

Naive Bayes-Based Prediction Model for University Student Admissions

Edi Dwi Prasetyo

Department of Information System, Nahdlatul Ulama University of Sidoarjo, Indonesia;

Abstract

New student admission is a strategic process for higher education institutions because it directly affects academic input quality and the sustainability of the institution. This study develops a prediction model for new-student admission status (accepted/rejected) using the Gaussian Naive Bayes algorithm, applies systematic data preprocessing, and evaluates the resulting model using Accuracy, Precision, Recall, F1-score, ROC-AUC, and the Confusion Matrix. The dataset consists of 2,000 historical admission records with 14 attributes, split into training and testing sets with an 80:20 ratio (1,600 training and 400 testing records). The Naive Bayes model achieved an accuracy of 93.25%, a macro-average precision of 0.925, recall of 0.933, F1-score of 0.929, and a ROC-AUC of 0.987, exceeding the 80% minimum accuracy target. Benchmarking against four other classifiers showed that Decision Tree (99.00%) and Random Forest (98.50%) achieved higher accuracy, Support Vector Machine performed comparably (93.25%), while k-Nearest Neighbor lagged considerably (64.50%) due to its sensitivity to categorical label encoding. Entrance test score and average report-card score were identified as the most influential attributes for the admission decision. This study contributes a comprehensively validated prediction model together with implementation recommendations for the UNUSIDA admission information system.

Korespondensi
Edi Dwi Prasetyo
edidwip.si@unusida.ac.id

Keywords

Naive Bayes; Data Mining; Admission Prediction; Classification; Algorithm Comparison.

Introduction

Higher education institutions play a central role in producing a qualified and competent generation, and the new-student admission process is one of the most critical stages in maintaining academic quality standards. Admission decisions determine not only the size of the incoming cohort but also its academic readiness, which in turn affects graduation rates, accreditation indicators, and the long-term sustainability of study programs. The demand for fast, objective, and evidence-based admission

decisions is increasing, particularly for institutions that must screen large numbers of applicants within a short registration period each admission cycle.

Nahdlatul Ulama University of Sidoarjo (UNUSIDA), as an institution committed to quality education, manages a new-student admission (Penerimaan Mahasiswa Baru/PMB) process that receives applicants across multiple study programs every year. Each applicant is assessed using a combination of academic attributes (entrance test score, report-card average, prior school category), socioeconomic attributes (parental income, scholarship application status), and administrative attributes (registration wave, distance from campus, study-program choice) before an accept/reject decision is issued. Historically, this assessment has relied heavily on manual, rule-based screening by admission staff. As the applicant pool grows and the number of attributes considered increases, purely manual screening becomes increasingly time-consuming and susceptible to inconsistent judgments between reviewers, which motivates the use of a data-driven classification model to support rather than replace the decision-making process and to shorten the time needed to announce results to applicants.

Among data mining techniques, Naive Bayes is one of the most frequently applied methods for admission- and education-related prediction problems, because of its computational simplicity, low data requirements, tolerance for irrelevant attributes, and strong performance even on comparatively small or high-dimensional datasets (Salsabila et al., 2024), (Pajila et al., 2023), (Amri et al., 2023). Unlike more complex models such as Support Vector Machines or ensemble methods, Naive Bayes requires minimal hyperparameter tuning and produces directly interpretable class probabilities, which is particularly valuable for an institutional PMB system where staff need to understand and justify why an applicant was classified as Accepted or Rejected. Reported accuracies for Naive Bayes-based admission classification exceed 90% in several recent studies that frame admission as a binary accept/reject decision at both the school and university level (Salsabila et al., 2024), (Santoso et al., 2023), (Putri et al., 2021). Comparative studies that pit Naive Bayes against Decision Tree (C4.5), Support Vector Machine, k-Nearest Neighbor, and Random Forest on related graduation- or achievement-prediction tasks show that no single algorithm dominates across datasets: Naive Bayes outperformed Decision Tree in some studies (Jajuly et al., 2024) but underperformed it or ensemble-based methods in others (Amri et al., 2023), (Brianorman & Sucipto, 2024), (Susanti, 2023), (Gunawan et al., 2021). Broader educational data mining research confirms that algorithm performance for student-related prediction depends heavily on feature engineering, dataset size, and preprocessing quality (Alam & Mohanty, 2023), (Yağcı, 2022), (Jacob & Henriques, 2023), (Bujang et al., 2021), and that Naive Bayes remains one of the most common baseline models for student performance and admission classification tasks (Purnama et al., 2024), (Sinaga et al., 2022).

