Naive Bayes-based Drop Out Recommendation System in Vocational College

Kartika Candra Kirana*, Slamet Wibawanto, Heru Wahyu Herwanto, Wahyu Sakti Gunawan Irianto, Wahyu Nur Hidayat, Muhamad Aqshal

Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Indonesia
Email: kartika.candra.ft@um.ac.id

ABSTRACT

When a student does not pass the requirements to graduate from a vocational college, the drop out (DO) system is usually utilized. The Naive Bayes technique make it simple to learn how graduation requirements might be modelled on prior evidence. As a result, this study suggests utilizing Naive Bayes to develop a Drop Out Recommendation System. We used 210 test data and 840 training data from the Kaggle dataset "Prediction of dropping out of school" for the testing phase. The proposed approach uses the Bayes technique to predict a student`s likelihood of dropping out based on their GPA and course enrolment in two semesters. The two value categories, high and low, of the GPA range from 1.6 to 4.6. However, the standard for the courses that are enrolled is based on the middle of the credit range, which runs from 5 to 20. The effectiveness of the Bayes Method is assessed using accuracy calculations. The test data shows five out of the seven data models. The test yielded 197 correct test results out of 210, with a maximum accuracy of 93.8%. It can be concluded that the Bayes technique can be used to recommend dropout strategies.

Keywords: *Drop out, Naïve bayes, Recommendation, Artificial intelligences.*

1. INTRODUCTION

According on statistics from the Indonesian Ministry of Education and Culture, out of the 8.483.213 registered students, there were 602.208 dropouts in 2020 c 7% of students fail to finish school [1]. The results of this survey demonstrate that the data are evenly distributed over all of Indonesia's islands and fall within the 0.04- 0.15 range. D3 and D4 are the categories where dropout rates are the highest. This indicates the need for university to regulate vocational levels in order to foresee dropouts [2].

The use of data mining in education and learning analytics has seen significant technological advancement [3]. To provide alerts for the "dropout" issue, computational techniques like machine learning can be used. The supervised learning paradigm in machine learning is thought to give models the ability to practice mapping data features to certain patterns. The Naive Bayes Algorithm is one of the supervised learning techniques that takes the probability of occurrences from other events into account [4].

Several studies have shown the superior performance of Naive Bayes compared to comparison methods [5]. Naive bayes probability can measure the features that are most likely to affect output [6], [7]. The Naive Bayes algorithm performs better than logical algorithms because it presumes that all data is independent. As a result, all data model can be computed [8]. Shynarbek (2022) achieves the highest accuracy rate for Naive Bayes in Dropout Detection Research compared to random forests and decision trees, which reach 80%, 77%, and 80%, respectively [9]. Naive Bayes' accuracy for predicting dropout at Brazilian universities is 80%, which is comparable to KNN and random forest [10]. Furthermore, demonstrates the superiority of Naive Bayes over SVM, Logistic Regression, and Neural Networks with accuracy reached 0.96 [11]. How naive bayes may be used to forecast dropouts in Bangladesh on par with SVM [12].

Naive Bayes performs well on patterned independent data, according to prior studies. Data with ambiguous patterns have not been put to the test. In several earlier research, the effectiveness of naive Bayes was solely assessed without feature selection based on the validity

A. Kusumastuti et al. (eds.), *5th Vocational Education International Conference (VEIC 2023)*,
Advances in Social Science, Education and Humanities Research 813,
https://doi.org/10.2991/978-2-38476-198-2_99

of the problem. MOOC with prediction of failure by considering the features that affect course graduation [13]. It inspires us to map features that can predict study failure based on features that are considered important. The probability of non-patterned data is supposed to be mapped by naive Bayes-based rational feature selection.

This research's proposed "Naive Bayes-based Drop Out Recommendation System in Vocational College". The contribution of this research is evaluating naive Bayes-based rational feature selection performs on non-patterned data. The goal of this project is to be able to offer features that are utilized in conjunction with naive bayes to make recommendations for schools to warn students who are indicated to have dropped out. These recommendations will be based on the problem reasoning.

