Predicting Senior High School Students' Higher Education Intention using Naive Bayes Algorithm: A Case Study in SMAN 6 Luwu Timur, Indonesia

Kadek Ayu Suryanti[1], Thresia Zita Ohoiwutun[2], Besse Helmi Mustawinar[3*], Nirma Wulandari[4], Raodah[5], Cindy S[6]

[1,2,3,4,5,6]*Department of Mathematics, Faculty of Sciences, Universitas Cokroaminoto Palopo*

Received: 10 November 2025

Accepted: 09 December 2025

***Correspondent Email:** emm.emm92@gmail.com.

Abstract:
Students' interest in continuing their education to a higher level is an essential indicator of improving the quality of human resources. However, the factors influencing this interest are diverse, encompassing academic, social, and economic aspects. This study aims to predict the interest of students at SMAN 6 Luwu Timur in continuing their studies at university using the Naive Bayes algorithm. Data were collected through questionnaires administered to students, covering attributes such as tuition fees, parental income, college location, motivation, environmental support, and the 94 12th-grade s' desire to continue. The analysis process included data cleaning, transformation of categorical variables, and dividing the data into training and test data. The Naive Bayes algorithm was used to classify students into two categories: "interested in continuing" and "not interested in continuing." The results showed that the model had an accuracy of 66.67%, indicating quite good performance for categorical education data. Sosio-environment and motivation proved to be the most influential factors influencing interest in continuing studies. These results suggest that the Naive Bayes algorithm can be an effective predictive tool for early identification and educational strategy design, encouraging students to progress to higher levels.
Keywords: Naive Bayes, Prediction Of Interest In Further Studies, Educational Analytics, Data Mining, High School Students

1. INTRODUCTION

Education plays a crucial role in enhancing the quality of human resources and driving a nation's progress. Through education, individuals not only develop thinking skills and technical skills, but also shape the character and capacities needed to face life's challenges and enter an increasingly competitive workforce. Senior high school is a particularly crucial stage in the educational cycle, as it marks a significant turning point for students. After completing this level, they face a vital decision: continuing their education in higher education or pursuing alternative paths, such as employment or entrepreneurship. This decision is not solely personal, but is heavily influenced by various factors—both internal and external to the student.

First, learning motivation is a key factor in determining the extent to which students view further education as a viable opportunity. Research in Indonesia by Khadijah et al. (2017) found that learning motivation is the most significant variable influencing students' interest in pursuing higher education. Intrinsic motivation, such as the desire for intellectual growth, and extrinsic motivation, including career prospects, play a crucial role in shaping aspirations for further education (Wang et al., 2024). Furthermore, parental support has also been shown to be crucial. A study by Kwartawaty et al. (2025) found that family support has a significant influence on institutional perceptions, which in turn affect students' intentions to pursue higher education. Family economic conditions are also a determining factor—students from families with higher incomes or parents with better educational backgrounds tend to be more interested in pursuing higher education (Dalimunthe et al., 2024). Fehintola et al. (2025) also explain that the school environment, peers, and community also shape students' choices; for example, positive or negative peer influence can trigger the decision to continue or discontinue formal education. In research by Norawati et al. (2022), social environmental factors and motivation were shown to jointly influence students' interest in pursuing further studies.

However, despite extensive research into these factors, schools often struggle to provide students with appropriate guidance. This is often because an understanding of the combination of motivational, social, economic, and institutional factors has not been systematically integrated. Schools often focus solely on academic aspects or report card grades without considering students' socio-psychological backgrounds. As a result, interventions tend to be general and non-specific, even though each student has a unique profile and needs. Therefore, a system is needed that can assist schools in analysing student data objectively and comprehensively, so that their inclinations and interests in further education can be identified early and targeted.

With advances in information technology and data analytics, the use of data mining methods has become a promising solution for conducting large-scale analyses of student variables and discovering hidden patterns within them. Through data mining analysis, educational institutions can uncover relationships between variables that may not be apparent through conventional analysis and then utilise these findings for more adaptive and evidence-based coaching strategies (Ramos et al., 2021). According to Kalita et al. (2025), the application of data mining techniques in education—known as educational data mining (EDM)—can help educational institutions predict student performance and behaviour more accurately, allowing for early and targeted interventions.

