Journal of Information Technology and Computer Science
Volume 11, Number 1, April 2026, pp. 127-138
Journal Homepage: www.jitecs.ub.ac.id

Naive Bayes with SMOTE for Predicting the Competitiveness of Vocational School Graduates on Imbalanced Data (Case Study: SMK Negeri 3 Malang)

Arsy Kurnia Fitri[1], Satrio Hadi Wijoyo*[2], Uun Hariyanti[3]
[1,2,3]Computer Science Faculty, Universitas Brawijaya, Malang
{arsykurniaa@student.ub.ac.id, satriohadi@ub.ac.id, uunhy@ub.ac.id}
*Corresponding Author

Received 23 January 2026; accepted 26 May 2026

Abstract. Vocational high schools (SMK) aim to produce work-ready graduates. However, the open unemployment rate (TPT) for SMK graduates remains high at 9.01%, indicating a significant competency gap. This study designs a model to predict graduates workforce competitiveness using the Naive Bayes algorithm combined with the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is employed to address the class imbalance between capable and incapable graduates. The study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, utilizing academic scores and tracer study datasets. Evaluation results demonstrate that applying SMOTE with a 70:30 train-test split successfully increased model accuracy to 97%. Notably, the model effectively detects the minority class with a Recall of 90%. Furthermore, cross-validation yielded an average accuracy of 97.66%, demonstrating stable performance. Finally, the model was implemented as a web-based dashboard to serve as an early warning system for schools.

Keywords: Data Mining, Naive Bayes, SMOTE, Work Readiness Prediction, Vocational High School, Data Imbalance

1 Introduction

Vocational education, particularly at the Vocational High School (SMK) level, plays a strategic role in preparing a skilled workforce ready to meet industrial demands. However, data from the Central Statistics Agency (BPS) reveals that in August 2024, the Open Unemployment Rate (TPT) for SMK graduates remained the highest among all education levels at 9.01% [1]. This statistic highlights a persistent gap between graduate competencies and industry requirements. A similar trend has been observed at SMK Negeri 3 Malang, where preliminary interviews suggest a decline in graduate competitiveness over the last two years.

Currently, assessments of work readiness in schools typically rely on academic transcripts and Field Work Practice (PKL) scores, which combine technical (hard skills) and non-technical (soft skills) evaluations. To address these challenges, this study employs a data mining approach using the Gaussian Naive Bayes algorithm (GNB). Gaussian Naive Bayes was selected because the dataset consists of continuous numerical variables, such as academic and internship scores, which are suitable for Gaussian distribution modeling. In addition, the algorithm is computationally efficient, simple to implement, and has demonstrated reliable performance in classification tasks [2].

The main challenge in the graduation dataset is class imbalance, where the number of graduates in the 'qualified' category far exceeds the number in the 'unqualified' category. This imbalance can lead to prediction bias in classification models. To address this issue, this study applies the SMOTE to balance class distributions before the modeling process. The SMOTE method was chosen because it can improve model performance such as accuracy, recall, F1-score, and AUC, and reduce the risk of overfitting by generating more varied synthetic data [3]. Thus, this study aims to design and evaluate a graduate competency prediction model and generate recommendations that can be used to support improvements in graduate quality.

2 Literature Review

Numerous studies have explored the use of the Naive Bayes algorithm for predicting graduate competency and academic performance. Jempol (2024) applied Naive Bayes to vocational high school graduate data with nine variables and achieved an accuracy of 95.6% [4]. Meanwhile, Hidayat (2021) combined Naive Bayes with Backward Elimination on 855 student records and obtained an accuracy of 96.95% [5]. These findings indicate that Naive Bayes performs well in educational data classification. However, educational datasets often face class imbalance issues, where one class contains significantly more data than another, potentially causing the model to become biased toward the majority class. Therefore, this study applies SMOTE to balance the data distribution before the classification process.

The use of SMOTE has been proven to enhance the performance of classification models. Research by Kurniadi, Nuraeni, and Lestari (2022) showed that the combination of SMOTE and Feature Forward Selection improved recall and precision for the minority class, despite achieving a final accuracy of 87.13% [6]. In addition, Herudin, Faqih, and Bahtiar (2022) applied SMOTE with Naive Bayes for new student classification and achieved an accuracy of 92.58% [7]. These studies indicate that SMOTE is effective in improving classification performance on imbalanced datasets.