2. LITERATURE REVIEW

A multi-class Monte Carlo simulation to demonstrate the application of naive Bayes. In both independent and functionally dependent characteristics, Naive Bayes achieved good accuracy [4]. In order to identify student dropouts at Columbia University, compare logistic regression and naive Bayes using Watson analytics. The outcome demonstrates that binary logistic regression is utilized on data that are dependent on dichotomous data, but naive Bayes assumes that the data are independent from other data without considering the data model [8].

Contrasted random forests, decision trees, and naïve bayes for detection of the university dropouts. According to the research, an accuracy rate reached 80%, 77%, and 80%. Furthermore, the learning process takes longer with decision trees and random forests than with naïve bayes [9]. Using academic data from the University of Brazil, also contrasted Naive Bayes with KNN and random forest to stop students from quitting. Naive Bayes produces outcomes that are equal to those of KNN and random forest [10].

Demonstrates the superiority of Naive Bayes over SVM, Logistic Regression, and Neural Networks. On a scale of [0.1], the outcomes of the comparison of the four approaches were 0.96, 0.89, 0.88, and 0.95, respectively [11]. How naive bayes may be used to forecast dropouts in Bangladesh on par with SVM. This study demonstrates how crucial it is to choose factors that affect students' decisions to drop out of school [12].

Investigates how naïve bayes might be used to assess students' academic performance. This experiment demonstrates the 76% accuracy of naive bayes. Nine features are employed in this study to classify data. This study demonstrates that some characteristics are thought to have no direct bearing on students' academic standing. Because of this, feature selection is thought to be crucial for enhancing naïve bayes performance [14]. Artificial neural networks frequently use the Bayes Algorithm to compare potential input opportunities to the output that has been given [15]. At the University of Lima, investigated the application of decision trees and Bayes-based neural networks on 500 data. According to this test, utilizing Bayes has 7% better performance than using a decision tree [16].

Optimized the Bayes function with priority hierarchy's dropout. In order to create hierarchical priorities, variance in the Gaussian model was used. The proposed method is superior than the prior method. Stochastic-based dropouts to the Bayes function [17].

3. METHOD

In this study, a dropout recommendation system is proposed based on the Naïve Bayes method.

3.1 Data

Predictive information for successful and dropout students was adapted from Kaggle for the simulation data [18]. In order to mapping the circumstances opportunities, 840 data are used as training data. Furthermore, 120 data were utilized as test data. The dataset includes 34 inputs, including marital status, application mode, application order, course, daytime and attendance, and previous experience, nationality, qualification of the mother, qualification of the father, occupation of the mother, occupation of the father displaced, special educational needs, the debtor current with tuition fees, gender, holder of a scholarship the enrolment age, international, course 1st (credited), course 1st (enrolled), course 1st (evaluations), course 1st (approved), course 1st (grade), course 1st (without evaluations), and course 2nd (credited), course 2nd (enrolled), course 2nd (evaluations), course 2nd (approved) (without evaluations), rates of unemployment and inflation. We proposed naive Bayes-based rational feature selection performs on non-patterned data. The GPA has an impact on graduation [19]. Data is shown in table 1 and table 2.

Long (2012) demonstrates that the course taken has an impact on graduation [20]. This study uses three features from the dataset and bases its decision on the two prior studies. These features are: the course enrolment taken in semesters 1 and 2, as well as the GPA earned. There were two types of registered curricula: those with a lot of students and those with few students. The students enrolled courses from 5 to 20. The GPA is a scale of 1.6 to 4.6. The mean value is applied as the threshold which calculate as Equation 1.

$$\theta_i = \frac{\sum_{j=1}^{jmax} x_j}{n} \tag{1}$$

Table 1. Training Data.