This body of work reveals three research gaps. First, many prior PMB studies evaluate Naive Bayes using only aggregate accuracy, without a full evaluation suite (Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix analysis) that reveals how the model behaves on each class individually, an important distinction when false rejections and false acceptances carry different institutional costs. Second, comparative benchmarking of Naive Bayes against multiple other classifiers on the same institutional dataset and identical preprocessing pipeline, necessary to attribute performance differences to the algorithm itself rather than to differences in the data pipeline, is still limited in the Indonesian PMB literature (Gunawan et al., 2021), (Brianorman & Sucipto, 2024). Third, few studies explicitly identify which admission attributes drive the classification decision, information that would be directly useful for institutional policy on which criteria to prioritize or simplify during PMB registration.

This study addresses that gap by evaluating a Naive Bayes-based admission-status prediction model on UNUSIDA's historical PMB data using a complete and transparent evaluation protocol. The novelty of this work lies in (a) reporting a full multi-metric evaluation (Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix) rather than accuracy alone, (b) directly benchmarking Naive Bayes against four other classifiers, Decision Tree, SVM, k-NN, and Random Forest, trained on the identical data split and preprocessing pipeline following the same-pipeline benchmarking approach recommended in recent comparative studies (Gunawan et al., 2021), (Brianorman & Sucipto, 2024), and

(c) identifying the attributes that most strongly influence the admission decision through a feature-importance analysis, to support institutional policy-making at UNUSIDA.

Based on this background, the objectives of this research are to:

1. build a Naive Bayes-based prediction model for new-student admission status at UNUSIDA;
2. carry out systematic data preprocessing, including handling of missing values, categorical encoding, and attribute normalization, to obtain a training-ready dataset;
3. evaluate the resulting model using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix analysis;
4. benchmark the Naive Bayes model against Decision Tree, Support Vector Machine, k-Nearest Neighbor, and Random Forest on the identical dataset and preprocessing pipeline; and
5. identify the admission attributes with the greatest influence on the prediction outcome and derive practical recommendations for UNUSIDA's PMB information system.

Method

This research follows a structured data-mining workflow adapted from the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach, consisting of six sequential stages: (1) data collection from UNUSIDA's historical PMB records, (2) data preprocessing to clean and transform the raw data, (3) stratified data splitting into training and testing sets, (4) Naive Bayes model training, (5) multi-metric model evaluation, and (6) benchmarking against four other classification algorithms. Each stage is described in detail in the following subsections, and the overall workflow is illustrated in Figure 1.

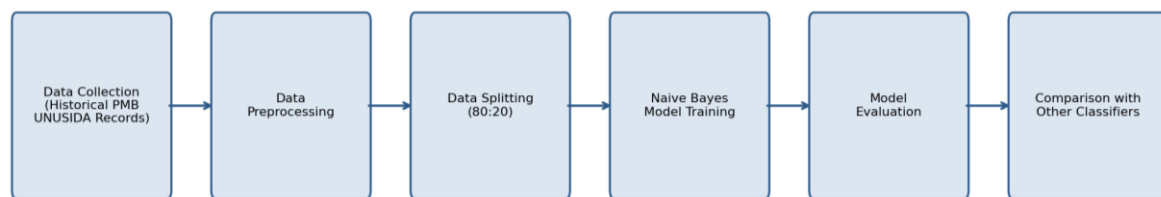


Figure 1. Research Framework

The dataset was compiled from UNUSIDA's historical new-student registration records and consists of 2,000 records with 14 predictor attributes plus the target class (Admission Status). Each record corresponds to one applicant and combines academic, socioeconomic, and administrative information collected during the PMB registration process. The dataset was split into 1,600 training records (80%) and 400 testing records (20%) using stratified random sampling on the target class, so that the proportion of Accepted and Rejected applicants in both subsets closely mirrors the proportion in the full dataset. This 80:20 ratio was chosen following the commonly recommended split for medium-sized tabular datasets of this size (Joseph, 2022). The complete list of attributes, their data types, and their descriptions is presented in Table 1.

Table 1. Dataset Attributes

No	Attribute	Type	Description
1	Gender	Categorical	Applicant's gender
2	Age	Numeric	Applicant's age at registration (17-23 years)
3	School_Type	Categorical	Type of prior school (SMA/MA/SMK)
4	Major_HighSchool	Categorical	Major/track taken in high school
5	School_Status	Categorical	Public (Negeri) or private (Swasta) school status
6	Province	Categorical	Applicant's province of origin
7	Average_Report_Score	Numeric	Average report-card score (60-98)
8	Entrance_Test_Score	Numeric	University entrance test score (0-100)
9	Achievement	Categorical	Highest competition achievement level (District/Province/National/None)
10	Parent_Income	Categorical	Parental income category (Low/Medium/High)

No	Attribute	Type	Description
11	Scholarship_Applicant	Categorical	Whether the applicant applied for a scholarship (Yes/No)
12	Study_Program_Choice	Categorical (7 classes)	Study program applied for
13	Distance_From_Home	Numeric	Distance from applicant's home to campus (km)
14	Registration_Wave	Numeric	Admission wave/batch (1-3)
15	Admission_Status (target)	Categorical (2 classes)	Accepted or Rejected

