Scholarship Eligibility Prediction Using Gaussian Naive Bayes Based on Academic Potential Test Results

Darma Juang, Zulham Sitorus

Abstract
A systematic and objective method is required to determine eligibility for scholarships. This study aims to develop a scholarship eligibility prediction model based on Academic Potential Test (TPA) results, utilizing the Gaussian Naive Bayes algorithm. The dataset comprises TPA attributes, including quantitative analysis scores, English proficiency scores, the number of correct answers, and total scores. Scholarship eligibility is classified into two categories: eligible and ineligible.The research methodology encompasses data preprocessing, model construction, prediction, and performance evaluation using metrics such as accuracy, precision, recall, and the F1-score. Experimental results indicate that the model achieved an accuracy of 98.21% on a test dataset consisting of 223 entries; for the "eligible" class, the model attained a precision of 96%, a recall of 100%, and an F1-score. Nevertheless, further validation using an independent dataset is necessary to assess the model's generalization capability and its reliability for real-world scholarship selection processes.
Keywords: Academic Potential Test, Scholarship Eligibility, Gaussian Naive Bayes, Classification, Prediction.

Darma Juang[1]
[1 2] Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: darma.juang55@gmail.com[1]

Zulham Sitorus[2]
Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: zulhamsitorus@dosen.pancabudi.ac.id[2]
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

Education is an important part of personal development in a cultured society [1]. Access to higher education is a key factor in improving the quality of human resources in Indonesia. However, certain challenges prevent some people from pursuing an education. Students facing financial difficulties can receive assistance through scholarship programs.

A scholarship is financial aid for students from low-income families, intended to help them continue their education and provide them with necessary resources. Scholarships make it easier for students to pursue their studies and avoid financial barriers. [1]

LP3I Polytechnic in Medan faces challenges in making decisions regarding which students are eligible for scholarships. The method must be capable of processing large amounts of data, recognizing patterns, and making quick and accurate predictive decisions. Data mining techniques can be used to address this need. Data mining is the process of extracting useful information from large amounts of data using mathematical, statistical, and artificial intelligence techniques. [2]

Supervised learning, which relies on labeled data as the basis for training models, is a commonly used machine learning technique for classification problems [4]. Decision Trees, Naive Bayes, Support Vector Machines (SVMs), Random Forests, and K-Nearest Neighbors (KNN) are some of the commonly used algorithms. Based on this progress, commonly used classification methods can be studied. By using the Naïve Bayes algorithm, Academic Potential Test (TPA) scores can be used to determine scholarship eligibility.

Based on previous research on the selection process and the algorithms used, Halizah and Kartikasari (2024) found that the Naive Bayes algorithm achieved an accuracy rate of 98.89% following the calculations [5], Khotimah (2021) used data mining to predict the selection results for scholarship candidates. This study employed the Decision Tree (C4.5) algorithm and compared its results with those of the Naive Bayes algorithm. The findings showed that the C4.5 algorithm had an accuracy of 100%, while the Naive Bayes algorithm had an accuracy of 90.16% [6]. A study conducted by Stiawan and Nugroho in 2023 focused on comparing the performance of decision tree algorithms in categorizing recipients of the Bank Indonesia Scholarship. The study compared the ID3, C4.5, and CART algorithms using data from 398 applicants in 2022. The results showed that CART had the highest accuracy at 72%, followed by C4.5 with an accuracy of 69.33% and ID3 with an accuracy of 68%. [7] , Thoriq et al. (2025) conducted a study titled "Implementation of the Naïve Bayes Algorithm in Predicting the Admission of KIP Scholarship Recipients at Adzkia University." The study used data from 829 scholarship applicants in 2024, with six main attributes. The best results were obtained with an 80:20 data split, with an accuracy of 77.11%, precision of 66.67%, and recall of 5.13%. The low recall value indicates a challenge in identifying the scholarship recipient class, which is significantly smaller in number.[8]. Putri et al. (2025) classified scholarship recipient eligibility using the Naive Bayes algorithm. They used K-Means-based attribute clustering with optimization. This study evaluated the model through several data partitioning scenarios and ten-fold cross-validation using 4,155 student records. The most accurate method was ten-fold cross-validation, which achieved an accuracy of 97.91%.. [3] This study is relevant as a reference on the development of Naïve Bayes models through attribute selection or optimization.

