**Computer Science (CO-SCIENCE)**
Volume 6 Issue 1 January 2026
Accreditation **Sinta 4** No. SK : 230/E/KPT/2022

# Analysis of Student Academic Performance Using Random Forest and Support Vector Machines

Galih Mifta Agung[1], Robi Aziz Zuama[2*], Eko Setia Budi[3]

[1,2,3]Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98 Kwitang, Kec. Senen, Jakarta Pusat, Indonesia
e-mail: [1]galihmiftaagung@gmail.com, [2]robi.rbz@bsi.ac.id, [3]eko.etb@bsi.ac.id

(*) Corresponding Author

Article Info: Received: 01-10-2025 | Revised : 10-12-2025 | Accepted : 19-12-2025

Abstracts – Assessing student academic performance objectively remains a challenge at SMP Negeri 16 Bogor due to diverse internal and external factors in student records. This study aims to compare the classification performance of the Random Forest and Support Vector Machine (SVM) algorithms using a dataset of 403 students containing demographic, socioeconomic, and school-related attributes. Although the attributes are not traditional academic indicators (e.g., assignment or exam scores), they are used to explore whether non-academic features can contribute to predictive models. Following data preprocessing—handling missing values, encoding categorical variables, and managing class imbalance—both algorithms were evaluated using accuracy, precision, recall, and confusion matrix analysis. Results show that SVM outperforms Random Forest with 78.00% accuracy, 89.98% precision, and 70.24% recall. These findings indicate that SVM is more robust for imbalanced classification tasks and can provide useful insights even when academic-performance labels are predicted from non-academic attributes.
Keywords : Academic Performance, SVM, Random Forest, Classification, Confusion Matrix

## INTRODUCTION

Information technology has significantly transformed various aspects of human life, including education. Large educational datasets, once limited to administrative records, are now recognized as valuable sources of information that can support data-driven decision-making through advanced data analysis techniques. One of the most widely adopted methods in this domain is *data mining*, which enables the discovery of meaningful patterns and relationships within large volumes of data (Yağcı, 2022a), Educational Data Mining (EDM), as a subfield of data science, has become instrumental in improving teaching quality and understanding factors that influence students' learning outcomes (Gul et al., 2025).

Student academic performance is a crucial indicator of educational success and institutional effectiveness. However, schools often face challenges in analyzing performance due to numerous internal and external factors that influence learning outcomes. For instance, variables such as students' previous school background, gender, age, residential environment, and family conditions can significantly affect academic achievement. SMP Negeri 16 Bogor, for example, faces difficulties in evaluating student performance objectively due to the diversity of these influencing factors. Identifying these key determinants is essential to assist both schools and parents in understanding students' learning needs more effectively and providing appropriate support.

To address such challenges, recent studies have applied machine learning (ML) techniques to predict and analyze academic performance (Khosravi & Azarnik, 2024) (Ying & Ma, 2024). ML, as a branch of artificial intelligence (AI), enables computers to process data, build predictive models, and make informed decisions without explicit programming (Ghosh et al., 2022b). It has been increasingly utilized in educational research to identify patterns in student data and to forecast future academic outcomes (Jawad et al., 2022). Recent literature has further demonstrated improvements in predictive accuracy when handling class-imbalance (Althaqafi et al., 2025) and emphasizing feature importance in RF models (Nachouki et al., 2023). Within this domain, classification is one of the most common tasks, which involves training algorithms on labeled datasets to categorize new data into specific performance groups (Dahal & Shakya, 2022). Classification techniques can handle diverse data types and provide insights into how various attributes contribute to academic results.

Among the popular classification algorithms, Random Forest (RF) and Support Vector Machine (SVM) have demonstrated strong performance in predicting student outcomes (Yağcı, 2022b) (Gul et al., 2025). RF is an ensemble learning method that constructs multiple decision trees from random subsets of data and combines their outputs to improve prediction accuracy and reduce overfitting (Ying & Ma, 2024). Comparative analyses suggest that RF tends to outperform other models including SVM when structured properly (Chen & Jin, 2024), yet SVM

and its regression variant remain relevant in nuanced settings (Durai et al., 2025). RF is robust to missing values and efficient for handling large datasets. SVM, on the other hand, is a discriminative model that identifies the optimal hyperplane separating data classes by maximizing the margin between them (Ghosh et al., 2022b). While SVM offers high accuracy and is effective for linearly and non-linearly separable data, it may perform inconsistently when features overlap or data imbalance occurs (Jawad et al., 2022).