In this study, Gaussian Naive Bayes algorithm was selected because the dataset consists of continuous numerical data, such as academic grades and internship scores, making it suitable for continuous data and computationally efficient [2]. Research conducted by Farmawati and Narti showed that Naive Bayes achieved higher accuracy compared to C4.5 decision tree algorithm [8], while a study by Asnawi et al. comparing Naive Bayes, K-NN, and SVM demonstrated that Naive Bayes obtained the best performance with an accuracy of 79.8% [9]. Although this algorithm has limitations related to the assumption of independence among variables, which is often not fulfilled in educational contexts, various studies within the last five years have shown that Naive Bayes can achieve prediction accuracies ranging from 80% to 88% on educational data. These findings indicate that Naive Bayes has strong classification performance and the potential to be used as a decision-support tool in academic settings [10].

SMOTE was selected because the dataset size was relatively limited, making oversampling more appropriate than undersampling. Undersampling can reduce the number of majority class samples and risk losing important information required during the classification process. In contrast, SMOTE generates synthetic samples for the minority class based on the k-nearest neighbors without removing the original data, resulting in a more balanced data distribution [11]. However, SMOTE has a limitation in that it may cause class overlapping, which can potentially affect model performance [12]. Therefore, this study compares the model using SMOTE with the baseline model trained on the original imbalanced dataset to evaluate the impact of SMOTE implementation on model performance.

Based on previous studies, the use of Naive Bayes and SMOTE has been proven to improve classification performance in educational data, particularly in handling imbalanced datasets. However, studies focusing on graduate competency prediction integrated with a school monitoring system are still limited. Therefore, this study develops a graduate competency prediction model using Gaussian Naive Bayes and SMOTE and presents it through a dashboard prototype as an initial step toward an early warning system.

3 Method

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. CRISP-DM is widely recognized as a comprehensive methodology for data mining projects and remains one of the most standard approaches in data science [13]. The workflow begins with the Business Understanding phase, which encompasses problem identification and a literature review. Subsequently, the process proceeds through Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, comprising six distinct phases [14]. The complete research workflow is illustrated in Figure 1.

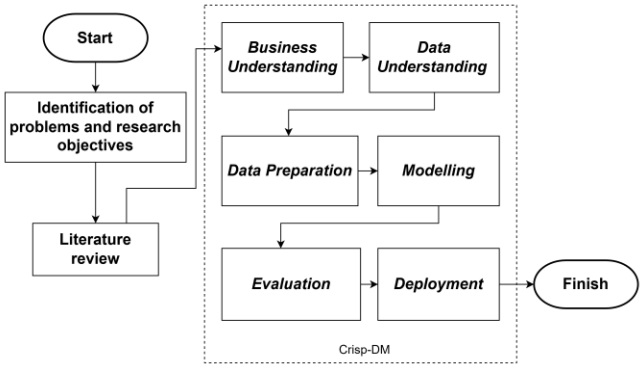

Figure 1. CRISP-DM research flow

3.1 Business Understanding

This initial phase focuses on understanding the project's objectives and requirements from a business perspective [15]. In this study, the Business Understanding phase began by identifying the primary problem: the low competitiveness of SMK Negeri 3 Malang graduates in the labor market. This issue was analyzed through consultations with teachers and an examination of available tracer study data. The analysis revealed that academic factors significantly influence graduates' work readiness. Consequently, this study aims to develop a prediction model using the Naive Bayes algorithm to provide an objective assessment of graduate competitiveness and serve as a basis for the school to formulate effective learning strategies.

3.2 Data Understanding

The Data Understanding phase began with the collection of academic report card data and tracer study data [15]. The dataset consisted of 119 graduate samples from the Computer and Network Engineering (TKJ) program for the 2024–2025 cohorts, based on the academic and tracer study records available from the school. Although the dataset size is relatively limited and may affect the generalizability of the findings, it reflects the actual data availability within the study context. Initial data exploration was

conducted to identify score distributions, missing values, and class imbalance between capable and incapable graduates.

The dataset included academic attributes representing graduates' hard skills and soft skills, such as knowledge scores, skill scores, internship scores, and average scores. The codes P and K represent knowledge and skill scores, while the numbers 2 and 3 indicate Grade 11 and Grade 12, respectively. A complete description of all attribute codes and their corresponding subjects is presented in Table 1.