Based on this research gap, the study aims to develop a Naïve Bayes-based classification model to predict scholarship eligibility using Academic Potential Test results. The proposed model uses four attributes: quantitative analysis, English language performance, number of correct answers, and overall test score, with scholarship eligibility represented by

Literature Review

2.1 Scholarships

A scholarship is financial aid provided to individuals, such as students, to support the continuation of their current education. Scholarships may be awarded by government agencies, companies, or foundations. Scholarships may be awarded free of charge or with a service obligation—that is, a requirement to work after completing one's education—the duration of which varies depending on the scholarship-granting institution. In addition, scholarships are also often awarded to Scholarships may be awarded free of charge or with a service obligation—that is, a requirement to work after completing one's education—the duration of which varies depending on the scholarship-granting institution. In addition, scholarships are also often awarded to **[9]**

2.2 Tes Potensi Akademik

The Academic Potential Test (TPA) is useful because it assesses a person's ability to analyze and solve problems, demonstrates their ability to succeed in their chosen subjects, and predicts their educational success. The Academic Potential Test (TPA) helps many people—including teachers, students, employees, and the general public—at various levels of educational institutions.[10]

2.3 Data Mining

Data mining is a data processing technique used to identify patterns, relationships, or useful information from large datasets. This study uses data mining to identify patterns in Academic Potential Test (TPA) scores associated with categories of eligible scholarship recipients. Subsequently, classification is performed based on the probability of each category using the Naive Bayes algorithm. [11]

2.4 Naive Bayes

Naive Bayes is a statistical classification technique based on Bayes' theorem and the assumption that each data point is independent of the others. The algorithm begins by using training data to calculate the probability of each feature being present in relation to the existing classes. Next, for the test data, the probability of each feature is multiplied by the class probability to identify the class with the highest probability of [12]

2.5 Gaussian Naive Bayes

Gaussian Naïve Bayes is an algorithm designed to handle numerical data and assumes that numerical features follow a normal (Gaussian) distribution. The likelihood function for each numerical feature xi [13]

Basic Concept:

Bayes' Theorem is used to calculate class probabilities:

$$P(C|X) = P(X) - \frac{P(X|C)\,P(C)}{P(X)}$$

The value of P(X|C)P(X|C) in Gaussian Naive Bayes can be calculated using the Gaussian probability density function.

$$P(xi|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\begin{matrix}-(\ xi\ -\mu\ c)^2 \\ 2\sigma^2{}_c\end{matrix}\right)$$

Keterangan:

1. xix_i = observed feature values
2. CC = kelas tertentu
3. μC\mu_C = the average feature value in the class CC
4. σC2\sigma_C^2 = Feature variance within a class CC

5. σC\sigma_C = Feature variance within a class C

2.6 Confusion matrix

After performing data mining using a classification model, the confusion matrix is used as a performance metric. Essentially, the confusion matrix provides information on how the algorithm's classification results compare to the actual classification results.[14]

Research Methodology

This study employed a quantitative approach using supervised machine learning methods to predict eligibility for scholarships based on students' Academic Potential Test (TPA) scores. The focus of this study is to develop a classification model using the Gaussian Naive Bayes algorithm. This study was conducted using a quantitative approach and supervised machine learning methods. The objective of the study is to determine whether students are eligible to receive scholarships based on their Academic Potential Test (TPA) results. Developing a classification model using the Gaussian Naive Bayes algorithm is the primary objective of this study. The prediction targets consist of two classes: Eligible and Ineligible. Data collection, data understanding, data preprocessing, exploratory analysis, data splitting, model development, prediction, and evaluation are all steps in the research process. The purpose of these stages is to transform unprocessed TPA data into a classification model that can be used to predict the likelihood of scholarship acceptance.