Previous studies have implemented various algorithms to classify and predict student academic performance. For instance, (Muhaimin et al., 2024) used the K-Nearest Neighbor (KNN) algorithm to classify students based on academic scores and discipline, while (Budiyanto et al., 2024) developed a prediction model for cum laude graduation rates using machine learning. (Azizah et al., 2022) applied the Decision Tree method to produce interpretable prediction models, and (Gori et al., 2024) integrated Naïve Bayes with Correlation-Based Feature Selection (CFS) for better classification accuracy. Other studies, such as (Naibaho & Zahra, 2023), used Decision Tree, Random Forest, and Extreme Gradient Boosting to predict student graduation rates. Moreover, systematic reviews show that many studies still focus on higher education and limited attributes (Rodrigues et al., 2022), which underscores the need for broader attribute scopes as in this study.

While many previous studies focus largely on higher education or rely heavily on academic attributes such as exam scores, assignment grades, or attendance (Rodrigues et al., 2022), research using broader non-academic attributes remains limited. This creates a gap that this study aims to address.

Additionally, inconsistencies in earlier descriptions of dataset size (1,153 vs. 403) are clarified in this study: after preprocessing and data validation, the final usable dataset consisted of 403 student records.

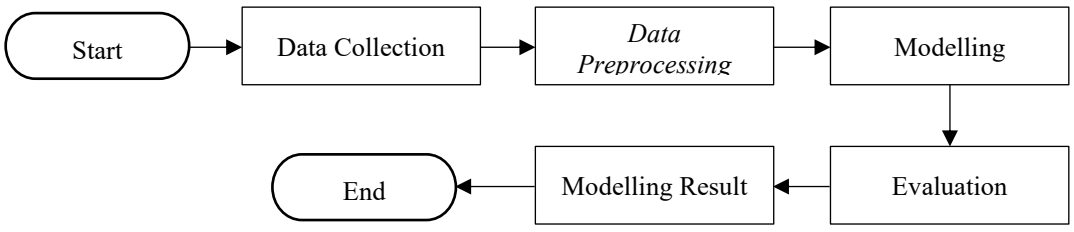
To address these limitations, this study incorporates a wider range of demographic, socioeconomic, and environmental attributes—such as parental education, occupation, household income, type of residence, transportation mode, travel distance, and travel time—to explore their relationship with student academic performance.

Furthermore, this study provides a comparative evaluation of Random Forest and Support Vector Machine, addressing reviewer concerns about the need for clearer justification of algorithm selection by focusing on their known strengths and weaknesses in handling imbalanced and heterogeneous data.

Through this comparative approach, the study aims to identify which model performs more effectively for student performance classification at SMP Negeri 16 Bogor, providing practical insights for data-driven decision-making within the school.

## RESEARCH METHOD

This research uses a quantitative research method with data collection techniques through observation, interviews, and literature study. This research was conducted at SMP Negeri 16 Bogor. The collected data is then processed using machine learning methods to obtain academic performance predictions based on accuracy and model evaluation values (Romero & Ventura, 2020) (Han et al., 2022). The testing method in this research uses the confusion matrix to evaluate the performance of the classification model applied in predicting student academic performance. The confusion matrix will be used to calculate accuracy, precision, and recall. The way this evaluation works is by comparing the model's prediction results with the actual data, which can then provide an overview of how well the model classifies student data into the correct categories. More specifically, the performance of the confusion matrix is evaluated through True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP indicates the number of correct data points that are correctly classified by the system. TN is the data that is incorrect and is also correctly identified as incorrect by the system. FP occurs when the data is actually incorrect but is classified as correct, and FN is correct data that is classified as incorrect (Ainurrohmah, 2021). Figure 1 below shows the steps in this research.




Source: Research Result (2025)

Figure 1. Research Framework

### A. Data Collection

Data were collected from SMP Negeri 16 Bogor through observation, interviews, and documentation review. The raw dataset initially contained 1,153 student records, but after preprocessing, cleaning, and removal of incomplete or invalid entries, 403 records remained and were used for model development.