One algorithm widely used in classification and prediction processes is the Naive Bayes algorithm. This algorithm is based on Bayesian probability theory and is known to have quite accurate prediction capabilities with efficient and straightforward computational processes (Kumar et al., 2024). However, it is important to note that Naive Bayes, like any other algorithm, has its limitations. For instance, it assumes that all features are independent, which may not always be the case in real-world scenarios. In the educational context, Naive Bayes has proven effective in predicting learning outcomes, the risk of dropping out of school, and students' interest in continuing their studies (Fitriani, 2024). Fitriani demonstrated that Naive Bayes achieved an accuracy of over 94% in classifying students' tendency to drop out.

In these studies, the attributes used included learning motivation, parental support, academic achievement, economic conditions, and peer influence, all of which are also relevant to the social context of students in Indonesia. This confirms that the Naive Bayes algorithm can be used to classify students, such as those who are "interested in continuing" versus those who are "not interested in continuing," based on these attributes.

The implementation of the Naive Bayes algorithm offers several practical advantages. First, because this method is probability-based, the resulting output can be clearly interpreted as the probability that a student will make a particular decision. This clarity provides reassurance about the interpretability of the output algorithm's Sujata, a crucial aspect in the field of educational data mining and prediction modelling.

Second, this algorithm works well with both categorical data and a combination of numeric and categorical data (after transformation), making it suitable for use in educational contexts that involve non-numeric variables. Third, the prediction results can be used by schools as a tool for early identification of low-risk students pursuing higher education. This process involves setting a threshold probability, below which a student is considered 'low-risk'. Once these students are identified, schools can implement interventions such as counselling, scholarships, or motivational learning training to help these students overcome potential barriers to higher education (Ermilian & Nugroho, 2024).

The application of the Naive Bayes method offers schools the opportunity to transition from a reactive to a proactive guidance system. By leveraging the power of data analysis for evidence-based decision-making, this approach can inspire a more strategic and measurable method for increasing student participation in higher education. The results of this classification can serve as a basis for educational institutions to design policies that are not only reactive but also proactive in nature.

Based on this, this study aims to apply the Naive Bayes method to determine the interest of high school students in continuing their education at SMAN 6 Luwu Timur, Indonesia. Through the analysis of final-year student data using the aforementioned variables, this study aims to develop a practical and actionable prediction model for the school. The results are expected to help schools understand student interest patterns more deeply, provide more targeted guidance recommendations, and inform educational policy decisions. Given the specific local characteristics, this study is also expected to contribute to the educational literature in Indonesia by utilising data mining methods to predict learning interests and inform further educational decisions.

Overall, by integrating an understanding of the factors influencing interest in continuing education with the application of the Naive Bayes algorithm as an analytical tool, this study seeks to provide a framework that is both theoretical and practical for schools. Thus, schools can enhance their role in facilitating students' transition from high school to higher education in a more optimal and evidence-based manner.

2. MATERIALS AND METHODS

This research employs a quantitative approach, utilising a computational experimental method based on data mining. The primary objective of the study is to develop and test a predictive model of student interest in continuing their education to higher education using the Naive Bayes algorithm. The quantitative approach was chosen because this research involves processing numerical and categorical data and statistically testing the model's accuracy.

Naïve Bayes is an algorithm used to classify data into specific categories. This algorithm employs statistical techniques to estimate the probability that data belong to a particular class. This method is widely used in data classification because it offers high speed and efficiency, especially when handling large datasets. Ananta et al. (2025) explain that the Naïve Bayes algorithm is capable of processing high-dimensional datasets, assuming independence between attributes, which is formulated as follows.

$$P(C|F_1 \dots F_n) = \frac{P(C)P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)} \quad (1)$$

Variable C describes the class, and variables F1…Fn describe the characteristics required for classification. This equation represents the Naive Bayes Theorem model that will be applied in the classification process.

The study was conducted at SMAN 6 Luwu Timur, with a population of all 12th-grade students. Five main attributes served as classification variables: tuition fees, parental income, college location, motivation, and environment.