3.1 Dataset

A dataset is a collection of data stored in digital media (memory), with each data point having a unique value. Different variables contribute to each of these values, and each variable serves a specific function or purpose depending on the type of data collected. A dataset can take the form of a document or a file.[18]. The dataset used in this study consists of students' TPA scores. The available attributes include Quantitative Analysis, English, Correct Answers, and Score. The target variable determines eligibility for a scholarship, which is classified into two categories: Eligible and Ineligible. Table 1 shows the attributes used in this study.

Table.1 attribut

No	Attribute	Description	Role
1	Quantitative Analysis	Student's quantitative analysis test result	Predictor
2	English	Student's English language test result	Predictor
3	Correct Answer	Number of correctly answered questions	Predictor
4	Value	Overall TPA score	Predictor
5	Feasibility	Scholarship eligibility status	Target

Since the predictor variables are numerical attributes, the Gaussian Naive Bayes classifier is appropriate. According to the assumptions of the Gaussian Naive Bayes classifier, numerical features follow a Gaussian distribution.

3.2 Data Preprocessing

To ensure that the dataset is suitable for classification, preprocessing is performed before model development. This preprocessing includes examining the data to determine the number, attributes, and data types of the observed variables; checking for missing values to identify incomplete information; checking for duplicates to find and remove duplicate records if necessary; checking data consistency to ensure that numeric attributes contain valid values; and verifying the target labels to ensure that the scholarship eligibility variables consist only of acceptable variables.

3.3 Explorative Datenanalyse

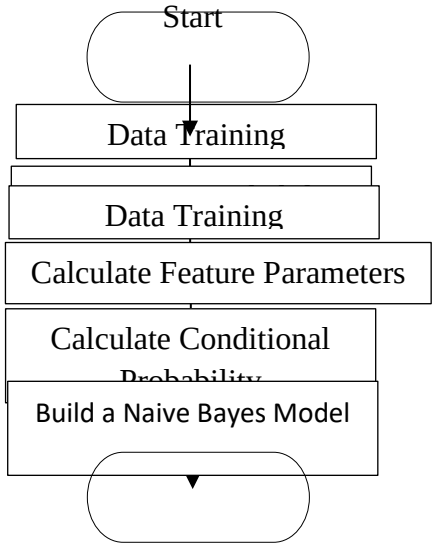
Exploratory data analysis was conducted to understand the characteristics of the dataset prior to classification. This analysis included the distribution of scholarship eligibility classes and descriptive statistics for the numerical attributes. The distribution of the target classes was examined to determine whether the dataset was balanced between the "Eligible" and "Ineligible" categories.

3.4 Dataset Split

The dataset is divided into training and testing subsets after preprocessing. The test dataset is used to evaluate the model's predictive performance, while the training dataset is used to build a Naive Bayes classification model. To evaluate the model, observations not used during the training process are utilized by splitting the data into training and test sets. An 80:20 ratio can be used to split the observations into 20% for the test data and 80% for the training data.

3.5 Model Training

The training data is used to create a Naive Bayes model during the training phase..



Description:

The model examines the relationship between Quantitative Analysis + English + Correct Answers + Score and the outcome of "Pass" or "Fail."

3.10 Model Testing

Once the model is built, the test data is used to evaluate the model's ability to classify data that was not used during the training process. For each data point in the test set, the model generates a prediction of whether it is acceptable or unacceptable. To obtain an evaluation score, the predictions are compared with the actual labels.