GDP	GPA	Enrolled (Sem 1)	Enrolled (Sem2)	Total
Graduate	High	High	High	25
Graduate	High	Low	High	180
Graduate	High	Low	Low	6
Graduate	Low	High	High	32
Graduate	Low	High	Low	0
Graduate	Low	Low	High	489
Graduate	Low	Low	Low	21
Dropout	High	High	High	10
Dropout	High	Low	High	1
Dropout	High	Low	Low	68
Dropout	Low	High	High	37
Dropout	Low	High	Low	1
Dropout	Low	Low	High	4
Dropout	Low	Low	Low	251

Table 2. Confusion Matrix.

Actual Class	Predicted	
	Dropout	Graduate
Dropout	*TP*	*FN*
Graduate	*FP*	*TN*

Table 3. Testing Data.

GPA	Enrolled (Sem 1)	Enrolled (Sem 2)	Total
High	High	High	92
High	Low	High	3
High	Low	Low	34
Low	High	High	37
Low	High	Low	14
Low	Low	High	1
Low	Low	Low	30

θ_i is defined as a threshold θ of feature-i. The mean result, which is used as the threshold, is obtained by dividing all data-x from i to jmax by total data-n. Data is shown in table 3.

3.2 Naïve Bayes Algorithm

A classification technique based on a conditional probability is called Naive Bayes. The explanatory variables must be assumed to be independent when applying Naive Bayes algorithm. The Naive Bayes algorithm has the benefit of quick calculations when used on larger datasets. The Naive Bayes algorithm does not involve statistical testing, in contrast to logistic regression approach.

The input training data is divided into high and low groups before beginning to calculate the probability value. Low values are initialized to zero, while high values are initialized to one. Equations 2 and 3 respectively indicate the grouping of GPA and enrolled curriculum units.

$$GPA_j = \begin{cases} 1\ \theta_{ipk} > 2.5 \\ 0\ \theta_{ipk} \le 2.5 \end{cases} \quad (2)$$

$$Enroll(smt)_j = \begin{cases} 1\ \theta_{enrolled} > 10 \\ 0\ \theta_{enrolled} \le 10 \end{cases} \quad (3)$$

A threshold of 2.5 is used to group the jth-GPA ratings, whereas a threshold of 10 is used to group the jth-course enrolment. Both thresholds are obtained from calculating the mean training data using Equation 1. Once the data has been grouped, predictions are calculated using Equation 4 as the Naive Bayes algorithm. Graduate and Dropout are the two output categories for GPD.

$$P(a|b) = \frac{P(b|a)\,P(a)}{\sum P(b|a)\,P(a)} \quad (4)$$

The likelihood of user input- P(a) and the likelihood that both input and output- P(b|a) occur at the same time are compared to all of the possible outcomes to determine the probability of GDP depending on user input- P(a|b).

3.3 Evaluation

We mapped the comparison between estimated outcomes and actual output in the confusion matrix displayed in Table 2 in order to evaluate the effectiveness of Naive Bayes as a recommendation engine. Table 2 assumes that the correctly projected dropout class is TP and the correctly expected graduation class is TN. It is assumed that the mistakenly anticipated dropout class is FP and the incorrectly predicted graduate class is FP.

The accuracy, precision, recall, and F1-score were then plotted. For both the graduation class and the dropout class, accuracy is used to assess how accurately the results were predicted. Recall is used to compare the estimated dropout results to all actual dropout data, whereas precision is used to compare the estimated dropout results to all forecasted dropout data. Equation 5 illustrates an accuracy calculation.

$$acc = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

Based on the quantity of graduate-*TN* and dropout-*TP* forecasts for all data, accuracy-*acc* is determined. Equation 6 illustrates Recall.

$$Recall = \frac{TP}{TP+FN}. \tag{6}$$

Calculating recall involves comparing the projected dropout rate to the real dropout class, which comprises of *TP* and *FN*. Equation 7 illustrates Precision.

$$Precision = \frac{TP}{TP+FP}. \tag{7}$$

Precision calculated by contrasting the *TP* class with the projected dropout class, which is made up of *TP* and *FP*. F1 score is the mean of recall and precision as shown in Equation 8.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{8}$$

Specificity calculations are also performed to assess the capability of detecting graduate class as shown in Equation 9.

$$TNR = \frac{TN}{TN+FP}, \tag{9}$$

where the number of predictions of the correct graduate class-*TN* to all actual grads is used to compute the true negative rate (*TNR*) or specificity.