Tuition fees, parental income, college location, motivation, and environment are labelled attributes, as shown in the following table.

Table 1: Data Description

Attribute	Description
Tuition Fees	Expensive
	Affordable
Parental Income	High
	Middle
	Low
	Poor
College Location	Far
	Near
Motivation	Family
	Teachers
	Community
	Peers
environment	Supportive
	Unsupportive

The analysis stages were conducted using Altair AI Studio software and the R programming language. The systematic research steps are as follows:

1. Data Cleaning
   This stage included checking for missing values, duplicate data, and format inconsistencies. Incomplete or illogical data were removed or adjusted using imputation techniques. After cleaning, 94 valid data sets were obtained for analysis.
2. Data Transformation
   Numeric variables, such as academic grades, are converted into categories (high, medium, low) using discretisation techniques. All attributes are converted to factor types to be compatible with the Naive Bayes algorithm.
3. Data Splitting
   The dataset is divided into two parts using the hold-out method, namely: 80% training data (training set) and 20% testing data (testing set). This division was conducted randomly (through random sampling) to maintain a balanced class distribution.
4. Model Building
   The model is built using the Naive Bayes algorithm (Gaussian NB and Multinomial NB) depending on the attribute distribution. This algorithm calculates the posterior probability using the formula:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad (2)$$

Where: $P(H|X)$ is the probability of students being interested in continuing their studies based on the attribute X; $P(X|H)\,P(H)$ is the probability of the attribute appearing in a particular class. $P(X)$ Is the initial (prior) probability of each class.

5. Model Evaluation
   Model evaluation is conducted to assess the accuracy and performance of the classification model using metrics such as accuracy, precision, and recall, using the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

6. Attribute Weighting using Information Gain
   After modelling the dataset, attribute weighting is performed using the Information Gain (IG) method to identify the most influential features in determining student interest. Information Gain measures the reduction in entropy after the dataset is split based on a particular attribute. Attributes with higher IG values are given greater weights in the classification model. The formula for Information Gain is as follows:

$$E(S) = \sum_{i=1}^{n} pi * log_2\, pi \quad (6)$$

$$IG(A) = E\,(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * E(S_i) \quad (7)$$

Where:
$IG(A)$ is the information Gain of the attribute A
$E(S)$ is the overall entropy of the dataset
S_v: subset of data where attribute Ahas value v
$\frac{|S_i|}{|S|}$ is the proportion of the subset S_v to the total dataset
The weight of each attribute is then normalised to produce the Weighted Attribute (WA**)** used in the classification process, defined as:

$$W\,(A_i) = \frac{IG(A_i)}{\sum_{j=1}^{n} IG(A_j)} \quad (8)$$

Where:
WA_i: Normalised weight of attribute A_i.
$IG(A_i)$: Information Gain value of attribute A_i.
$\sum_{j=1}^{n} IG\,(A_j)$: Total Information Gain of all attributes used in the model.
This formula ensures that each attribute weight is scaled proportionally based on its importance. This weighting process ensures that attributes contributing more to class differentiation have a more substantial influence on the Naive Bayes probability calculation.

The final result of this research is a predictive model that identifies the factors most influential in influencing students' interest in continuing their education at university. This model is expected to serve as a basis for schools in providing career guidance and determining strategies to increase students' interest in continuing their education.

3. Results and Discussions

The initial step in the analysis process is data cleaning, which is performed to ensure data integrity and quality before the modelling stage. This process includes correcting structural errors, such as duplicate and inconsistent entries, and handling missing values through imputation or removing cases with incomplete information.

The data used in this study consisted of 94 records, each containing five predictor attributes: tuition fees, environment, motivation, parental income, and college location. All attributes were categorised according to modelling requirements. Preprocessing was then performed, including checking for missing values, ensuring category consistency, and aligning data to eliminate ambiguity in the probability calculations. After the data preparation process was complete, the dataset was divided into training data (80%) and test data (20%) using random sampling to maintain representative class distribution.

In the modelling stage, the Naïve Bayes algorithm was used to calculate prior probabilities for each student interest class, followed by calculating conditional probabilities for each attribute category based on the training data. This combination of probabilities formed a predictive model based on a probabilistic function that estimated students' likelihood of pursuing higher education. When tested using the test data, each instance was evaluated through a posterior probability calculation, and the predicted class was determined based on the highest probability value. These results served as the basis for identifying students' interest in pursuing higher education.