Tabel 1. Attribute code mapping

Code	Subject Name/Meaning
DTKJ	Fundamentals of Computer Network and Telecommunications Engineering
DTK	Fundamentals of Computer Engineering
DDG	Fundamentals of Graphic Design
DDM	Fundamentals of Digital Marketing
Jarnil	Wireless Networks
KPJ	Network Device Configuration
ASJ	Network System Administration
TLJ	Network Service Technology
KKA	Coding and Artificial Intelligence
KJ	Network Security
PKK	Creative Products and Entrepreneurship
PKL	Field Work Practice (Internship)
AVG	Average Academic Score

3.3 Data Preparation

The Data Preparation phase was carried out to prepare the dataset before the modeling process [15]. This stage included data cleaning, such as removing irrelevant attributes, handling missing values, and standardizing data formats. Feature scaling was then applied to normalize the range of variables, and the dataset was divided into training data (70%) and testing data (30%).

To handle class imbalance and reduce prediction bias, the SMOTE was applied only to the training data, while the testing data remained in its original distribution. In this study, SMOTE used 5 nearest neighbors with a sampling strategy of 1.0 to balance the capable and incapable classes. A `random_state` value of 42 was also used to ensure reproducible results. This phase aimed to produce a clean, balanced, and consistent dataset before the modeling stage.

3.4 Modelling

The balanced training data was used to develop a Gaussian Naive Bayes classification model. The model was implemented using the Scikit-learn library by calculating prior probabilities for each class and conditional probabilities for each attribute. Although Gaussian Naive Bayes assumes independence among features, which may not always be fulfilled due to correlations between academic attributes, previous studies have shown that the algorithm still provides stable and effective classification performance in educational data mining. To address this, the model's practical robustness is strictly verified during the evaluation phase. During the classification process, the model assigns the class with the highest posterior probability to a new instance to classify graduates into "capable" or "incapable" categories.

3.5 Evaluation

The developed models were evaluated using the testing dataset to measure classification performance. Evaluation was conducted using a confusion matrix and several performance metrics, including Accuracy, Precision, Recall, and F1-Score. The evaluation compared three experimental scenarios: Gaussian Naive Bayes trained on the original imbalanced dataset, Gaussian Naive Bayes trained on the SMOTE-balanced dataset, and a Decision Tree classifier trained on the same SMOTE-balanced dataset.

The confusion matrix was used to analyze prediction results based on four components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [16][17]. Comparative confusion matrices were analyzed to evaluate the impact of SMOTE on minority class recognition and overall model performance. Furthermore, Stratified 5-Fold Cross-Validation was applied to assess the stability and generalization capability of the best-performing model across different data partitions.

3.6 Deployment

According to the CRISP-DM methodology, the knowledge generated from the model should be presented in a form that is understandable and useful for stakeholders. In this study, the prediction results were presented through a scientific report and a web-based dashboard prototype. The dashboard was developed as a Proof of Concept (PoC) for an early warning system to support schools in monitoring graduate competency predictions. To address deployment sustainability, the current implementation is intentionally limited to a local tunneling environment (via Pyngrok) for demonstration and preliminary validation purposes. User validation for this prototype was conducted through structured interviews with school administrators to gather objective feedback regarding the interface's utility and functional limitations before full-scale production deployment.

4 Results and Discussion

4.1 Business Understanding Phase Results

The Business Understanding phase commenced by establishing the primary business objective: to address the competency gap among SMKN 3 Malang graduates through the development of a data-driven early warning system. This system is designed to provide actionable intervention recommendations. Recognizing the challenge of class imbalance within the academic and internship (PKL) dataset, the technical data mining goal focused on constructing a prediction model using the Gaussian Naive Bayes algorithm integrated with the Synthetic Minority Over-sampling Technique (SMOTE). This integration aims to preserve the model's sensitivity towards the minority class (incapable). The project plan was systematically structured according to the CRISP-DM lifecycle ranging from data preparation and model training to stability evaluation via cross-validation. Finally, the solution was deployed as an interactive Streamlit-based dashboard to serve as a practical decision-support tool for the school.

4.2 Data Understanding Phase Results

4.2.1 Description and Feature Type Identification

The initial dataset was constructed by integrating academic transcripts and tracer study records from 119 graduates within the TKJ program (2024–2025 cohorts). The resulting dataset comprises 33 attributes, including knowledge scores (P), skill scores (K), and Field Work Practice (PKL) marks. To provide a comprehensive baseline representation of student competencies, all 33 original attributes were retained for the initial modeling phase without applying preliminary feature selection. This decision

aligns with the foundational baseline assumption that every vocational competency indicator potentially correlates with workforce competitiveness. The structural identification of these data types and their operational ranges is systematized in Table 2.