4. RESULT AND DISCUSSION

We map the test data first. Data for the test is chosen at random. Random selection aims to assess naïve Bayes' performance on non-patterned test data. In Table 3, the test table is displayed. The highest number, which reached 92, was dominated by high GPA and course enrolment. Three sets of circumstances (1) high GPA and low course enrolment; (2) low GPA and course enrolment, and (3) low GPA and high course enrolment-dominate the group of about 30 students. Then we present a confusion matrix to represent the test results, as seen in Table 4. A ratio of 1:1.2 for graduates to dropouts. The test results are displayed in Table 5 based on the result of confusion matrix and the computations of Equations 5–9.

The system's ability to recognize graduate and dropout classes is indicated by the accuracy value, which hits 93.8%. In the meantime, the recall value, which demonstrates the system's capacity to identify dropouts from all classes, reaches 88.4%. The recall value, which measures how well the system performs in identifying dropout classes for all data discovered by dropouts, however, reaches 93.5%. This demonstrates that the dropout class is not overrepresented in the system. The F1-Score, which hits 90.9%, illustrates how precision and recall are balanced. the graduate class's detection performance was also evaluated and registered a TNR of 92.7%.

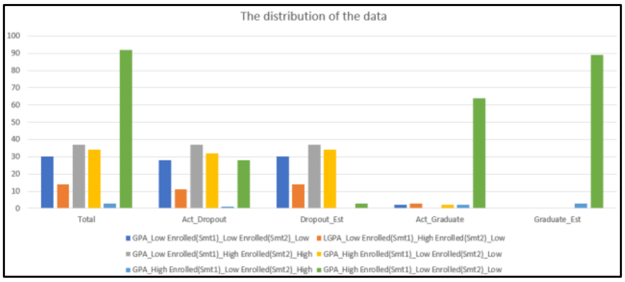
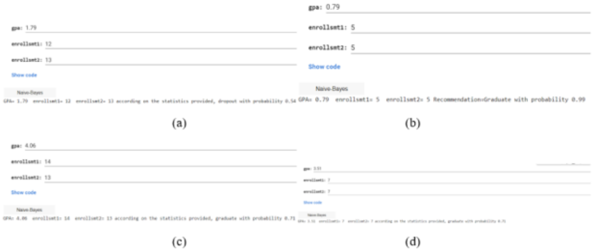


Figure 1. The Result of Confusion Matrix.

In further analysis we observe the distribution of the data on the predicted results shown in Figure 1. Further investigation reveals the data's dispersion regarding the results that were predicted, as seen in Figure 1. Bayes investigations using probability do better through historical probabilities. These four chances demonstrate this: There are four categories: GPA Low, Enrolled (Smt1) Low, Enrolled (Smt2) Low; LGPA Low, Enrolled (Smt1) High, enrolled (Smt2) Low; and GPA Low, Enrolled (Smt1) High, Enrolled (Smt2) High. The four opportunities demonstrate that there is more real evidence supporting dropout than graduate outcomes, but there is also real data supporting the possibility of graduate output. Naive Bayes cannot handle this situation effectively.


gpa: 1.79
enrollsmt1: 12
enrollsmt2: 13
Show code
Naive-Bayes
```
GPA= 1.79  enrollsmt1= 12  enrollsmt2= 13 according on the statistics provided, dropout with probability 0.54
```
(a)

gpa: 0.79
enrollsmt1: 5
enrollsmt2: 5
Show code
Naive-Bayes
```
GPA= 0.79  enrollsmt1= 5  enrollsmt2= 5 Recommendation=Graduate with probability 0.99
```
(b)

gpa: 4.06
enrollsmt1: 14
enrollsmt2: 13
Show code
Naive-Bayes
```
GPA= 4.06  enrollsmt1= 14  enrollsmt2= 13 according on the statistics provided, graduate with probability 0.71
```
(c)

gpa: 3.51
enrollsmt1: 7
enrollsmt2: 7
Show code
Naive-Bayes
```
GPA= 3.51  enrollsmt1= 7  enrollsmt2= 7 according on the statistics provided, graduate with probability 0.71
```
(d)

Figure 2. The Result of Observation (a) True Positive (b) False Negative (c) True Negative(d) False Positive.