The model implementation and testing process was conducted using Altair AI Studio. The process begins by importing the dataset and assigning the "Interest" attribute as the label. The Split Data operator is used to divide the dataset into training and test portions. Naïve Bayes modelling is performed using the Training operator, then the model is applied to the test data using the Apply Model operator. Performance measurement is performed using the Performance operator, which generates an accuracy metric by comparing the actual labels with the predicted results.

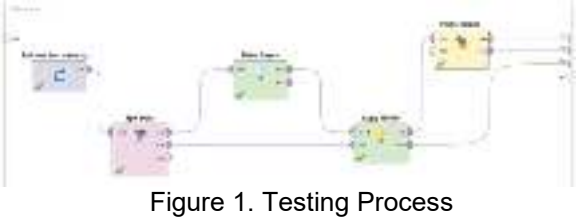

Figure 1. Testing Process

Based on Figure 1, all operators are arranged in an integrated workflow, resulting in output in the form of model accuracy and predicted probability estimates. These results are then analysed to assess the effectiveness of the Naïve Bayes algorithm in predicting students' interest in continuing their education to college, specifically among students at SMAN 6 Luwu Timur.

Figure 2. Performance vector test results

Figure 2 shows that the Naïve Bayes model achieved an accuracy rate of 66.67%, a precision of 68.75%, and a recall of 91.67%. These three evaluation values illustrate the model's ability to classify students' interest in continuing their education at the university level. Although the model's accuracy was moderate—with approximately two-thirds of predictions matching the actual situation—a deeper understanding of the precision and recall values provides a more comprehensive picture of the quality of the projections.

Further analysis using a confusion matrix provides a more detailed explanation of how the model makes decisions. Based on the model's output, predictions for the "continue" category showed that 11 data points were correctly classified as students genuinely interested in continuing their education. This indicates that the majority of the model's optimistic predictions came from this category. However, the model also made five errors, where it predicted that a student would continue their education when, in reality, the student showed no interest in doing so. These classification errors are recorded as false positives and directly contribute to the model's precision. With a precision of 50%, the interpretation is that only half of the "continue" predictions actually matched the actual situation.

Meanwhile, for the "do not continue" category, the model was only able to correctly predict one instance, indicating that it still struggled to recognise patterns of students who did not wish to continue their education. Furthermore, the model also incorrectly predicted one example that should have been in the "continue" category but was classified as "do not continue." This resulted in a very low recall value of only 16.67%, indicating that the model failed to recognise a significant portion of students who honestly did not wish to continue their education. Low recall suggests that the model has a strong tendency to predict students into the majority class, in this case, the "continue" class.

This combination of low precision and recall values indicates that, despite its seemingly high accuracy, the model was not effective in

distinguishing between the two classes. This suggests a possible class imbalance problem in the training data, where the number of students interested in continuing their education significantly exceeds that of those who are not interested. In such conditions, the model tends to "play it safe" by predicting the majority class, resulting in seemingly high accuracy but actually inadequate predictive performance.

Furthermore, model performance can also be affected by the quality and relevance of the attributes used. If the attributes are unable to capture factors that significantly differentiate student interests, probabilistic models like Naïve Bayes will struggle to form accurate patterns. This situation opens up opportunities for improvement through attribute selection (feature selection).

In addition to model testing, this study also conducted attribute weight analysis using the Weights by Information Gain operator to identify the most influential variables in shaping the model's decisions. This operator calculates the amount of information each attribute provides in distinguishing between "advanced" and "non-advanced" classes. The higher the information gain value of an attribute, the greater its contribution to the classification process.

The test results, which identified the most influential attributes on student interest labels, are presented in Table 2.

Table 2. Attribute Weight Results

Attribute	Weight
Tuition Fees	0,002
College Location	0,002
Parental Income	0,010
Motivation	0,030
Environment	0,088

The results from Table 2 show the level of contribution of each variable in influencing students' decisions to pursue higher education. These weight values reflect the extent to which each attribute contributes to the model's ability to distinguish between classes that are "advancing" and those that are "non-advancing." Therefore, the higher the weight, the more critical the attribute is in the classification process.