Tabel 2. Identification of dataset attributes

Attribute Category	Data Type	Measurement Range / Indicator
Student Identifiers (No, Name)	String	Unique
Knowledge and Skill Scores	Numeric	Range 1 - 100
PKL (Internship) Score	Numeric	Range 1 -100
Work Readiness Status	Binary	Target Class Label: Capable/Incapable

4.2.2 Data Verification

A data integrity check was conducted to evaluate the completeness of the dataset. The results confirmed that there were no missing values across all 33 attributes and 119 samples, eliminating the need for data imputation or data removal techniques. An exploratory analysis of the target variable distribution was then performed to examine class balance. The analysis showed a class imbalance in the dataset, where 90 graduates (75.6%) were categorized as “Capable” and only 29 graduates (24.4%) were categorized as “Incapable.” This imbalance indicates the need for an oversampling technique during the data preparation phase to reduce majority class bias in the classification model. The baseline distribution of the target variable is presented in Figure 2.

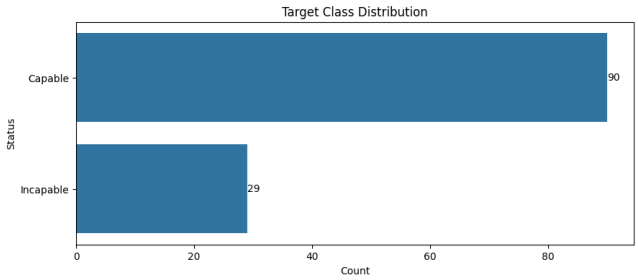

Figure 2. Visualization of target attribute distribution

4.3 Data Preparation Phase Results

4.3.1 Data Cleaning, Encoding, and Feature - Target Separation

The data preparation phase commenced with cleaning and formatting operations to ensure the matrix was structurally optimized for machine learning. Irrelevant identifying attributes, specifically the No. and Name columns, were excluded to eliminate non-predictive features that could potentially introduce statistical noise. Following feature removal, label encoding was applied to the categorical target variable (Status), mapping the text label "Capable" to 1 and "Incapable" to 0. This conversion transformed the target variable into a numerical format compatible with algorithmic processing. Finally, the preprocessed dataset was structurally partitioned into an independent feature matrix (x), encompassing all 33 academic subject grades and Field Work Practice (PKL) scores, and a dependent target vector (y)

4.3.2 Dataset Splitting and SMOTE Application

Following the feature-target separation, the dataset was partitioned into training and testing sets. Based on empirical analysis and strong support from prior literature, the 70:30 ratio was established as the optimal configuration. This decision is grounded in statistical analysis indicating that a 70:30 proportion offers the most effective balance for data splitting [18]. Furthermore, related studies implementing the Naive Bayes algorithm with SMOTE have demonstrated that a 70:30 split yields higher accuracy rates compared to alternative ratios [6]. The detailed distribution of samples across both partitions, including the baseline imbalanced state, is structured in Table 3.

Tabel 3. Sample Distribution Across Training and Testing Partitions

Target Class Label	Original Dataset	Training (70%)	Testing (30%)
Class 1 (Capable)	90	64	26
Class 0 (Incapable)	29	19	10
Total Samples	119	83	36

To mitigate prediction bias without compromising test data authenticity, SMOTE was applied solely to the 83 training samples listed in Table 3. Through this process, the 19 "Incapable" training instances were successfully upsampled to 64, achieving a perfectly balanced training distribution of 128 total samples. A visualization comparing the training data matrix pre and post SMOTE application is presented in Figure 3.

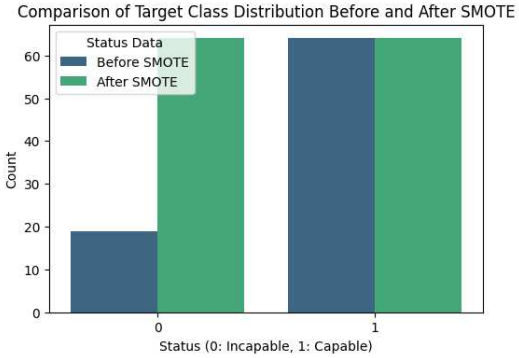

Figure 3. Comparison of data before and after SMOTE

4.4 Modelling Phase Results

The modeling phase evaluated the Gaussian Naive Bayes (GNB) algorithm, chosen for its computational efficiency and stability on small datasets. GNB assumes feature independence [19]. However, Pearson correlation analysis on the 33 academic attributes showed a moderate localized correlation between knowledge (P) and skill (K) scores. While this violates strict independence, prior literature shows GNB remains highly robust because feature dependencies rarely shift the final classification boundary.