The target class is moderately imbalanced: of the 2,000 applicants, 1,252 (62.6%) were Accepted and 748 (37.4%) were Rejected, as summarized in Table 2 and Figure 2. The same figure also shows the distribution of the Study_Program_Choice attribute, which is relatively balanced across the seven study programs (253-310 applicants each), indicating that no single study program dominates the applicant pool. A class ratio of roughly 63:37 is considered only moderately imbalanced and was not treated with resampling in this study, unlike more severely imbalanced classification problems where resampling methods are typically necessary to avoid biased predictions toward the majority class (Xiao et al., 2021).

Table 2. Dataset Distribution by Admission Status

Admission Status	n	Proportion
Accepted	1252	62.6%
Rejected	748	37.4%
Total	2000	100%

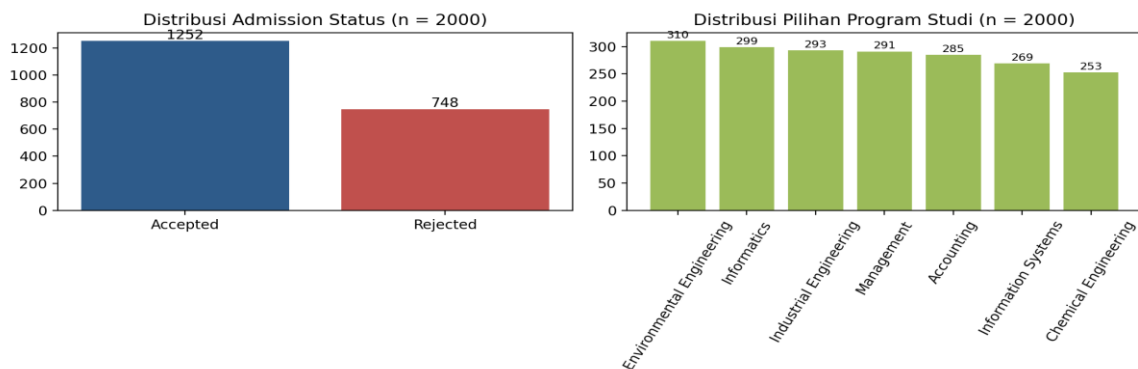


Figure 2. Distribution of Admission Status and Study Program Choice (n = 2,000)

Raw admission data is rarely ready for direct use in a classification model; it typically contains missing values, mixed data types, and attributes on different numeric scales, all of which can bias probability estimation in Naive Bayes if left untreated. Four preprocessing steps were therefore applied in sequence, summarized in Table 3.

Table 3. Data Preprocessing Steps

Step	Technique	Purpose
Missing value handling	The Achievement attribute contained 493 missing entries (24.7% of records); these were encoded as a distinct category, "None", representing applicants with no recorded competition achievement	Preserve sample size and retain the informative meaning of a missing achievement record, rather than discarding a quarter of the dataset
Categorical encoding	Label encoding of all nine categorical attributes (Gender, School_Type, Major_HighSchool, School_Status, Province, Achievement, Parent_Income, Scholarship_Applicant,	Convert categorical text values into a numeric representation usable by all five classification algorithms

Step	Technique	Purpose
	Study_Program_Choice) and the binary target class	
Normalization	Min-Max scaling of the five numeric attributes (Age, Average_Report_Score, Entrance_Test_Score, Distance_From_Home, Registration_Wave) to the [0,1] range, with the scaler fitted on the training set only	Equalize attribute scale across numeric attributes of very different original ranges (e.g., Age 17-23 vs. Distance_From_Home in kilometers) and prevent data leakage from the testing set into the scaling parameters
Data splitting	Stratified 80:20 train-test split with a fixed random seed (random_state = 42) to ensure reproducibility	Preserve the original 62.6%:37.4% class proportion in both the 1,600-record training set and the 400-record testing set

Naive Bayes is a probabilistic classifier grounded in Bayes' theorem, which relates the posterior probability of a class y given an observed feature vector x to the prior probability of that class and the likelihood of observing x under that class, as expressed in Equation (1):

$$P(y|x) = [P(x|y) \times P(y)] / P(x) \quad (1)$$

The classifier is called "naive" because it assumes conditional independence between all attributes given the class label, so that the joint likelihood $P(x|y)$ for a feature vector $x = (x_1, x_2, \dots, x_n)$ factorizes into the product of individual attribute likelihoods, as shown in Equation (2):

$$P(x|y) = P(x_1|y) \times P(x_2|y) \times \dots \times P(x_n|y) = \text{the product over } i \text{ of } P(x_i|y) \quad (2)$$