The same evidence occurs under the following two circumstances: (1) GPA High, enrolled (Smt1) Low, enrolled (Smt2) High; and (2) GPA High Enrolled (Smt1) Low, enrolled (Smt2) Low, which both suggest a better likelihood of graduating than dropping out. As a result, information having the greatest chance of going undetected has different features.

The trial sample is shown in Figure 2. The GPA input in Figure 4 (a) is 1.79, and the number of students enrolled in the two semesters is 12 and 13, respectively. As a dropout is seen in both the actual and anticipated output, it is classified as a true positive. Figure 4 (b) shows that the GPA input reaches 0.79 and that the total number of credits taken in both semesters is 5. The outcome is actually a dropout, and graduates are found, hence it is classified as a false negative.

The GPA input in Figure 4 (c) is 4.06, and the number of students enrolled in the curriculum in both semesters is 14 and 13, respectively. The three inputs are the high categorized. The true negative is applied since the real output is a graduate and is recognized as such. Figure 4 (d) shows that the GPA input was 3.51 and that the total number of credits taken in both semesters was 7. The three inputs are the high categorization. Graduates were the genuine output, and when they were discovered, they were classified as false positives. Based on these trials, the process of considering conditions beyond the highest probability needs to be designed to improve naive bayes.

Table 4. The Result of Confusion Matrix.

Actual Class	**Predicted**	
	Dropout	**Graduate**
Dropout	*108*	*13*
Graduate	*7*	*89*

Table 5. The Result of Confusion Matrix.

Test	**Result (%)**
Accuracy	93.8
Recall	88.6
Precision	93.5
F1-Score	90.9
TNR	92.7

5. CONCLUSION

In this study, naive bayes is recommended as a method for vocational dropout students. The curriculum that is enrolled in the first and second semesters as well as the most recent GPA are inputs that are used as material for consideration. The Graduation and Dropout outputs are the ones that are defined. 120 test data were assessed using 840 training data, with metrics of accuracy, recall, precision, F1-score, and TNR reaching 93.8%, 88.6%, 93.5%, 90.9% and 92.7%. It can be concluded that the Bayes technique can be used to recommend dropout strategies. Based on several trials, the process of considering conditions beyond the highest probability needs to be designed to improve naive bayes in overfitting data.

ACKNOWLEDGMENTS

The head of the researcher's thanks goes to the research members who have contributed to the preparation of the research manuscript. Thanks also go to the Devastators Team for sharing the dataset for this research. Funding for this research is funding from the Department of Electrical Engineering and Informatics.

REFERENCES

[1] Anonymous, Statistik Pendidikan Tinggi (Higer Education Statistic), Kemendikbud, 2020.

[2] R. M. Isiaka and S. O. Abdulsalam, A Machine Learning Approach to Dropout Early Warning System Modeling, Int. J. Adv. Stud. Comput. Sci. Eng., vol. 8, 2019, pp. 1–12.

[3] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, Predicting Student Performance Using Advanced Learning Analytics, Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 415–421. DOI: 10.1145/3041021.3054164.

[4] I. Rish, An Empirical Study of the Naive Bayes Classifier, Phys. Chem. Chem. Phys., vol. 3, 2001, pp. 4863–4869. DOI: 10.1039/b104835j.

[5] M. Otair, S. Zacout, L. Abualigah and M. Omari, Chapter Eleven-Hybrid Arabic Classification Techniques Based on Naïve Bayes Algorithm for Multidisciplinary Applications, Artificial Neural Networks for Renewable Energy Systems and Real-World Applications, pp. 239–265. 2022. DOI: https://doi.org/10.1016/B978-0-12-820793-.00004-5

[6] A. A. M. Shaheen and N. Naheed, Relevance-Diversity Algorithm for Feature Selection and Modified Bayes for Prediction, Alexandria Eng. J., vol. 66, 2023, pp. 329–342, DOI: https://doi.org/10.1016/j.aej.2022.11.002.