To systematically evaluate the impact of class balancing and verify the superiority of the proposed framework, the modeling process was executed across three experimental scenarios. The first scenario established a baseline by training the GNB model directly on the original, imbalanced training data consisting of 64 Capable and 19 Incapable samples. The second scenario represents the proposed framework, where the GNB model was trained on the training dataset balanced via SMOTE using five nearest neighbors, resulting in an equalized distribution of 64 Capable and 64 Incapable

samples. The third scenario served as an algorithmic benchmark by training a Decision Tree classifier on the identical SMOTE. All three models were subsequently prepared for the evaluation phase to analyze their predictive performance on the test data.

4.5 Evaluation Phase Results

4.5.1 Impact of SMOTE on Classification Performance

To objectively evaluate the impact of the SMOTE oversampling technique, the performance of the Gaussian Naive Bayes model was assessed under two different scenarios: Scenario 1 (Baseline GNB without SMOTE) and Scenario 2 (Proposed GNB with SMOTE). The empirical differences in classification performance between these scenarios are illustrated through comparative confusion matrices and classification results presented in Table 4.

Tabel 4. Comparison Before and After SMOTE

Model	Accuracy	Label	Precision	Recall	F1-Score
GNB without SMOTE	0.94	0	0.90	0.90	0.90
		1	0.96	0.96	0.96
GNB + SMOTE	0.97	0	1.00	0.90	0.95
		1	0.96	1.00	0.98

Based on the evaluation results of Scenario 1 (Baseline GNB without SMOTE), the model achieved an overall accuracy of 94%. Under the original imbalanced data condition, the majority class (label 1/capable) obtained precision, recall, and F1-score values of 0.96. Meanwhile, the minority class (label 0/incapable) achieved a precision of 0.90, recall of 0.90, and F1-score of 0.90. Although the baseline model demonstrated relatively strong performance, several misclassifications were still observed due to the tendency of the decision boundary to favor the majority class distribution.

In contrast, the implementation of SMOTE in Scenario 2 significantly improved the classification capability of the Gaussian Naive Bayes model. The application of SMOTE increased the overall model accuracy to 97%. More importantly, the balanced training data improved minority class recognition, increasing the minority class F1-score to 0.95 and achieving a precision value of 1.00. In addition, the majority class recall reached 1.00, indicating that all capable graduates were correctly classified. These findings demonstrate that SMOTE effectively enhanced the model's sensitivity toward minority class detection without disrupting the original data distribution, resulting in more balanced and reliable classification performance.

4.5.2 Algorithm Benchmark Comparison

Tabel 5 Comparison Before and After SMOTE

Model	Accuracy	Label	Precision	Recall	F1-Score
Decision Tree + SMOTE	0.91	0	0.89	0.80	0.84
		1	0.93	0.96	0.94
GNB + SMOTE	0.97	0	1.00	0.90	0.95
		1	0.96	1.00	0.98

To validate the effectiveness of the proposed model, a benchmark experiment was conducted using Scenario 3 (Decision Tree + SMOTE). Both models were evaluated using the same testing dataset and experimental configuration to ensure fair

comparison. The comprehensive performance metrics for this algorithm comparison are summarized in Table 5.

Table 5 presents the benchmark comparison between the Gaussian Naive Bayes (GNB) and Decision Tree classifiers trained using the same SMOTE-balanced dataset. The results indicate that GNB + SMOTE achieved superior overall performance, obtaining an accuracy of 97%, compared to 91% achieved by Decision Tree + SMOTE. In addition, GNB demonstrated better minority class recognition, particularly for the incapable class (label 0), with higher precision, recall, and F1-score values. These findings suggest that Gaussian Naive Bayes is more suitable and consistent for predicting graduate competitiveness within the current educational dataset.