Results of the Discussion

4.1 Data Preprocessing

For the purpose of this study, a dataset from an Academic Potential Test (TPA)—comprising 1,117 rows and fifteen attributes—was utilized. The research began with an examination of the data structure, missing values, and duplicates. This assessment revealed that certain entries lacked values for the TPA score attribute; consequently, these were excluded from the classification process. Data preprocessing was then performed, which involved formatting numerical attributes appropriately and selecting those relevant to the prediction task. The input variables consisted of scores in English, analytical reasoning, and quantitative reasoning. Personal identification details—such as names, email addresses, WhatsApp numbers, educational background, and other administrative attributes—were omitted from the model, as they bear no direct relation to the academic quality metrics used for prediction. Following data cleaning, 498 records with complete TPA scores were retained for the modeling phase. Since the original dataset lacked a target attribute for scholarship eligibility, classes were

defined based on the TPA scores: participants with a score below 60 were classified as "ineligible," while those with a score of 60 or higher were classified as "eligible." The classification results yielded 107 participants in the "eligible" category and 112 in the "ineligible" category. An analysis of the distribution confirmed that the dataset does not suffer from extreme class imbalance, as the proportions of the two classes are relatively balanced

4.2 Gaussian Naive Bayes Classification

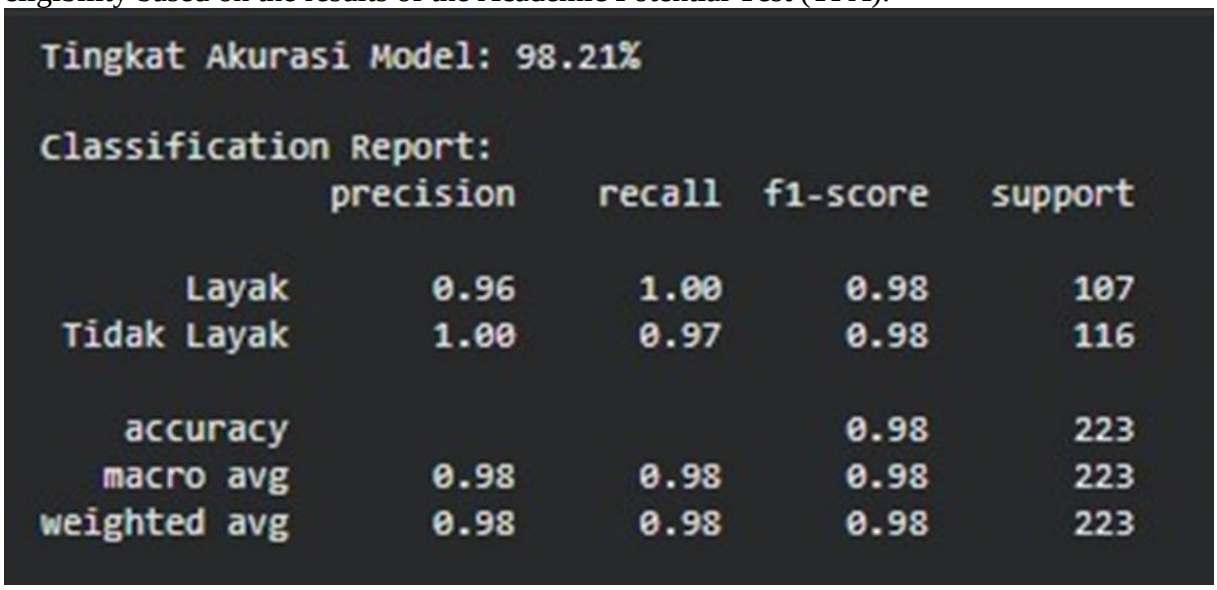
The classification model employed is Gaussian Naive Bayes. This algorithm was selected because the input variables consist of numerical data—specifically, scores for Analytical, Quantitative, and English language assessments. The dataset was partitioned into training and testing sets using an 80:20 ratio via stratified sampling, resulting in 890 samples for training and 223 samples for testing. The Gaussian Naive Bayes model was then trained on the training set and subsequently used to predict the eligibility class for the data in the testing set.