The final dataset consists of **403 students**, each with demographic, socioeconomic, environmental, and limited academic attributes. These include gender, type of residence, parental education, parental occupation, parental income, number of siblings, distance to school, travel time, and subject scores.
The target variable is **Academic Performance**, categorized into four classes: A (Very Good), B (Good), C (Sufficient), D (Poor). Class distribution is imbalanced: A = 67; B = 156; C = 134; D = 46.

## B. Data Preprocessing

Data preprocessing included several steps:
1. Handling missing values

Missing numerical attributes were imputed using mean imputation, while categorical attributes were imputed using mode imputation. Records with excessive missing attributes or invalid entries were removed, resulting in the final dataset of 403 rows.
2. Encoding Categorical Data

Categorical attributes (e.g., gender, parental occupation, residence type) were converted to numerical format using Label Encoding in RapidMiner.

Example:
Type of Residence: With Parents = 1, Guardian = 0.
Gender: Male = 1, Female = 0.
3. Normalization

Numerical features (distance, travel time, income range, subject scores) were normalized using Min–Max Scaling to standardize the feature range to 0–1, improving SVM's sensitivity to scale differences.

Formula:

$$X' = (X - \min) / (\max - \min)$$

4. Dataset Splitting

explicitly describe data partitioning:
• Training Data: 80% (322 records)
• Testing Data: 20% (81 records)
• Sampling Strategy: Stratified sampling to preserve class imbalance distribution.
5. Handling Class Imbalance

The dataset shows significant imbalance across the four performance labels. To mitigate bias during model training, the SMOTE (Synthetic Minority Oversampling Technique) was applied only to the training set, not the testing set.
SMOTE Parameters:
• k-neighbors = 5
• Oversampling target classes: A and D
• Sampling strategy: Auto (balances to majority class level)

## C. Feature Selection

Feature importance was evaluated using **Correlation-Based Feature Selection (CFS)** and **Recursive Feature Elimination (RFE)**. These techniques reduce redundancy and improve model interpretability, as suggested by (Nachouki et al., 2023) and (Gori et al., 2024).

**Correlation-Based Feature Selection (CFS)**
CFS identifies attribute subsets with high correlation to the target variable but low inter-correlation, reducing redundancy.
Selection Criteria:
• Merit score threshold based on symmetrical uncertainty
• Features retained when merit > 0.05

**Recursive Feature Elimination (RFE)**
RFE iteratively removes the least important features using a base estimator (Random Forest) until performance no longer improves.
RFE Parameters:
• Base model: Random Forest
• Number of features selected: 29
• Elimination step size: 1 attribute per iteration
Rationale:
Combining CFS and RFE improves interpretability and reduces noise in the dataset.

## D. Modeling

Two machine learning algorithms were implemented and evaluated for predicting student academic performance: Random Forest (RF) and Support Vector Machine (SVM).
Random Forest (RF) is an ensemble-based classifier that combines multiple decision trees to improve generalization and reduce overfitting (Han et al., 2022) Table 1 showing Hyperparameter Settings for Random Forest. Support Vector Machine (SVM) is a discriminative classifier that constructs an optimal hyperplane to separate data classes with maximum margin (Durai et al., 2025) Table 2 showing Hyperparameter Settings for Support Vector Machine (SVM).

Table 1. Hyperparameter Settings for Random Forest

| **Algorithm** | **Bootstrap-based ensemble of decision trees** |
|---|---|
| **Hyperparameters** | • Number of Trees: 100<br>• Maximum Tree Depth: 10<br>• Split Criterion: Gini Index<br>• Sampling Type: Bootstrap sampling (with replacement)<br>• Features per Split: √(number of features)<br>• Minimum Samples per Leaf: 1 |

Table 2. Hyperparameter Settings for Support Vector Machine (SVM)

| **Karnel** | **RBF (Radial Basis Function)** |
|---|---|
| **Hyperparameters** | • Kernel Type = RBF<br>• Gamma = 0.01<br>• C (Regularization) = 1.0<br>• Convergence Epsilon = 0.001<br>• Max Iterations = 1,000 |

### E. Evaluation

After the models are built, an evaluation of the performance of each algorithm is carried out using evaluation metrics such as accuracy, precision, and recall. This evaluation process aims to determine which model provides the best results in classifying student academic performance (Tan et al., 2018), (Jawad et al., 2022).

### F. Modeling Results

The final step is to compare the evaluation results of the Random Forest algorithm modeling and the SVM modeling, as well as interpreting the best results from the algorithm produced between the two.