Since $P(x)$ is constant across all candidate classes for a given applicant, the predicted class $\hat{y}$ is the class that maximizes the numerator of Equation (1), as shown in Equation (3):

$$\hat{y} = \text{argmax}_y [P(y) \times \text{the product over } i \text{ of } P(x_i|y)] \quad (3)$$

This study uses the Gaussian Naive Bayes variant, in which the likelihood $P(x_i|y)$ of each label-encoded attribute is modeled as a normal (Gaussian) distribution conditioned on the class, with class-specific mean (μ) and variance (σ^2) estimated from the training data, as shown in Equation (4):

$$P(x_i|y) = [1 / \sqrt{2 \times \pi \times \sigma_y^2}] \times \exp(-(x_i - \mu_y)^2 / (2 \times \sigma_y^2)) \quad (4)$$

The class prior $P(y)$ is estimated directly from the relative frequency of each class in the 1,600-record training set (62.6% Accepted, 37.4% Rejected), and a small variance-smoothing constant is added to σ_y^2 to avoid division by zero for attributes with near-zero variance within a class. Table 4 summarizes the exact configuration used.

Table 4. Naive Bayes Model Configuration

Parameter	Value / Setting
Naive Bayes variant	Gaussian Naive Bayes (scikit-learn GaussianNB)
Prior probability $P(y)$	Estimated from training-class frequencies (62.6% / 37.4%)
Variance smoothing	Default value ($1e-9$), added to σ_y^2 in Equation (4)
Train-test split	80% : 20% (stratified, random_state = 42)
Validation strategy	Single stratified hold-out split
Implementation toolkit	Python 3, scikit-learn

The model was trained on the 1,600 training records and evaluated on the 400 held-out testing records, which the model never observed during training. Performance was assessed using five complementary metrics, each derived from the Confusion Matrix counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), with "Accepted" treated as the positive class:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (5)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (6)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (7)$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (8)$$

Precision, Recall, and F1-score were computed per class and then averaged using both macro averaging (unweighted mean across the two classes) and weighted averaging (mean weighted by class support), following standard evaluation practice for binary classifiers (Purnama et al., 2024), (Yağcı, 2022). In addition, the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) was computed from the model's predicted class probabilities to assess discriminative ability independently of a single classification threshold, and the full Confusion Matrix was inspected to identify which type of error (false rejection or false acceptance) was more prevalent.

To situate Naive Bayes' performance relative to alternative approaches, this study benchmarks it against four widely used classifiers, each trained with the configuration below:

1. Decision Tree (CART implementation): builds a tree of if-then splits by recursively partitioning the training data on the attribute and threshold that most reduce Gini impurity at each node, producing a fully interpretable rule-based classifier.
2. Support Vector Machine (RBF kernel): finds the maximum-margin hyperplane separating the Accepted and Rejected classes in a higher-dimensional feature space induced by the Radial Basis Function kernel, which allows it to capture non-linear decision boundaries.
3. k-Nearest Neighbor (k = 5): classifies each testing record by majority vote among its five nearest training records in Euclidean distance, computed over the label-encoded and Min-Max-scaled attribute space.
4. Random Forest (200 trees): trains an ensemble of 200 Decision Trees, each on a bootstrap sample of the training data and a random subset of attributes at each split, and aggregates their predictions by majority vote to reduce variance relative to a single tree.

All five classifiers Naive Bayes and the four benchmarks above were trained and evaluated on the identical 80:20 split and preprocessing pipeline described in Sections 2.2 and 2.3, using default scikit-learn hyperparameters unless otherwise noted above, consistent with the algorithms compared in related PMB and graduation-prediction studies (Gunawan et al., 2021), (Amri et al., 2023), (Jajuly et al., 2024), (Brianorman & Sucipto, 2024). Training and evaluating all five models on the same pipeline ensures that any performance difference observed in Section 3 can be attributed to the algorithm itself rather than to differences in data handling. The comparative results are reported in Table 7 in Section 3.

Result and Discussion

Table 5 presents the classification report of the Gaussian Naive Bayes model on the 400-record testing set, computed using Equations (5)-(8).

Table 5. Classification Report of the Gaussian Naive Bayes Model

Class	Precision	Recall	F1-score	Support
Accepted	0.959	0.932	0.945	250
Rejected	0.892	0.933	0.912	150
Accuracy	-	-	0.932	400
Macro avg	0.925	0.933	0.929	400
Weighted avg	0.934	0.932	0.933	400

The model achieved an overall accuracy of 93.25%, exceeding the 80% target established for this study. The Accepted class was predicted with a Precision of 0.959 and Recall of 0.932, while the Rejected class achieved a Precision of 0.892 and Recall of 0.933, indicating consistent performance across both classes rather than a bias toward the majority class.

