, 2000, Bayesian Network for Data Mining. Journal of Software, 11(5): 660–666.

[3] Han L, 2018, Application of the Bayesian Classification Model in the Forecast of Student Achievement. Computer & Digital Engineering, 46(10): 2039–2056.

[4] Yang A, Jiang X, Liu P, et al., 2016, Sparse Bayesian Multinomial Probit Regression Model with Correlation Prior for High-Dimensional Data Classification. Statistics & Probability Letters, 119: 241–247.

[4] Huang H, 2010, Principle and Application of Rough Set Algorithm in Library Data Mining Server System. Hebei Journal of Industrial Science and Technology, 27(5): 344–347.

[5] Han L, Han G, 2017, Application of the Decision Tree Algorithm in the Analysis of Student Achievement. Electronic Design Engineering, 25(2): 18–21.

[6] Bhardwaj BK, Pal S, 2011, Data Mining: A Prediction for Performance Improvement Using Classification. International Journal of Computer Science and Information Security, 9(4): 136–140.

[7] Arias J, Martínez J, Gámez JA, et al., 2016, Medical Image Modality Classification Using Discrete Bayesian Networks. Computer Vision and Image Understanding, 151: 61–71.

[8] Bataineh M, Qudah Z, 2017, A Novel Gene Identification Algorithm with Bayesian Classification. Biomedical Signal Processing and Control, 31: 6–15.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.