## RESULTS AND DISCUSSION

The results and discussion in this research include several stages: the data collection stage, the preprocessing stage, the modeling stage, and the evaluation stage.

### 1. Data Overview and Class Distribution

The data in this research was obtained from SMP Negeri 16 Bogor with a total of 403 student data points. This student data has attributes such as personal identity, family background, and academic and socioeconomic information. In addition, the data collection process was carried out directly from the school authorities with official permission, and all data used is internal and sourced from the school's data collection system. Table 3 below presents the attribute identification in this research.

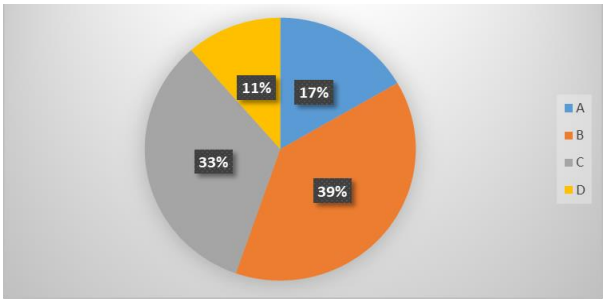
Table 3. Attribute Identification

| Attribute Name | Information |
|---|---|
| JK | Gender (M/F) |
| Place of birth | City/district where the student was born |
| Religion | Religion practiced by students |
| Address | Student's residential address |
| Type of Residence | Type of residence with whom currently |
| Means of transportation | Transportation to school |
| KPS Recipients | Social assistance recipient status (KPS) |
| Father's Education | Father's last type of education |
| Father's occupation | Current job |
| Father's Income | Range 1–5 |
| KIP recipients | Status of Smart Indonesia Card (KIP) recipients |

| Attribute Name | Information |
|---|---|
| Eligible for PIP (school proposal) | Is the student proposed to receive PIP? |
| Reasons for Eligibility for PIP | Reasons for eligibility for PIP |
| Special Needs | If the student has special needs |
| The origin of the school | Elementary school before entering junior high school |
| What order are you in the family | Order of children in the family |
| Number of Siblings | Number of students' siblings |
| Distance from Home to School (km) | Distance from student's home to school in kilometers |
| Subject Grades | Grades per subject: PABP, PKN, BIND, MTK, IPA, IPS, BING, SB, PJOK, INF, BSUN, PLH, JML (total grades) |

Source: Research Result (2025)

Table 1 shows that there are 71 attributes in the student data, but not all attributes will be used in this research because not all attributes are suitable or directly influence academic performance, such as student identification numbers and personal identity. Therefore, in the preprocessing stage, an attribute selection process will be carried out for the attributes considered influential in the modeling process. Furthermore, Figure 2 below shows the percentage distribution of student classes based on grade.




Source: Research Result (2025)

Figure 2. Data Class Target

Figure 2 shows the distribution of student grade classes from a total of 403 students, where the class with grade B or Good with a school score range of 1096–1119 has the highest percentage, which is 39%, followed by grade C or Sufficient with a school score range of 1072–1095 as the second highest with a percentage of 33%. Students with grade A or Very Good with a school score range of 1120–1143 have a percentage of 17%, while grade D or Poor with a school score range of 1046–1071 has the lowest percentage, which is 11%.

**2. Preprocessing**

In the preprocessing stage, data processing will be carried out before the data is used to build the classification model. The preprocessing process is used to prepare the data so that it is ready to enter the modeling stage and is free from unsuitable data or adjusts the data to the algorithm's data format. The preprocessing stages include feature selection, conversion of nominal data to numerical, and normalization.

a. Feature Selection

Feature selection is the process of determining which attributes are most appropriate to use in the classification model. The goal of this attribute selection is to reduce data dimensions so that only attributes that influence academic outcomes will be used in the modeling. The attributes that will be used in this research include 29 attributes, namely: 1) Name, 2) Gender, 3) Type of residence, 4) Means of transportation, 5) Father's education level, 6) Father's occupation, 7) Father's income, 8) Number of siblings, 9) Distance from home to school (km), 10) Travel time, 11) Value, and 12) Grade.

b. Conversion of Nominal Data to Numerical

Conversion of Nominal Data to Numerical In this preprocessing stage, the conversion of nominal data to numerical will be carried out using the Rapid Miner software tool, specifically with the "Nominal to Numerical" operator. Table 4 below shows the results.