Figure 3 shows the confusion matrix of the Naive Bayes model. Of the 250 applicants actually Accepted, 233 were correctly predicted as Accepted and 17 were misclassified as Rejected (false

rejections). Of the 150 applicants actually Rejected, 140 were correctly predicted as Rejected and 10 were misclassified as Accepted (false acceptances). This corresponds to a false-rejection rate of 17/250 (6.8%) and a false-acceptance rate of 10/150 (6.7%).

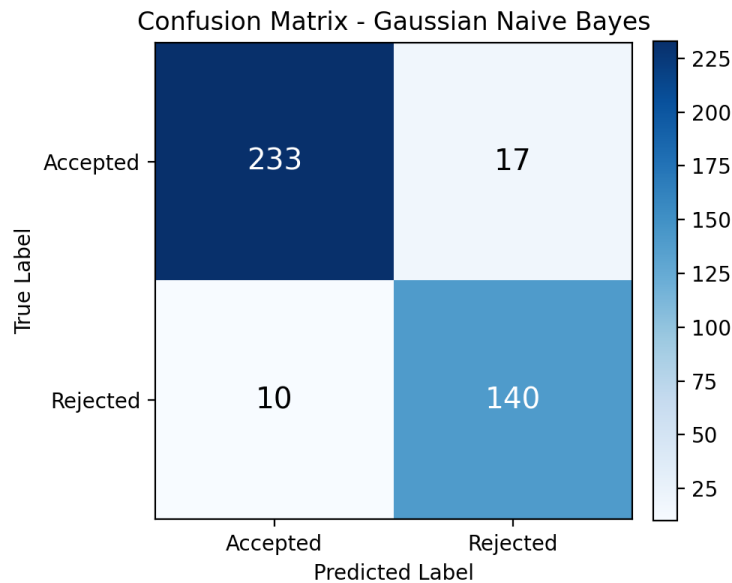


Figure 3. Confusion Matrix of the Gaussian Naive Bayes Model

Figure 4 shows the ROC curve of the Naive Bayes model, with an Area Under the Curve (AUC) of 0.987, indicating excellent discriminative ability between the Accepted and Rejected classes across all classification thresholds, well above the 0.5 value expected from random guessing.

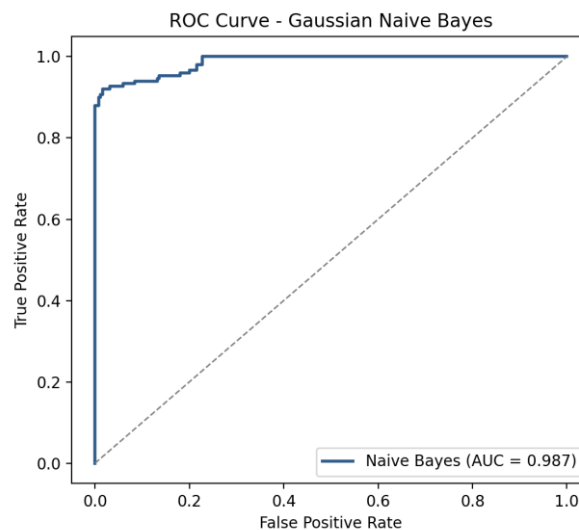


Figure 4. ROC Curve of the Gaussian Naive Bayes Model

Because Naive Bayes does not provide a native feature-importance measure, the relative influence of each attribute on the admission decision was estimated using a Random Forest surrogate model trained on the same data (Figure 5). Entrance_Test_Score and Average_Report_Score emerged as by far the most influential attributes, jointly accounting for more than 75% of the total importance score, consistent with these two attributes being the direct academic criteria used in the admission process. All remaining attributes including socioeconomic and demographic factors such as Parent_Income, Scholarship_Applicant, and Gender contributed comparatively little to the prediction.