The attribute with the highest weight is Environment (0.088). This relatively high value indicates that environmental factors—including the school environment, social interactions, and learning atmosphere—make the most contribution in influencing student interest. A conducive, supportive, and encouraging environment for educational aspirations has been shown to play a significant role in shaping students' orientation toward higher education. This finding is also consistent with various previous studies that emphasise the importance of academic climate and social support as predictors of educational aspirations.

The following attribute is Motivation (0.030), which also has a reasonably significant weight value. This confirms that students' internal motivation, including intrinsic motivation such as the desire to develop, and extrinsic motivation, such as career aspirations, significantly influences their decision to continue their education. Although its value is lower than that of environmental factors, motivation remains a key factor with a clear pattern in the data.

Parental Income (0.010) has a medium weighting, indicating that family economic factors remain relevant but are not overly dominant. This could mean that while family financial ability influences the decision to continue college, this factor may not be as strong as motivation and environment. Some students may receive alternative support, such as scholarships, so economic factors are not the sole determinant of their academic success.

Meanwhile, Tuition Fees (0.002) and College Location (0.002) have very low weightings. These small values indicate that these two variables do not contribute significantly to the model's classification process. This means that information regarding tuition fees and campus location does not considerably differentiate between students who are interested in continuing their studies and those who are not. This could be due to students' perception that costs can be covered with educational financial assistance, or because access to higher education is relatively equitable, making location a less significant factor in their decision.

Overall, the Information Gain calculation results indicate that environmental factors and motivation are the primary determinants of student interest. In contrast, economic factors and external aspects such as cost and campus location have a lesser influence. These findings can serve as a basis for schools in designing intervention strategies, such as improving the academic climate, strengthening motivational counselling programs, and providing education about scholarship opportunities, to encourage students' interest in continuing their education in higher education more effectively.

4. CONCLUSIONS

Based on the analysis using the Naive Bayes algorithm and attribute weighting with the Information Gain method, it can be concluded that the model's ability to predict student interest in continuing their education to higher education remains limited. The model achieved an accuracy of 66.67%, with a precision of 50% and a recall of 16.67%. This suggests that while the model is quite capable of predicting some students who are interested in continuing their education (the "advanced" class), its ability to identify all genuinely interested students remains low. This is reflected in the distribution of predictions in the confusion matrix, where the model correctly predicted 11 students who continued, incorrectly predicted five students, and only slightly identified students who

did not continue. The low recall indicates significant challenges in capturing pattern variations in the positive class, which is likely influenced by an imbalanced data distribution or attribute patterns that do not sufficiently differentiate the two classes.

In terms of variable contribution, the Information Gain results indicate that the Environment factor (0.088) is the most influential attribute in determining student interest, followed by Motivation (0.030) and Parental Income (0.010). Meanwhile, Tuition Fees and College Location, each with a weight of 0.002, had minimal impact on the classification process. This finding confirms that students' decisions to continue their education are more influenced by internal and social factors, particularly a supportive learning environment and the students' own motivation, than external factors such as cost and college location.

Overall, the analysis results indicate that the Naïve Bayes model can provide an initial overview of student interest patterns. However, its performance still needs improvement, either through data balancing, the addition of more informative attributes, or the exploration of alternative algorithms. Furthermore, the Information Gain findings provide important insight that school interventions should focus more on improving the learning environment and strengthening student motivation, as these two factors are the most significant in influencing interest in continuing their education to college. Thus, this study not only provides an overview of the prediction model's performance but also offers a strategic direction for student development efforts at SMAN 6 Luwu Timur.

ACKNOWLEDGMENTS

The authors would like to thank the SMAN 6 Luwu Timur for their support, cooperation, and assistance throughout the research process. The availability of accurate data and ease of coordination with the SMAN 6 Luwu Timur were key factors in ensuring the smooth running of the research.

REFERENCES

Ananta, A., Wulandari, N., Mustawinar, B. H., & Putri, F. G. (2025). Klasifikasi Pembelian Produk Rumah Tangga Melalui Metode Naïve Bayes. indonesian journal of material and applied physics, 1(1), 16-21.