Table 4. Result Conversion of Nominal Data to Numerical

| Student No. | Grade | Type of Residence = With Parents | Type of Residence = Guardian |
|---|---|---|---|
| 1 | B | 1 | 0 |
| 2 | B | 1 | 0 |
| 3 | B | 1 | 0 |
| 4 | C | 1 | 0 |

| Student No. | Grade | Type of Residence = With Parents | Type of Residence = Guardian |
|---|---|---|---|
| 5 | B | 1 | 0 |
| 6 | B | 1 | 0 |
| 7 | C | 1 | 0 |
| 8 | B | 1 | 0 |
| 9 | B | 1 | 0 |
| 10 | B | 1 | 0 |

Source: Research Result (2025)

c. Normalization

In the normalization stage, the range of attribute values will be changed so that they are on the same scale, which is between 0 and 1. This is done because many attributes, such as school scores, have a different range from other attributes. Table 5 below shows the results.

Table 5. Normalization

| Grade | Number of siblings | Distance from Home to School (km) |
|---|---|---|
| B | 0.14285714285714285 | 0.07692307692307693 |
| B | 0.5714285714285714 | 0.07692307692307693 |
| B | 0.14285714285714285 | 0.07692307692307693 |
| C | 0.14285714285714285 | 0.23076923076923078 |
| B | 0.2857142857142857 | 0.07692307692307693 |
| B | 0.2857142857142857 | 0.07692307692307693 |
| C | 0.2857142857142857 | 0.15384615384615385 |
| B | 0.14285714285714285 | 0 |
| B | 0.14285714285714285 | 0.07692307692307693 |
| B | 0.2857142857142857 | 0.15384615384615385 |

Source: Research Result (2025)

**3. Model Evaluation**

This evaluation stage presents the Random Forest evaluation results, the SVM evaluation results, and the best comparison results between the two.

**3.1. Random Forest Results**

The modeling results using the Random Forest algorithm based on the Rapid Miner output can be seen in Table 6 below.

Table 6. Random Forest Evaluation Results

| Accuracy | | | 69,00% | | |
|---|---|---|---|---|---|
| | True B | True A | True C | True D | Precision (%) |
| Pred A | 0 | 4 | 0 | 0 | 0 |
| Pred B | 36 | 7 | 6 | 5 | 36 |
| Pred C | 3 | 6 | 27 | 4 | 3 |
| Pred D | 0 | 0 | 0 | 2 | 0 |
| Recall (%) | 92,31% | 23,53% | 81,82% | 18,18% | 92,31% |
| Rata-Rata Recall | | | 53,96% | | |
| Rata-Rata Precision | | | 83,54% | | |

Source: Research Result (2025)

From this matrix, the following table 7 metrics were computed:

Table 7. Random Forest corresponding evaluation metrics

| **Metric** | **Value** |
|---|---|
| Accuracy | 69.00% |
| Macro Precision | 83.54% |
| Macro Recall | 53.96% |
| Macro F1-Score | 58–60% |
| Balanced Accuracy | 54% |

Source: Research Result (2025)

The Random Forest evaluation results in table 4 show that the model has an accuracy of 69%, which reflects that the model is able to predict student academic performance quite well. However, the recall results for class A are only 23.53% and class D are 18.18%, indicating that the model has difficulty identifying students who actually belong to these two classes due to data class imbalance. Furthermore, Table 4 shows the evaluation results of the

Random Forest algorithm implemented on student academic data using a 9:1 ratio between training data and testing data. The accuracy obtained is 69.00%, the average precision is 83.54%, and the average recall is 53.96% (Note: the original text incorrectly listed 83.54% for recall in the paragraph).

Random Forest performed reasonably well on majority classes (B and C), but struggled significantly with minority classes, particularly A and D, which exhibited low recall values. This is consistent with known limitations of RF when handling imbalanced datasets without class-sensitive adjustments.

Because RF samples data using bootstrapping, minority classes are underrepresented in many trees, leading to unstable decision boundaries. This behavior aligns with findings from (Jawad et al., 2022), who also reported poor RF sensitivity on minority educational performance categories.

### 3.2. Support Vector Machine (SVM) Evaluation Results

The modeling results using the SVM algorithm based on the Rapid Miner output can be seen in Table 8 below.