4.3 Prediction Results

The prediction results from 223 test data points show that the model can accurately classify the majority of participants. Based on the confusion matrix, we found 107 false negatives, 112 true positives, 4 false positives, and 0 false negatives. The results show that the model was able to successfully identify 107 participants who truly belonged in the “Eligible” category and classify them as such. For now, the 112 participants deemed ineligible were correctly classified.

4.4 Model Performance

The performance of the Gaussian Naive Bayes model was evaluated using accuracy, precision, recall, and F1-score. The test results are presented in the following table. The model achieved an accuracy of 98.00%, indicating that 107 out of 223 test data points were correctly classified. A precision value of 96.00% suggests that the majority of participants predicted to be in the "Eligible" category were indeed eligible. A recall value of 1%—which is quite high—demonstrates that the model is capable of identifying most of the participants who truly belong to the "Eligible" category. In the context of a scholarship decision-support system, recall serves as a crucial indicator, as the ability to identify candidates who meet the established criteria is of great importance. Furthermore, an F1-score of 98.00% reflects a strong balance between precision and recall. Consequently, based on the test results from the research dataset, the Gaussian Naive Bayes model demonstrates good performance in classifying scholarship eligibility based on the results of the Academic Potential Test (TPA).

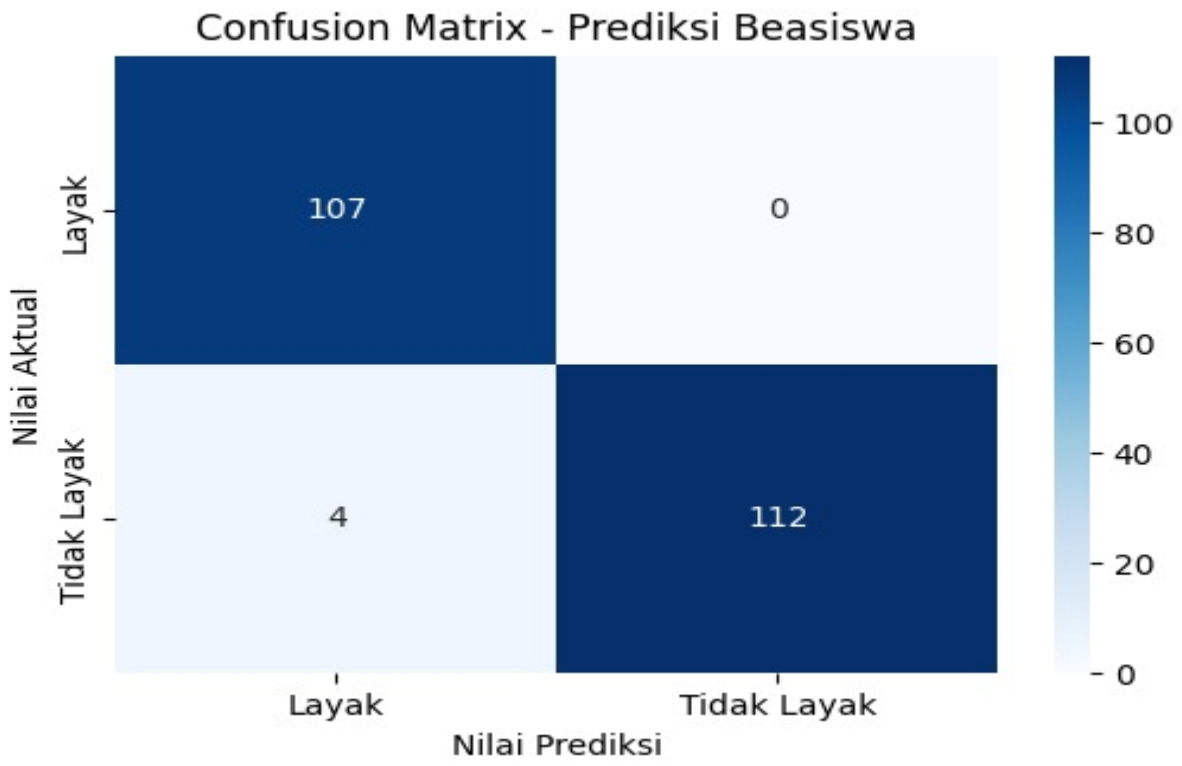


```
Tingkat Akurasi Model: 98.21%

Classification Report:
              precision    recall  f1-score   support

       Layak       0.96      1.00      0.98       107
 Tidak Layak       1.00      0.97      0.98       116

    accuracy                           0.98       223
   macro avg       0.98      0.98      0.98       223
weighted avg       0.98      0.98      0.98       223
```