[7] S. Farhana, Classification of Academic Performance for University Research Evaluation by Implementing Modified Naive Bayes Algorithm, Procedia Computer Science, pp. 224–228, 2021, DOI: https://doi.org/10.1016/j.procs.2021.10.077.

[8] B. Perez, C. Castellanos, and D. Correal, Applying Data Mining Techniques to Predict Student Dropout: A Case Study, IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI), pp. 1–6, 2018, DOI: 10.1109/ColCACI.2018.8484847.

[9] N. Shynarbek, A. Saduakassova, N. Sagyndyk, Y. Sapazhanov and A. Orynbassar, Forecasting Dropout in University Based on Students' Background Profile Data Through Automated Machine Learning Approach, International Conference on Smart Information Systems and Technologies (SIST), pp. 1–5, 2022 doi: 10.1109/SIST54437.2022.9945715.

[10] K. J. de O. Santos, A. G. Menezes, A. B. de Carvalho and C. A. E. Montesco, Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout, vol. 2161–377X, IEEE 19th International Conference on Advanced Learning Technologies (ICALT), 2019, pp. 207–208. DOI: 10.1109/ICALT.2019.00068.

[11] N. Shynarbek, A. Orynbassar, Y. Sapazhanov and S. Kadyrov, Prediction of Student's Dropout from a University Program, International Conference on Electronics Computer and Computation (ICECCO), 2021, pp. 1–4. DOI: 10.1109/ICECCO53203.2021.9663763.

[12] S. A. Ahmed, M. A. Billah, and S. I. Khan, A Machine Learning Approach to Performance and Dropout prediction in Computer Science: Bangladesh Perspective, International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1–6, DOI: 10.1109/ICCCNT49239.2020.9225356.

[13] V. Muthukumar and D. B. N., MOOCVERSITY - Deep Learning Based Dropout Prediction in MOOCs over Weeks, J. Soft Comput. Paradig., vol. 2, 2020, pp. 140–152. DOI: 10.36548/jscp.2020.3.001.

[14] Haviluddin, N. Dengen, E. Budiman, M. Wati, and U. Hairah, Student Academic Evaluation using Naïve Bayes Classifier Algorithm, 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT), 2018, pp. 104–107. DOI: 10.1109/EIConCIT.2018.8878626.

[15] N. Kesireddy, W. Khokher, F. Safi, W. Barnett, and R. Assaly, Detection of Ventilator-Associated Events Among Adults Using Naive Bayes Classifer, Chest, vol. 160, 2021, p. A1418. DOI: https://doi.org/10.1016/j.chest.2021.07.1297.

[16] E. C. Medina, C. B. Chunga, J. Armas-Aguirre, and E. E. Grandon, Predictive Model to Reduce the Dropout Rate of University Students in Perú: Bayesian Networks vs. Decision Trees, 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020, pp. 1–7. DOI: 10.23919/CISTI49556.2020.9141095.

[17] Y. Liu, W. Dong, L. Zhang, D. Gong, and Q. Shi, Variational Bayesian Dropout with a Hierarchical Prior, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7117-7126. DOI: 10.1109/CVPR.2019.00729.

[18] V. Realinho, J. Machado, L. Baptista and M. V. Martins, Predict Students' Dropout and Academic Success, UCI Machine Learning Repository, vol. 7, 2022, pp. 1-17. DOI: https://doi.org/10.3390/data7110146.

[19] J. A. Gipson, Predicting Graduation and College Gpa: a Multilevel Analysis Investigating the Contextual Effect of College Major, Department of Educational Studies West Lafayette-Indiana, 2018.

[20] M. C. Long, D. Conger and P. Iatarola, Effects of High School Course-Taking on Secondary and Postsecondary Success, Am. Educ. Res. J., vol. 49, 2012, pp. 285–322. DOI: 10.3102/0002831211431952.