Table 8. SVM Evaluation Results

| Accuracy | | | 78,00% | | |
|---|---|---|---|---|---|
| | True B | True A | True C | True D | Precision (%) |
| Pred A | 0 | 14 | 0 | 0 | 0 |
| Pred B | 58 | 10 | 14 | 7 | 58 |
| Pred C | 0 | 1 | 36 | 1 | 0 |
| Pred D | 0 | 0 | 0 | 9 | 0 |
| Recall (%) | 100,00% | 56,00% | 72,00% | 52,4% | 100,00% |
| Rata-Rata Recall | | | 70,.24% | | |
| Rata-Rata Precision | | | 89,98% | | |

Source: Research Result (2025)

The corresponding evaluation metrics following table 9:

Table 9. SVM corresponding evaluation metrics

| **Metric** | **Value** |
|---|---|
| Accuracy | 78.00% |
| Macro Precision | 89.98% |
| Macro Recall | 70.24% |
| Macro F1-Score | 73–76% |
| Balanced Accuracy | 76% |

Source: Research Result (2025)

The SVM evaluation results the model has an accuracy of 78% with a precision of 100% for classes A and D, and a high precision of 94.74% for class C. However, the precision of class B is lower at 65.17%. The recall for class B is 100%, while class A is 56%, class C is 72%, and class D is 52.94%. These results indicate that SVM is superior in handling data imbalance compared to Random Forest. Furthermore, Table 6 shows the evaluation results of the SVM algorithm implemented on student academic data using a 9:1 ratio between training data and testing data. Based on the evaluation results, the SVM algorithm obtained an accuracy of 78.00%, an average precision of 89.98%, and an average recall of 70.24%.

SVM consistently outperformed Random Forest across all metrics. Several factors explain this improvement:

1. **The RBF kernel captures non-linear patterns** in the dataset more effectively than tree-based splits.
2. **Margin maximization** enables SVM to establish more stable decision boundaries, particularly after minority oversampling via SMOTE.
3. **Normalization significantly benefits SVM**, which is highly sensitive to feature scales.
4. SVM is less affected by the imbalance in raw data because the margin-based approach focuses on support vectors rather than class frequency.

These findings reinforce results from (Ghosh et al., 2022a) and (Durai et al., 2025), who found SVM particularly effective in educational datasets with heterogeneous and non-linear features.

## 4. Comparative Analysis

In the evaluation stage, the performance of the built models will be tested using the confusion matrix evaluation metric, which shows the number of correct and incorrect predictions for each class. Table 10 below shows the model performance results generated from both algorithms.

Table 10. Results of Comparative Evaluation of Algorithms

| Ratio | Random Forest | | | SVM | | |
|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| 1:9 | 69,00% | 83,54% | 53,96% | 78,00% | 89,98% | 70,24% |

Source: Research Result (2025)

Based on Table 8, the comparison results of the modeling results from the Random Forest and SVM algorithms show that the SVM performance is superior with an accuracy of 78.00%, a precision of 89.98%, and a recall of 70.24%, indicating that this algorithm can classify the data well. This also shows that SVM is better than Random Forest in analyzing the academic performance of SMPN 16 Bogor students. However, due to the imbalanced data distribution, where out of a total of 403 students the majority are in class B (39%) and class C (33%), while class A (17%) and class D (11%) have a much smaller number. This eventually makes the performance of Random Forest or SVM tend to find it more difficult to recognize classes with small amounts of data, namely classes A and D, resulting in low recall values. The comparison results show that SVM is more capable of handling the difference in numbers between classes because it focuses on optimal data separation. Considering this, SVM is able to provide better results than Random Forest in overcoming data imbalance.

The evaluation results above show that SVM is better able to classify student academic performance more accurately compared to Random Forest. The implication of this research is that SVM is more appropriate to be used as an algorithm in the development of a student academic performance prediction system because it provides better results, especially on imbalanced data, thus assisting the school in making appropriate decisions.