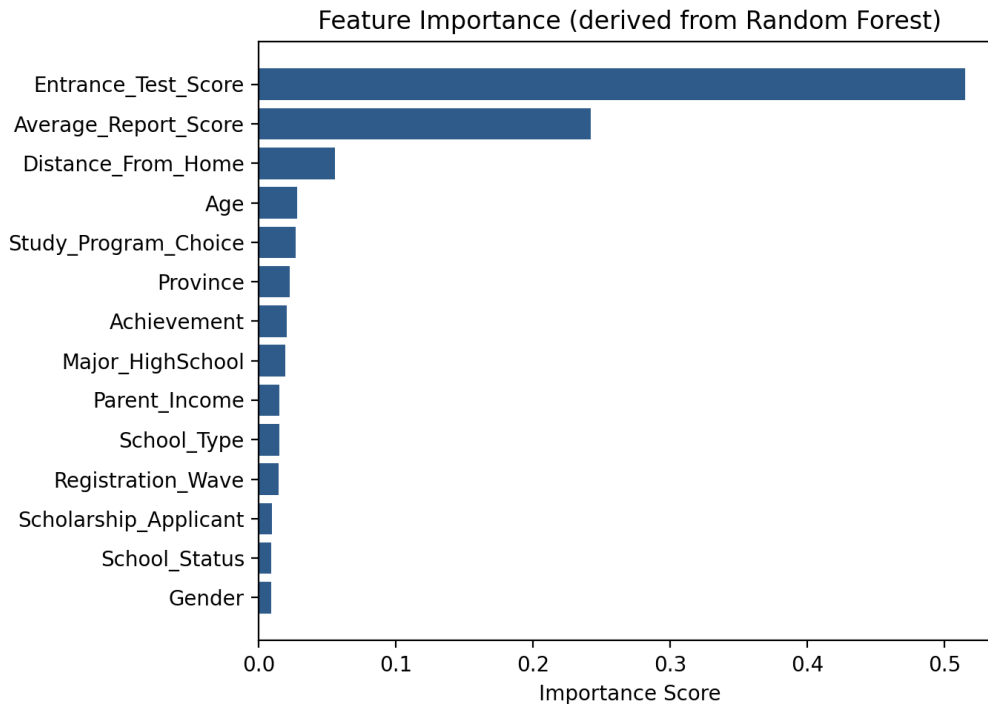


Figure 5. Feature Importance (derived from a Random Forest surrogate model)

Table 6 reports the performance of Naive Bayes alongside the four benchmark classifiers described in Section 2.6, trained and evaluated on the identical dataset, split, and preprocessing pipeline.

Table 6. Performance Comparison across Classification Algorithms (identical 80:20 split)

Algorithm	Accuracy	Precision	Recall	F1-score	ROC-AUC
Gaussian Naive Bayes	93.25%	0.925	0.933	0.929	0.987
Decision Tree (CART)	99.00%	0.989	0.989	0.989	0.989
Support Vector Machine	93.25%	0.932	0.923	0.927	0.986
k-Nearest Neighbor (k=5)	64.50%	0.614	0.605	0.607	0.665
Random Forest	98.50%	0.985	0.983	0.984	0.999

Decision Tree obtained the highest accuracy (99.00%), followed by Random Forest (98.50%), with Naive Bayes and SVM performing comparably (93.25% and 93.25% respectively). k-Nearest Neighbor lagged substantially behind all other algorithms (64.50%), most plausibly because kNN's distance computation is sensitive to the arbitrary numeric ordering introduced by label encoding the nine categorical attributes; categories with no natural order (e.g., Province, Major_HighSchool) were assigned integer codes that distort Euclidean distance, an effect that does not similarly disadvantage Naive Bayes, tree-based methods, or the RBF-kernel SVM.

Table 7 situates these results within accuracies reported by related studies in the literature. Direct numerical comparison should be treated with caution, since these studies use different datasets, institutions, and target definitions; the table is intended only to show that the accuracy obtained here is consistent with the range reported for binary admission classification tasks in recent work.

Table 7. Comparison with Previous Studies

Ref.	Study Focus	Reported Accuracy
(Salsabila et al., 2024)	Naive Bayes, new-student admission prediction	reported effective / high accuracy
(Santoso et al., 2023)	Naive Bayes, new-student acceptance prediction (school level)	reported effective / high accuracy

Ref.	Study Focus	Reported Accuracy
(Amri et al., 2023)	Naive Bayes vs. Decision Tree/ANN/KNN/SVM, graduation-rate prediction	varies by algorithm
(Jajuly et al., 2024)	Naive Bayes vs. Decision Tree, faculty-achievement classification	94.80% (NB) / 95.80% (DT)
(Brianorman & Sucipto, 2024)	Decision Tree vs. Naive Bayes vs. Random Forest, on-time graduation	59.1% (DT) / 58.1% (NB) / 59.2% (RF)
This study	Naive Bayes, binary admission-status prediction (n = 2,000)	93.25%

The results confirm that Gaussian Naive Bayes is well suited to this admission-status prediction task, achieving an accuracy of 93.25% and a ROC-AUC of 0.987, comfortably exceeding the 80% target and consistent with the high-accuracy range reported for binary-admission Naive Bayes studies in Table 7 (Salsabila et al., 2024), (Santoso et al., 2023). This supports the general finding that Naive Bayes performs competitively when the conditional-independence assumption in Equation (2) is not severely violated and when a small number of attributes carry most of the discriminative signal, as is the case here with *Entrance_Test_Score* and *Average_Report_Score* (Figure 5). Where this assumption is more severely violated, filter-based attribute-selection approaches have been shown to improve Naive Bayes performance by reducing redundancy among correlated attributes before training (Subramanian et al., 2023), a strategy that could be explored in future extensions of this model.

