Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

Classification of Student Grading Using Naïve Bayes Method with Under-sampling Approach to Handle Imbalance

Alif Abdul Aziz[1*], Budi Prasetiyo[2]

[1,2]*Department of Computer Science, Universitas Negeri Semarang, Indonesia*

DOI: https://doi.org/10.52465/joiser.v3i1.537

Received 25 January 2025; Accepted 30 January 2025; Available online 30 January 2025

Article Info

Keywords:
Classification;
Naïve Bayes;
Under-sampling;
Student grades;
Data analysis

Abstract

This study explores the application of the Naive Bayes classification method to predict student grades based on important attributes such as timeliness of assignment submission, attendance rate, and quality of work. This research uses a dataset that includes three attributes, namely timeliness of submission, attendance level in learning, and evaluation of the quality of assignments collected by students. The pre-processing is performed to clean the data, followed by an under-sampling stage to balance the class distribution. Then, the classification model is evaluated and tested using specific data samples to measure prediction accuracy. The results showed a significant improvement in model accuracy after applying under-sampling, highlighting the importance of handling data imbalance in predictive analysis. The implications of these findings are not only relevant in the context of higher education, but also offer opportunities for further development in data-driven decision-making in various fields.

1. Introduction

Higher education is part of the education system that has the task of educating the nation's life and improving the quality of teaching and learning interactions. Indonesian universities are equipped with a curriculum management system, which serves as a foundation for higher education management policies. This system embodies a philosophy that influences societal formation, academic ambiance, instructional patterns, and the overall atmosphere resulting from teaching and learning interactions [1].

The college student is the result of the educational process in higher education which is a milestone for the formation of the nation's future generation. The importance of the quality of higher education is highlighted because of the close relationship between students and these educational institutions. Universities act as a place of education, while students are the product of the education system. It is expected that students, as products of higher education, can have a significant positive impact on the

* ***Corresponding Author:***

Alif Abdul Aziz,
Department of Computer Science,
Universitas Negeri Semarang,
Semarang, Indonesia.
Email: alifabdulaziz24@students.unnes.ac.id

progress of the nation and state. As the young generation, students reflect the direction of the country's future development [2].

Higher education plays an important role in shaping one's career and professional future. In addition to providing the necessary knowledge and skills, higher education institutions are also responsible for assessing students' learning progress. This evaluation is often done through various forms of assessment, one of which is the awarding of grades. However, the process of giving grades is not an easy thing. Lecturers are often faced with challenges in assessing student performance objectively and consistently [3].

In practice, factors such as timeliness of assignment submission, level of attendance in learning, and quality of work submitted by students can affect the final assessment given by lecturers. Therefore, it is important to comprehensively understand and identify these factors to ensure fairness in the assessment process.

In this digital era, technology and data analytics can make a significant contribution to understanding and optimizing the grading process. By utilizing advanced data analysis techniques, we can explore patterns hidden in student assessment data and identify factors that influence grading more precisely [4].

One of the popular methods in data analysis is the classification method. This method allows us to predict or classify grade classes based on the attributes of the students. With classification methods, we can identify patterns in student grading data and identify factors that influence grading more precisely [5], [6]. One of the popular classification methods is the Naive Bayes method [7]. Previous research [8] was conducted using the Naive Bayes method to predict prospective students who will re-register. The study used two datasets, namely a dataset with unbalanced classes and a dataset with balanced classes, which were compiled using the Under-sampling method. The results showed that the Naive Bayes method can improve accuracy on datasets with unbalanced classes up to 63.83%.

Another study [9] employed the Naive Bayes approach to evaluate the precision of a decision support system for scholarship acceptance. The decision support system has an accuracy rate of 92.7%, which shows that the scholarship acceptance decision support system using the Naive Bayes method is very feasible to use. Thus, the Naive Bayes technique, also utilized in this investigation regarding student grading classification, demonstrates its efficacy not only in addressing data imbalance but also in yielding dependable and valuable outcomes across diverse decision-making scenarios.