According to Table 5, the GNB + SMOTE model achieved strong classification performance across both classes. The incapable class (label 0) obtained a precision of 1.00, recall of 0.90, and F1-score of 0.95, while the capable class (label 1) achieved a precision of 0.96, recall of 1.00, and F1-score of 0.98. These results indicate that the model effectively classified both majority and minority classes with minimal misclassification. The confusion matrix visualization is presented in Figure 4.

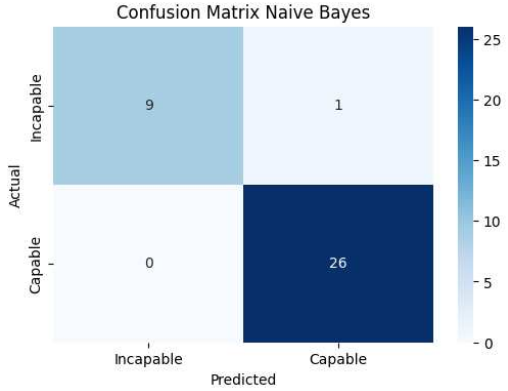

Figure 4. Heatmap confusion matrix

Figure 4 shows that the model correctly classified most test instances. The model produced 26 True Positives (TP) and 9 True Negatives (TN), with only one False Positive (FP). No False Negatives (FN) were detected, indicating reliable classification performance.

4.5.3 Cross Validation Results

Further evaluation was conducted using Stratified 5-Fold Cross-Validation to assess the stability and generalization capability of the best-performing model, namely GNB + SMOTE. This method partitions the dataset into five folds while maintaining balanced class distributions in each fold [20] [21]. The experimental results showed that the model achieved a mean accuracy of 97.66%, with fold accuracies ranging from 96.15% to 100% and a standard deviation of 0.0191. These results indicate low performance variance, demonstrating that the model is stable and capable of generalizing consistently across different data partitions.

4.6 Deployment Phase Results

The classification model was implemented through a Streamlit-based dashboard prototype. The interface allows users to input student academic data and obtain graduate competitiveness predictions in real time. The trained model, stored in .pkl format, was integrated into the application as the prediction engine. For demonstration and preliminary testing purposes, the dashboard was hosted using Google Colaboratory and

exposed through a Pyngrok tunnel. The dashboard interface is illustrated in Figure 6.

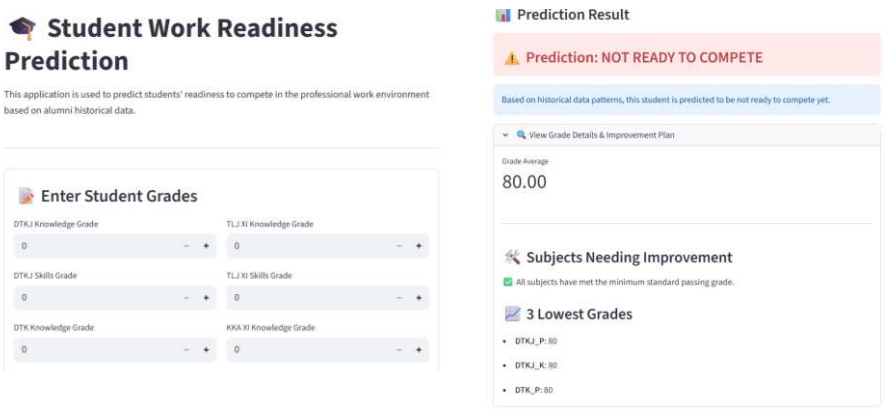

Figure 6. Prediction Dashboard Interface

Figure 6 presents the interface of the prediction dashboard. The left section provides an input form for student academic scores, while the right section displays prediction results, grade summaries, and subject improvement recommendations generated by the model. Initial functional testing indicated that the dashboard operated properly in generating predictions and displaying recommendation outputs. Based on preliminary feedback from school representatives, the interface was considered sufficiently clear and easy to use for monitoring student competency predictions. However, the current implementation remains a prototype and has not yet been deployed in a permanent production environment. Therefore, further development and structured usability evaluation are recommended before large-scale implementation in schools.

4.7 Discussion of Analysis Results

Initial exploratory analysis revealed a substantial imbalance between the capable and incapable graduate classes, where the majority class dominated the dataset distribution. This condition may lead to prediction bias toward the majority class and reduce minority class recognition performance [11], [22]. To address this issue, SMOTE was applied to balance the training data distribution before the modeling process. The experimental results demonstrated that the application of SMOTE improved the classification performance of the Gaussian Naive Bayes model, particularly in detecting the minority class. Compared to the baseline model trained on the original imbalanced dataset, the SMOTE-balanced model achieved higher overall accuracy and more stable precision, recall, and F1-score values across both classes. These findings indicate that synthetic minority oversampling effectively reduced classification bias without significantly altering the original data characteristics.