At the same time, Decision Tree and Random Forest both outperformed Naive Bayes by four to six percentage points, consistent with prior comparative findings that tree-based and ensemble methods can outperform Naive Bayes on similar prediction tasks (Amri et al., 2023), (Brianorman & Sucipto, 2024). This is consistent with the general expectation that tree-based ensembles can model non-linear thresholds and attribute interactions (e.g., a joint cutoff on both *Entrance_Test_Score* and *Average_Report_Score*) more flexibly than a model that assumes conditional independence between attributes. The practical implication is a trade-off: Naive Bayes offers a simpler, faster, and more directly interpretable probabilistic model with only a modest accuracy cost relative to Decision Tree and Random Forest, whereas the latter two offer the highest raw accuracy at the cost of a more complex, less directly interpretable decision process. For an operational PMB decision-support tool where interpretability and transparency of the accept/reject rationale matter particularly for applicant-facing explanations, Naive Bayes remains an attractive choice despite not achieving the very highest accuracy.

The comparatively poor performance of k-Nearest Neighbor is a methodological finding in its own right: it illustrates that label encoding, while adequate for Naive Bayes and tree-based models, is not a safe default preprocessing choice for distance-based algorithms applied to nominal categorical data. Future replications of this comparison should use one-hot encoding (or a Gower-type distance) specifically for kNN and other distance-based methods before drawing conclusions about their relative performance.

This research has limitations that should be addressed in future work, including: (1) the model was evaluated on a single stratified 80:20 hold-out split rather than k-fold cross-validation, so the reported metrics may be somewhat sensitive to the particular split used, although the high ROC-AUC suggests the result is unlikely to be an artifact of a favorable split; (2) all categorical attributes were label-encoded uniformly across all five algorithms, which is a plausible explanation for k-Nearest Neighbor's low accuracy and should be revisited with encoding schemes tailored to each algorithm; (3) missing values in the Achievement attribute were treated as a distinct "None" category, which is a defensible modeling choice but assumes that a missing record genuinely reflects the absence of a competition achievement rather than a data-entry gap; (4) the dataset originates from a single institution (UNUSIDA), which may limit the generalizability of the findings to other universities with different admission criteria; and (5) hyperparameters for the benchmark classifiers were left at their

default or commonly used values rather than being tuned, so the comparison in Table 6 reflects out-of-the-box performance rather than each algorithm's best achievable result.

Conclusion

This study developed and evaluated a Gaussian Naive Bayes model for predicting new-student admission status (Accepted/Rejected) at UNUSIDA using 2,000 historical PMB records with 14 predictor attributes, benchmarked against four alternative classifiers on an identical data pipeline. The detailed findings of this research are summarized as follows:

1. The Gaussian Naive Bayes model achieved an overall accuracy of 93.25%, a macro-average Precision of 0.925, Recall of 0.933, F1-score of 0.929, and a ROC-AUC of 0.987 on the 400-record testing set, exceeding the 80% accuracy target set for this study.
2. The confusion matrix showed a false-rejection rate of 17/250 (6.8%) and a false-acceptance rate of 10/150 (6.7%), indicating that the model performs consistently across both the Accepted and Rejected classes rather than being biased toward the majority class.
3. Benchmarking against four other classifiers on the identical dataset and preprocessing pipeline showed that Decision Tree (99.00%) and Random Forest (98.50%) achieved the highest raw accuracy, Support Vector Machine performed comparably to Naive Bayes (93.25%), and k-Nearest Neighbor underperformed substantially (64.50%), most likely due to its sensitivity to label-encoded categorical attributes rather than to a genuine limitation of distance-based classification in general.
4. Feature-importance analysis using a Random Forest surrogate model identified Entrance_Test_Score and Average_Report_Score as the dominant predictors of admission outcome, jointly accounting for more than 75% of total importance, while socioeconomic and demographic attributes contributed comparatively little.
5. Naive Bayes is recommended as a practical baseline model for an admission decision-support module within UNUSIDA's PMB information system, given its competitive accuracy, low computational cost, and directly interpretable class probabilities, while Decision Tree or Random Forest may be considered in scenarios where maximizing raw accuracy is the institution's priority.

The main contribution of this research is a comprehensively validated, multi-metric evaluation of Naive Bayes for admission-status prediction covering Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix analysis combined with a direct, same-pipeline benchmark against four alternative classifiers and an explicit feature-importance analysis, which together provide a more complete picture than accuracy alone. As noted in Section 3.2, this study is limited by its single-institution scope, its use of a single train-test split, and its uniform categorical encoding across algorithms with different sensitivity to that encoding. Future work should apply k-fold cross-validation, tune hyperparameters for each benchmark classifier, test encoding schemes tailored to distance-based methods such as k-Nearest Neighbor, and validate the model on admission data from additional institutions before operational deployment.