2. Method

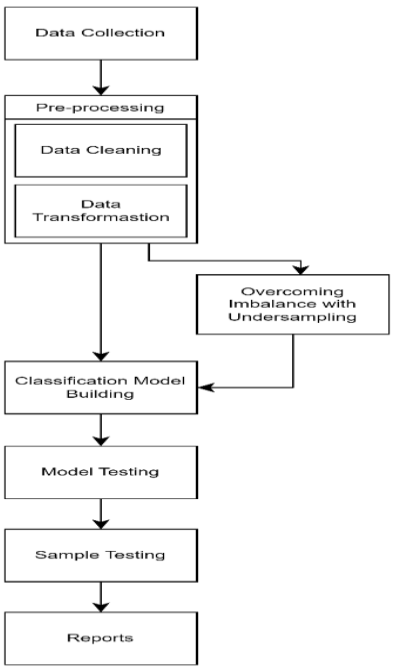

Figure 1. Research method

This research process begins with collecting data on student grades based on aspects of attendance, timeliness of submission, and quality of work. This process is then followed by data cleaning, data

transformation, and class balancing using under-sampling techniques. The next steps include building a classification model, analyzing mutual information, testing the model, and experimenting using various scenarios. Evaluation of experimental results is an important part of this research, and the process ends with reporting and publication. The flow of the research method can be seen in Figure 1.

2.1. Data Collection

The dataset employed in this study was acquired through the generation of a dataset using artificial intelligence algorithms, resulting in files formatted as CSV. This dataset includes three supporting attributes that have an important role in the analysis, namely the timeliness of submission, the level of attendance in learning, and the evaluation of the quality of work submitted by students. The timeliness of submission is divided into two categories, namely late and on-time, while the attendance rate is classified as fair (when attendance reaches 75%), good (when attendance ranges between 75%-85%), and excellent (when attendance exceeds 85%). On the other hand, the quality of work is rated in three categories, namely poor, fair, and good. Furthermore, there is one main target variable that is the focus of the analysis, namely student grades, which are classified into A, AB, and B categories.

2.2. Data Pre-processing

Data pre-processing is the process of processing data before analysis. This process aims to produce data that matches the criteria for the analysis to be carried out [10]. Data pre-processing can include several steps, such as cleaning, normalization, transformation, and feature selection [11]. The pre-processing step is carried out to transform the data into a format that suits the needs of the system. In the context of this research, several pre-processing steps are performed, including data cleansing and data transformation so that it can be processed further. In the data cleaning stage, missing value cleaning is done by removing rows that have missing values. In the data transformation stage, data modification is carried out by converting the target data type into categories and modifying other data using one-hot techniques to improve the model's ability to understand and process information.

2.3. Under-sampling

In this step, the transformed dataset is adjusted for distribution using the under-sampling method to balance the dataset classes. Under-sampling is a data collection technique used to reduce the amount of data from more classes in imbalanced data. This technique can be used to reduce data redundancy and noise, which can improve analysis performance. Under-sampling can be done by removing or reducing data from more classes, such as with the Random Under-Sampling (RUS) method [12] and customized instance random under-sampling [13].

2.4. One-hot Encoding

One-hot encoding is an important technique in data processing that converts categories into unique numerical representations. Through this technique, the categories are converted into binary vectors with different values, allowing data processing systems to recognize and process information more effectively. The use of one-hot encoding is extensive in the development of various information systems, such as data processing systems, pattern recognition, testing, and image analysis. By utilizing one-hot encoding, these systems can improve their ability to process data and make more accurate predictions [14].

2.4. Naïve Bayes Classification

Naive Bayes is a classifier method that uses Bayes' theorem to calculate the likelihood of a document with a certain level of certainty based on the likelihood that there is a certain level of measured features [15]. This method uses Bayes theory, which uses the posterior distribution of classes obtained from the prior distribution of classes and the Conditional Probability Distribution (CPD) of features. Naive Bayes uses the assumption that features are independent of each other in classifying examples, which is not always true, but in practice, this method often produces the same or better results than some more complicated learning methods. [16]. The general form of Bayes' theorem is as follows [17]:

Bayes' Theorem for one proof and one hypothesis is stated in Equation (1).

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)} \quad (1)$$

Description:

- *P(H/E)* is the probability of hypothesis *H* given evidence *E*.
- *P(E/H)* is the probability of evidence *E* if hypothesis *H* is known.
- *P(H)* is the probability of *H* regardless of any evidence.
- *P(E)* is the probability of proof *E*.

Bayes' Theorem for one proof and multiple hypotheses can be formulated in Equation (2):

$$P(Hi|E) = \frac{P(E|Hi).P(Hi)}{\sum_{k=1}^{n} P(E|Hk).P(Hk)} \quad (2)$$

Description:

- *P(Hi/E)* is the probability of hypothesis *Hi* being true given evidence *E*.
- *P(E/Hi)* is the probability of evidence *E* if hypothesis *Hi* is known to be true.
- *P(Hi)* is the probability of hypothesis *Hi* without considering any evidence.

n is the number of hypotheses that occur.