Benchmark comparison results further showed that Gaussian Naive Bayes outperformed the Decision Tree classifier under the same SMOTE-balanced configuration. The GNB + SMOTE model achieved higher accuracy and better minority class recognition, indicating that Gaussian Naive Bayes was more suitable for the current educational dataset. These findings are consistent with previous studies conducted by Farmawati and Narti [8] as well as Asnawi et al. [9], which reported competitive performance of Naive Bayes compared to other classification algorithms. Although Gaussian Naive Bayes assumes independence among predictor variables, several academic attributes in this study may exhibit correlations, particularly between

knowledge and skill scores within related subjects. Nevertheless, the evaluation results indicate that the model maintained stable and reliable performance despite potential violations of this assumption. Despite achieving strong classification performance, this study remains limited by the relatively small dataset size consisting of 119 graduate records from a single vocational program. Therefore, the findings may not yet fully generalize to broader educational contexts and should be interpreted within the scope of the current study.

5 Conclusion

This study developed a graduate work readiness prediction model for vocational high school students using the Gaussian Naive Bayes algorithm combined with the SMOTE technique to address class imbalance. Experimental results showed that the proposed GNB + SMOTE model achieved the best performance, obtaining an accuracy of 97% with strong precision, recall, and F1-score values across both majority and minority classes. In addition, Cross-Validation produced a mean accuracy of 97.66% with low performance variance, indicating that the model demonstrated stable and consistent classification capability.

Benchmark evaluation further showed that Gaussian Naive Bayes outperformed the Decision Tree classifier under the same SMOTE-balanced configuration, indicating that Gaussian Naive Bayes was more suitable for the current educational dataset. The developed model was subsequently implemented into a Streamlit-based dashboard prototype to support graduate competitiveness monitoring and prediction visualization.

Despite the promising results, this study remains limited by the relatively small dataset consisting of 119 graduate records from a single vocational program. Therefore, future research is recommended to expand the dataset scope, incorporate non-academic variables, and explore alternative balancing techniques or additional classification algorithms to further improve prediction performance and model generalizability. Future studies may also focus on transforming the current dashboard prototype into a fully integrated production system supported by cloud infrastructure and school information systems for sustainable real-world implementation.

References

1. BPS, ‘Tingkat Pengangguran Terbuka Berdasarkan Tingkat Pendidikan, 2024’, Badan Pusat Statistik.
2. J. T. Haqquesda and E. A. Pambudi, ‘Gaussian Naïve Bayes dengan Harmonic Mean untuk Klasifikasi Hasil Produksi Gula Merah’, Jurnal Media Pratama, vol. 18, no. 2, pp. 14–26, 2024.
3. J. Pardede and D. P. Pamungkas, ‘The Impact of Balanced Data Techniques on Classification Model Performance’, Scientific Journal of Informatics, vol. 11, no. 2, pp. 401–412, May 2024, doi: 10.15294/sji.v11i2.3649.
4. N. K. Jempol, ‘Penerapan Algoritma Naïve Bayes dalam Memprediksi Kemampuan Lulusan SMK untuk Bersaing Mendapatkan Pekerjaan’, Bali, 2024.
5. H. Hidayat, ‘Algoritma Naïve Bayes Berbasis Backward Elimination untuk Prediksi Kesiapan Kerja pada Siswa SMK’, Smart Comp, vol. 10, no. 2, 2021, doi: https://doi.org/10.30591/SMARTCOMP.V10I2.2492.
6. D. Kurniadi, F. Nuraeni, and S. M. Lestari, ‘Implementasi Algoritma Naïve Bayes Menggunakan Feature Forward Selection dan SMOTE Untuk Memprediksi Ketepatan Masa Studi Mahasiswa Sarjana’, Jurnal Sistem Cerdas, vol. 5, no. 2, pp. 63–82, 2022, doi: https://doi.org/10.37396/jsc.v5i2.215.
7. D. Herudin, A. Faqih, and A. Bahtiar, ‘Klasifikasi Penerimaan Peserta Didik Baru Menggunakan Algoritma Naïve Bayes dengan SMOTE Pada SMKN 1 Jamblang’, Jurnal Sistem Informasi dan Manajemen, vol. 10, no. 2, Aug. 2022, Accessed: Aug. 23, 2025. [Online]. Available: https://ejournal.stmikgici.ac.id/