References

- Alam, A., & Mohanty, A. (2023). Predicting students' performance employing educational data mining techniques, machine learning, and learning analytics. In *Communication, networks and computing (CNC 2022)* (Vol. 1893). Springer. https://doi.org/10.1007/978-3-031-43140-1_15
- Amri, Z., Kusriani, & Kusnawi. (2023). Prediksi tingkat kelulusan mahasiswa menggunakan algoritma Naive Bayes, Decision Tree, ANN, KNN, dan SVM. *Edumatic: Jurnal Pendidikan Informatika*, 7(2), 187–196. <https://doi.org/10.29408/edumatic.v7i2.18620>
- Brianorman, Y., & Sucipto, S. (2024). Prediction of prospective new students using Decision Tree, Random Forest, and Naive Bayes. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 18(3), 1433–1446.

- Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera Viedma, E., Fujita, H., & Md. Ghani, N. A. (2021). Multiclass prediction model for student grade prediction using machine learning. *IEEE Access*, 9, 95608–95621. <https://doi.org/10.1109/ACCESS.2021.3094422>
- Gunawan, G., Hanes, H., & Catherine, C. (2021). C4.5, K-nearest neighbor, Naïve Bayes, and Random Forest algorithms comparison to predict students' on time graduation. *Indonesian Journal of Artificial Intelligence and Data Mining*, 4(2), 62–71. <https://doi.org/10.24014/ijaidm.v4i2.10833>
- Jacob, D., & Henriques, R. (2023). Educational data mining to predict bachelors students' success. *Emerging Science Journal*, 7, 159–171.
- Jajuly, R., Maryono, R., Gunawan, A. A., Heriyana, N., & Saifudin, A. (2024). Analisis perbandingan kinerja algoritma C4.5 dan Naive Bayes dalam prediksi penerimaan mahasiswa baru di Universitas Pamulang. *JRIIN: Jurnal Riset Informatika dan Inovasi*, 2(4).
- Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4), 531–538. <https://doi.org/10.1002/sam.11583>
- Pajila, P. J. B., Sheena, B. G., Gayathri, A., Aswini, J., Nalini, M., & Subramanian, R. S. (2023). A comprehensive survey on Naive Bayes algorithm: Advantages, limitations and applications. In 2023 4th International Conference on Smart Electronics and Communication (ICOSEC) (pp. 1228–1234). *IEEE*. <https://doi.org/10.1109/ICOSEC58147.2023.10276274>
- Purnama, P. A. W., Pohan, N., & Restiady, A. (2024). Naive Bayes algorithm classification for predicting graduation rate. *Jurnal Teknologi dan Open Source*, 7(2), 138–144. <https://doi.org/10.36378/jtos.v7i2.3866>
- Putri, S. E. Y., Saniati, & Surahman, A. (2021). Penerapan model Naive Bayes untuk memprediksi potensi pendaftaran siswa di SMK Taman Siswa Teluk Betung berbasis web. *Journal of Engineering, Computer Science and Information Technology (JECSIT)*, 1(1), 82–87.
- Salsabila, A., Nasution, M., & Irmayanti, I. (2024). Implementation of the Naïve Bayes algorithm to predict new student admissions. *Building of Informatics, Technology and Science (BITS)*, 6(1), 421–429. <https://doi.org/10.47065/bits.v6i1.5363>
- Santoso, F., Sunardi, & Lukman, H. Z. (2023). Implementation of data mining using the Naive Bayes method to predict acceptance of new students at MTS NU Islamiyah Asembagus. *G-Tech: Jurnal Teknologi Terapan*, 7(4), 1355–1366. <https://doi.org/10.33379/gtech.v7i4.3086>
- Sinaga, S., Sembiring, R. W., & Sumarno, S. (2022). Penerapan algoritma Naive Bayes untuk klasifikasi prediksi penerimaan siswa baru. *Journal of Machine Learning and Data Analytics (Journal Malda)*, 1(2).
- Subramanian, R. S., Girija, P., Sudha, K., Aswini, J., SivaKumar, S., & Nattesan, N. V. S. (2023). Alleviating the Naive Bayes assumption using filter approaches. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1430–1436). *IEEE*.
- Susanti. (2023). Perbandingan penerapan algoritma Decision Tree C4.5 dan Naive Bayes dalam analisa kelulusan siswa pada SMK Swadhipa 2 Natar Kabupaten Lampung Selatan. *Jurnal SIMADA*, 6(2).
- Xiao, J., Wang, Y., Chen, J., Xie, L., & Huang, J. (2021). Impact of resampling methods and classification models on the imbalanced credit scoring problems. *Information Sciences*, 569, 508–526.
- Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), 1–19. <https://doi.org/10.1186/s40561-022-00192-z>