3. Result and Discussion

3.1. Pre-processing Stage

The initial process of data analysis begins with the preprocessing stage, where important steps are taken to ensure the cleanliness and readiness of the data before classification. This stage starts with the removal of data rows that contain missing values. This step is done to maintain data quality and ensure that the classification process can be done properly. Next, the target data type, which was originally an object, was changed to a category. This was done to facilitate the data analysis and classification process. In addition, the remaining data is encrypted using the one-hot encoding technique. This technique aims to convert categorical data into numerical representations that are more easily processed by the classification model.

3.2. Data Splitting and Naïve Bayes Classification

Once the preprocessing phase concludes, the dataset is split into two main segments: the training set and the testing set. This division allocates 70% of the data to the training set and 30% to the testing set. A Naïve Bayes classification model employing the Gaussian NB type is then applied to the training set. Model assessment is conducted using the testing set to evaluate the model's performance in predicting the target classes. The evaluation outcomes reveal a perceived inadequacy in accuracy. This could potentially stem from data imbalance, where the sample sizes across target classes are uneven. The results of the model evaluation are presented in the following Table 1.

Table 1. Classification model evaluation

Accuracy			**0.78**	59
Macro avg	0.85	0.89	**0.83**	59
Weighted avg	0.88	0.78	**0.79**	59

3.3. Handling Imbalance with Under-sampling

To handle the problem of data imbalance, under-sampling techniques were used. Under-sampling aims to balance the number of samples in each target class. After under-sampling, the number of samples in each class is balanced. This step aims to improve the model's performance in classifying minority classes. The Naïve Bayes classification model was then reapplied to the under-sampled data. The model evaluation results show a significant improvement in classification accuracy, indicating that the under-sampling step has successfully improved model performance. The model evaluation results after under-sampling can be seen in Table 2.

Table 2. Model Evaluation After Under-sampling

Accuracy			0.87	39
Macro avg	0.85	0.89	0.83	39
Weighted avg	0.88	0.78	0.79	39

3.4. Testing with Specified Samples

After obtaining a model with satisfactory accuracy, testing is carried out with certain samples to test the predictive ability of the model. An example is testing with cases of on-time assignment submission, sufficient attendance, and good assignment quality. The prediction results of the model show that the predicted value for the sample is category A. This indicates that the classification model has succeeded in predicting the value based on the given attributes.

3.5. Classification Evaluation

Table 3 displays the classification evaluation results of the Naive Bayes model before and after under-sampling.

Table 3. Classification evaluation before under-sampling

	Precision	**Recall**	**F1-Score**	**Support**
A	0.54	1.00	0.70	15
AB	1.00	1.00	1.00	5
B	1.00	0.67	0.80	39

Table 4. Classification Evaluation After Under-sampling

	Precision	**Recall**	**F1-Score**	**Support**
A	0.92	1.00	0.96	11
AB	0.50	1.00	0.67	4
B	1.00	0.79	0.88	24

Table 4 illustrates that following under-sampling, the Naive Bayes model effectively enhances classification performance, particularly in categorizing minority classes. This underscores the significance of addressing data imbalance to enhance the classification model's effectiveness.

4. Conclusion

In this study, employing the Naive Bayes classification method, particularly with the utilization of under-sampling techniques, has demonstrated effectiveness in enhancing the precision of student grade predictions. Implementing strategies to address data imbalance has notably bolstered the classification model's performance, particularly in categorizing minority groups. Consequently, this approach holds promise for enhancing equity and precision in the student evaluation process. The results of this study also open up opportunities for further development in the field of higher education data analysis. Measures for handling data imbalance, such as under-sampling, can be further explored to improve the accuracy of classification models. In addition, more sophisticated data analysis techniques can also be developed to identify factors that influence grading more precisely.

References

[1] C. S. Yusrie, E. Ernawati, D. Suherman, and U. C. Barlian, “Pengembangan Kurikulum dan Proses Pembelajaran Pendidikan Tinggi:,” *Reslaj : Religion Education Social Laa Roiba Journal*, vol. 3, no. 1, pp. 52–69, Feb. 2021, doi: 10.47467/reslaj.v3i1.276.

[2] R. F. Ramadhan and A. A. Widodo, “Penilaian Mahasiswa Berprestasi Menggunakan Metode Simple Additive Weighting Berbasis Decision Support System,” *Jurnal Sistem Informasi dan Informatika (JUSIFOR)*, vol. 1, no. 2, pp. 90–97, Dec. 2022, doi: 10.33379/jusifor.v1i2.1695.