Results

The research results indicate that a classification model for scholarship eligibility can be developed using data from the Academic Potential Test (APT) components. The English language, analytical, and quantitative sections provide numerical information that can be processed by the Gaussian Naive Bayes algorithm to determine the probability of an applicant falling into an eligibility category. The model's performance, which achieved an accuracy of 98.00%, demonstrates that Gaussian Naive Bayes possesses strong classification capabilities for this research dataset. Furthermore, a high recall rate of 96.30% indicates that the model is effectively able to identify applicants who fall into the "Eligible" category. However, these findings must be interpreted in the context of the labeling mechanism used. An APT score threshold of 60 was applied to categorize applicants as either "eligible" or "ineligible." Consequently, the model essentially examines how the scores from the APT components correlate with the categories defined by this threshold. Future research could enhance the model's validity by incorporating actual scholarship decision data—meaning that the "Eligible/Ineligible" labels would be derived from the real-world outcomes of scholarship selection processes. Additionally, academic and non-academic variables—such as Grade Point Average (GPA), economic status, student achievements, and other criteria established by the institution—could be integrated into the model's development. By including these variables, the model could predict eligibility based on APT performance while reflecting a more comprehensive set of scholarship selection criteria.

Research Limitations

When interpreting the findings of this research, several limitations should be considered

1. Variable limitations: The research data is limited to TPA (Academic Potential Test) scores, including quantitative analysis, English language proficiency, the number of correct answers, and total scores. Additional factors—such as family socioeconomic

status, as well as academic and non-academic achievements—have not been accounted for

2. Methodological limitations: This study employs the Gaussian Naive Bayes algorithm, which relies on a Gaussian distribution for continuous features and assumes conditional independence among predictor variables. The relationships between variables in the dataset may not be fully captured by these assumptions.
3. Dataset limitations: The quality, size, and representativeness of the data, as well as the accuracy of the available scholarship eligibility labels, influence the model's performance.
4. Generalization limitations: The research results cannot necessarily be applied directly to other institutions that utilize different scholarship selection criteria.
5. Future research could potentially compare the Gaussian Naive Bayes classification algorithm with other algorithms, expand the dataset, incorporate relevant additional variables, and employ various validation methods.

Conclusion

This study develops a prediction model for scholarship eligibility based on Academic Potential Test (TPA) results, utilizing the Gaussian Naive Bayes algorithm. Experimental results demonstrate an accuracy rate of 98.0% on a test dataset consisting of 223 samples. The classification report shows precision, recall, and F1-score values of approximately 0.98 for both eligibility classes. For the class of eligible candidates, the model achieved a precision of 0.96, a recall of 1.00, and an F1-score of 0.98. For the class of ineligible candidates, the model achieved a precision of 1.00, a recall of 0.97, and an F1-score of 0.98. These results indicate that the model exhibits strong classification performance on the evaluated dataset, particularly in identifying eligible candidates.

References

[1] K. Words, "許剣 1 ・山上 路生 2 ・岡本 隆明 3 ・角 哲也 4," vol. 4, no. 16, pp. 1–7, 2024, [Online]. Available: https://www.bajangjournal.com/index.php/JIRK/article/view/8744

[2] P. T. Informasi, F. Teknik, U. N. Surabaya, P. T. Informasi, F. Teknik, and U. N. Surabaya, "Dodik Arwin Dermawan," vol. 07, pp. 58–66, 2022.