Andika, A. W., Nurhakim, L., & Andas, N. H. (2025). PENGGUNAAN DEEP LEARNING UNTUK MEMPREDIKSI KINERJA AKADEMIK DAN MEMBERI DUKUNGAN YANG TEPAT BAGI SISWA. SIBATIK JOURNAL: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, Dan Pendidikan, 4(7), 1647-1664. https://doi.org/10.54443/sibatik.v4i7.3152

Dalimunthe, A. S., Sitepu, E., Yuriska, K., & Ichayu, V. (2024). Analisis Faktor-Faktor yang Mempengaruhi Minat Mahasiswa Melanjutkan Pendidikan ke Perguruan Tinggi. Lebah, 18(1), 1-10.

Ermillian, A., & Nugroho, K. (2024). Perancangan Model Deteksi Potensi Siswa Putus Sekolah Menggunakan Metode Logistic Regression Dan Decision Tree. Jurnal Informatika: Jurnal Pengembangan IT, 9(3), 281-295. https://doi.org/10.30591/jpit.v9i3.8007

FEHINTOLA, V. A., OGUNNIYI, T. M., & FEHINTOLA, F. C. (2025). Peer Influence and Home Environment as Predictors of School Dropout Risk among Senior Secondary Students in Oyo East LGA, Oyo State, Nigeria.

Fitriana, S. (2024). a PREDIKSI SISWA PUTUS SEKOLAH DAN KEBERHASILAN AKADEMIK MENGGUNAKAN MACHINE LEARNING: Prediksi Siswa Putus Sekolah dan Keberhasilan Akademik. The Indonesian Journal of Computer Science, 13(6). https://doi.org/10.33022/ijcs.v13i6.4453

Kalita, E., Oyelere, S. S., Gaftandzhieva, S., Rajesh, K. N., Jagatheesaperumal, S. K., Mohamed, A., ... & Ali, T. (2025). Educational data mining: a 10-year review. Discover Computing, 28(1), 81. https://doi.org/10.1007/s10791-025-09589-z

Khadijah, S., Indrawati, H., & Suarman, S. The Factors That Influence Student's Interest in Continuing Higher Education. International Journal of Economic, Business & Applications, 2(1), 23–30. https://doi.org/10.31258/ijeba.11

Kumar, R., Goswami, B., Mhatre, S. M., & Agrawal, S. (2024). Naive bayes in focus: a thorough examination of its algorithmic foundations and use cases. Int. J. Innov. Sci. Res. Technol, 9(5), 2078-2081. https://doi.org/10.38124/ijisrt/IJISRT24MAY1438

Kwartawaty, N. N., Prajanti, S. D. W., Pramono, S. E., & Khafid, M. (2025). MULTIDIMENSIONAL FACTORS INFLUENCING SENIOR HIGH SCHOOL STUDENTS INTENTION TO PURSUE HIGHER EDUCATION: A STRUCTURAL EQUATION MODELLING APPROACH. Lex Localis, 23(10), 743–749. https://doi.org/10.52152/800972

Premalatha, N., & Sujatha, S. (2022). Prediction of students' employability using clustering algorithm: A hybrid approach. International Journal of Modeling, Simulation, and Scientific Computing, 13(06), 2250049.

Norawati, S., Zulher, Z., Arman, A., & Usman, U. (2022). Determinant Factors Affecting Student Interest In Continue Education To Higher Education. International Journal of Economics, Business and Accounting Research (IJEBAR), 6(4). https://doi.org/10.29040/ijebar.v6i4.8399

Ramos, S., Soares, J., Cembranel, S. S., Tavares,

I., Foroozandeh, Z., Vale, Z., & Fernandes, R. (2021). Data Mining Techniques for Electricity Customer Characterisation. Procedia Computer Science, 186, 475–488.

Wang, X., Dai, M., & Short, K. M. (2024). One size doesn't fit all: how different types of learning motivations influence engineering undergraduate students' success outcomes. International Journal of STEM Education, 11(1), 41. https://doi.org/10.1186/s40594-024-00502-6