[3] I. Pero, “Keterampilan Literasi Informasi Mahasiswa Fakultas Kedokteran Umum UNBRAH dalam Proses Pembelajaran,” *Shaut Al-Maktabah : Jurnal Perpustakaan, Arsip dan Dokumentasi*, vol. 11, no. 2, pp. 170–184, Jan. 2020, doi: 10.37108/shaut.v11i2.249.

[4] N. M. Firdaus and B. Robandi, “EFEKTIVITAS PENGGUNAAN TEKNOLOGI INTERNET DALAM MENCARI PENGETAHUAN DAN KETERAMPILAN BAGI WARGA BELAJAR PKBM,” *Comm-Edu (Community Education Journal)*, vol. 6, no. 1, p. 6, Jan. 2023, doi: 10.22460/comm-edu.v6i1.7527.

[5] A. Septiana, G. Dwilestari, and A. Bahtiar, “PERBANDINGAN METODE KLASIFIKASI DENGAN MENERAPKAN ADABOOST DALAM ANALISIS SENTIMEN PENGGUNA TWITTER X TERHADAP PENERAPAN KURIKULUM MERDEKA,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp. 323–330, Feb. 2024, doi: 10.36040/jati.v8i1.8453.

[6] R. Muzayanah, A. D. Lestari, B. Prasetiyo, and D. A. A. Pertiwi, “Comparative Study of Imbalanced Data Oversampling Techniques for Peer-to-Peer Landing Loan Prediction,” *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 245–254, 2024, doi: 10.15294/sji.v11i1.50274.

[7] A. Nurdina and A. B. I. Puspita, “Naive Bayes and KNN for Airline Passenger Satisfaction Classification: Comparative Analysis,” *Journal of Information System Exploration and Research*, vol. 1, no. 2, pp. 83–92, 2023, doi: 10.52465/joiser.v1i2.167.

[8] R. Pikriyansah, F. R. Umbara, and P. N. Sabrina, “Klasifikasi Daftar Ulang Calon Mahasiswa Baru Dengan Menggunakan Metode Klasifikasi Naive Bayes,” *Informatics and Digital Expert (INDEX)*, vol. 4, no. 2, pp. 70–74, Jan. 2023, doi: 10.36423/index.v4i2.912.

[9] A. U. Kurnia, A. S. Budi, and P. H. Susilo, “SISTEM PENDUKUNG KEPUTUSAN PENERIMAAN BEASISWA MENGGUNAKAN METODE NAIVE BAYES,” *Joutica*, vol. 5, no. 2, p. 397, Sep. 2020, doi: 10.30736/jti.v5i2.484.

[10] D. Maulina and M. Corry Andhara, “Perbandingan Pre-Processing Opini Netizen Terhadap RUU PKS Menggunakan Algoritma Naive Bayes Classifier,” *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 12, no. 1, Jan. 2023, doi: 10.30591/smartcomp.v12i1.4610.

[11] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[12] Z. Al Faridzi, D. Pramesti, and R. Y. Fa’rifah, “A Comparison of Oversampling and Undersampling Methods in Sentiment Analysis Regarding Indonesia Fuel Price Increase Using Support Vector Machine,” in *2023 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*, IEEE, Aug. 2023, pp. 1–6. doi: 10.1109/ICADEIS58666.2023.10270851.

[13] C. C. Tusell-Rey, O. Camacho-Nieto, C. Yáñez-Márquez, and Y. Villuendas-Rey, “Customized Instance Random Undersampling to Increase Knowledge Management for Multiclass Imbalanced Data Classification,” *Sustainability*, vol. 14, no. 21, p. 14398, Nov. 2022, doi: 10.3390/su142114398.

[14] A. S. Almajid, “Multilayer Perceptron Optimization on Imbalanced Data Using SVM-SMOTE and One-Hot Encoding for Credit Card Default Prediction,” *Journal of Advances in Information Systems and Technology*, vol. 3, no. 2, pp. 67–74, Sep. 2022, doi: 10.15294/jaist.v3i2.57061.

[15] D. D. Lewis, “Naive (Bayes) at forty: The independence assumption in information retrieval,” 1998, pp. 4–15. doi: 10.1007/BFb0026666.

[16] I. Rish, “An empirical study of the naive Bayes classifier,” in *IIJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, pp. 41–46.

[17] Y. Junaedi, B. N. Sari, and A. S. Y. Irawan, “Sistem Pakar Untuk Diagnosis Hama Pada Tanaman Jambu Air Menggunakan Metode Theorema Bayes,” *Jurnal Ilmiah Informatika*, vol. 5, no. 2, pp. 168–178, Dec. 2020, doi: 10.35316/jimi.v5i2.960.