[3] Putri Azhiah, Jasmir, and Purnama Benni, "Klasifikasi Kelayakan Penerima Beasiswa Menggunakan Naive Bayes dengan Optimasi Atribut Berbasis K-Means," *Inst. Ris. dan Publ. Indones.*, vol. 5, no. October, pp. 1450–1462, 2025.

[4] M. R. Rizaldi *et al.*, "Penerapan Machine Learning Berbasis Supervised Learning Menggunakan Metode Naive Bayes Untuk Klasifikasi Prestasi," vol. 6, no. 1, pp. 101–108, 2026.

[5] C. Fitri Nuril Halizah and M. Kartikasari, "Implementasi Algoritma Naïve Bayes Dalam Penentuan Pemberian Beasiswa Kip Kuliah (Studi Kasus Stiki Malang)," *J. Simantec*, vol. 12, no. 2, pp. 1–6, 2024, doi: 10.21107/simantec.v12i2.21889.

[6] K. Khotimah, "Teknik Data Mining menggunakan Algoritma Decision Tree (C4.5) untuk Prediksi Seleksi Beasiswa Jalur KIP pada Universitas Muhammadiyah Kotabumi," *J. SIMADA (Sistem Inf. dan Manaj. Basis Data)*, vol. 4, no. 2, pp. 145–152, 2022, doi: 10.30873/simada.v4i2.3064.

[7] Dimas Bayu Stiawan and Yusuf Sulistyo Nugroho, "Perbandingan Performa Algoritma Decision Tree," *Indones. J. Comput. Sci. Attrib.*, vol. 12, no. 4, p. 2108, 2023.

[8] M. Penerima and B. Kip, "Implementasi Algoritma Naïve Bayes dalam Prediksi Penerimaan," vol. 15, no. 1, pp. 108–114, 2025.

[9] M. Andry Juniawan and A. Sudrajat, "Prediksi Seleksi Penerima Beasiswa Kip Kuliah Menggunakan Algoritma Naïve Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 6, pp. 9965–9972, 2025, doi: 10.36040/jati.v9i6.15839.

[10] Sinta, Sandie, and U. D. Susiaty, "Analisis Penalaran Spasial Mahasiswa dalam Menyelesaikan Tes Potensi Akademik," *J. Prodi Pendidik. Mat.*, vol. 4, no. 2, pp. 432–441, 2022.

[11] A. Nasrullah, "Data Mining Algoritma Naive Bayes Untuk Memprediksi Jumlah Siswa Baru," *EJECTS J. Comput. Technol. ...*, vol. 2, no. 2, pp. 62–67, 2023, [Online]. Available: https://www.jurnal.unda.ac.id/index.php/ejects/article/view/378%0Ahttps://www.jurnal.unda.ac.id/index.php/ejects/article/download/378/296

[12] A. Pebdika, R. Herdiana, and D. Solihudin, "Klasifikasi Menggunakan Metode Naive Bayes Untuk Menentukan Calon Penerima Pip," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 452–458, 2023, doi: 10.36040/jati.v7i1.6303.

[13] P. Binuryan, N. M. Satianingsih, and C. E. Murwaningtyas, "Analisis Komparasi Metode Gaussian Naive Bayes dan Bernoulli Naive Bayes dalam Klasifikasi Pemilihan Program Peminatan Siswa SMA," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 11, no. 2, pp. 314–325, 2025, [Online]. Available: https://jurnal.untan.ac.id/index.php/jepin/article/view/89583

[14] P. Romadloni, B. Kusuma Adhi, and W. Baihaqi Maulana, "Komparasi Metode Pembelajaran Mesin Untuk Implementasi," vol. 6, no. 2, pp. 622–628, 2022.