8. F. Fatmawati and N. Narti, 'Perbandingan Algoritma C4.5 dan Naive Bayes Dalam Klasifikasi Tingkat Kepuasan Mahasiswa Terhadap Pembelajaran Daring', JTIM : Jurnal Teknologi Informasi dan Multimedia, vol. 4, no. 1, pp. 1–12, May 2022, doi: 10.35746/jtim.v4i1.196.
9. M. H. Asnawi, I. Firmansyah, R. Novian, and R. S. Pontoh, 'Perbandingan Algoritma Naïve Bayes, K-NN, dan SVM dalam Pengklasifikasian Sentimen Media Sosial', 2021. [Online]. Available: http://prosiding.statistics.unpad.ac.id
10. G. H. Hilmawan, 'Literatur Review: Efektifitas Penerapan Metode Naïve Bayes Untuk Prediksi Kelulusan', Jurnal Media Akademik, vol. 3, no. 6, pp. 3031–5220, 2025, doi: 10.62281.
11. I. K. Dharmendra, I. M. A. W. Putra, and Y. P. Atmojo, 'Evaluasi Efektivitas SMOTE dan Random Under Sampling pada Klasifikasi Emosi Tweet', Informatics for Educators And Professionals: Journal of Informatics, vol. 9, no. 2, pp. 192–193, Dec. 2024.
12. A. A. G. W. S. Erlangga, P. P. O. J. KW, and Adnan, 'Perbandingan Performa SMOTE dan Smote-Enn Pada Data Tidak Seimbang', Jurnal Ramatekno, vol. 5, no. 2, pp. 150–158, Nov. 2025.
13. M. Yasir and R. Suraji, 'Perbandingan Metode Klasifikasi Naïve Bayes, Decision Tree, Random Forest Terhadap Analisis Sentimen Kenaikan Biaya Haji 2023 pada Media Sosial Youtube', Jurnal Cahaya Mandalika, vol. 3, no. 2, pp. 180–192, 2023, doi: 10.36312/jcm.v3i2.1520.
14. Y. A. Singgalen, 'Analisis Performa Algoritma NBC, DT, SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CHRISP-DM', Building of Informatics Technology and Science (BITS), vol. 4, no. 3, pp. 1634–1646, 2022, doi: 10.47065/bits.v4i3.2766.
15. P. Chapman et al., CRISP-DM 1.0. DaimlerChrysler, 2000.
16. S. Swaminathan and B. R. Tantri, 'Confusion Matrix-Based Performance Evaluation Metrics', African Journal of Biomedical Research, vol. 27, no. 4, pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.
17. A. Tasnim, Md. Saiduzzaman, M. A. Rahman, J. Akhter, and A. S. Md. M. Rahaman, 'Performance Evaluation of Multiple Classifiers for Predicting Fake News', Journal of Computer and Communications, vol. 10, no. 09, pp. 1–21, 2022, doi: 10.4236/jcc.2022.109001.
18. K. M. Kahloot and P. Ekler, 'Algorithmic Splitting: A Method for Dataset Preparation', IEEE Access, vol. 9, pp. 125229–125237, 2021, doi: 10.1109/ACCESS.2021.3110745.
19. A. Mulyoto, Naïve Bayes pada Google Colabs. Purbalingga: EUREKA MEDIA AKSARA, 2024.
20. Wijiyanto, A. I. Pradana, Sopingi, and V. Atina, 'Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa', Jurnal Algoritma, vol. 21, no. 1, May 2024, doi: 10.33364/algoritma/v.21-1.1618.
21. S. Prusty, S. Patnaik, and S. K. Dash, 'SKCV: Stratified K-fold Cross Validation on ML Classifiers for Predicting Cervical Cancer', Frontiers in Nanotechnology, vol. 4, Aug. 2022, doi: 10.3389/fnano.2022.972421.
22. V. H. Barella, L. P. F. Garcia, M. C. P. de Souto, A. C. Lorena, and A. C. P. L. F. de Carvalho, 'Assessing The Data Complexity of Imbalanced Datasets', Inf. Sci. (N. Y)., vol. 553, pp. 83–109, Apr. 2021, doi: 10.1016/j.ins.2020.12.006.