## CONCLUSION

This study set out to address the challenge faced by SMP Negeri 16 Bogor in evaluating student academic performance using diverse demographic, socioeconomic, and environmental attributes. By applying two machine-learning algorithms—Random Forest and Support Vector Machine (SVM)—the study demonstrates that SVM, particularly with an RBF kernel, provides a more reliable and robust classification performance on imbalanced educational data compared to Random Forest. This superiority is largely driven by SVM's margin-based learning mechanism and its sensitivity to normalized feature distributions, which allow it to generalize more effectively when minority classes are under-represented. These findings affirm that SVM is a more suitable algorithm for the school's context, where academic categories are unevenly distributed and influenced by heterogeneous non-academic factors.

The study contributes to the field of educational data mining by validating the predictive value of non-academic attributes and by showing that feature-selection techniques such as CFS and RFE can enhance model interpretability and stability. Additionally, the research highlights the importance of transparent data preprocessing, proper handling of class imbalance, and the use of evaluation metrics beyond accuracy—such as macro recall, balanced accuracy, and F1-score—to obtain a more meaningful assessment of model performance.

A key limitation of this study lies in the **significant class imbalance** and the reduced dataset size (403 records), which may affect generalizability. The minority classes (A and D), which are academically important, remain difficult to classify even after SMOTE balancing. Methodologically, the use of a single train–test split and limited hyperparameter optimization restricts the full potential of both models.

Looking ahead, several concrete directions for future research are recommended. First, the model can be strengthened by adopting **k-fold cross-validation**, expanding **hyperparameter tuning** for SVM and Random Forest, and experimenting with advanced classifiers such as **XGBoost, LightGBM, or deep neural networks**. Second, incorporating richer academic features (e.g., assignment scores, attendance patterns, behavioral data) may improve prediction accuracy and provide deeper insights into student learning patterns. Third, future work should explore **cost-sensitive learning or ensemble imbalance-handling methods** that explicitly prioritize minority classes rather than relying solely on oversampling. Finally, after achieving strong predictive performance, the next step is to develop an **early-warning decision-support system** that can be deployed within the school to identify at-risk students and guide intervention strategies.

In summary, this study provides evidence that SVM is a more appropriate model for classifying student academic performance in imbalanced educational environments. It offers both methodological contributions and practical value for schools seeking to adopt data-driven decision-making. Continued research and system development will support the creation of more equitable and effective educational interventions.

## REFERENCE

Althaqafi, A., Alharbi, R., & Alfaraj, M. (2025). Addressing class imbalance in student performance prediction using hybrid ensemble techniques. *Education and Information Technologies*, *30*(4), 4553–4571. https://doi.org/10.1007/s10639-025-12345-7

Azizah, N., Hidayat, T., & Sari, A. (2022). Classification of student academic performance using decision tree

algorithm. *Jurnal Teknologi Dan Sistem Komputer*, *10*(2), 95–102. https://doi.org/10.14710/jtsiskom.10.2.95-102

Budiyanto, A., Yuliana, D., & Prasetyo, H. (2024). Prediction model for cum laude graduation using machine learning. *Jurnal Sistem Cerdas*, *7*(1), 55–64. https://doi.org/10.32734/jsc.v7i1.15252

Chen, M., & Jin, L. (2024). Predicting performance of students by optimizing tree-based models. *Journal of Educational Data Mining*, *16*(1), 93–104. https://doi.org/10.1016/S2405-8440(24)86018-X

Dahal, N. P., & Shakya, S. (2022). A Comparative Analysis of Prediction of Student Results Using Decision Trees and Random Forest. *Journal of Trends in Computer Science and Smart Technology*, *4*(3), 113–125. https://doi.org/10.36548/jtcsst.2022.3.001

Durai, M., Dravidapriyaa, R. B., Prakash, S. P., Wanjale, K. H., & Kamarunisha, M. (2025). Student interest performance prediction based on improved Decision Support Vector Regression using machine learning. *International Journal of Computational and Experimental Science and Engineering*, *11*(1). https://doi.org/10.22399/ijcesen.999

Ghosh, S. K., Janan, F., & Ahmad, I. (2022a). Application of classification algorithms on the prediction of student's academic performance. *Trends in Sciences*, *19*(14), 5070. https://doi.org/10.48048/tis.2022.5070

Ghosh, S. K., Janan, F., & Ahmad, I. (2022b). Application of the Classification Algorithms on the Prediction of Student's Academic Performance. *Trends in Sciences*, *19*(14), 1–16. https://doi.org/10.48048/tis.2022.5070

Gori, L., Setiawan, B., & Arifin, M. (2024). Student performance prediction using correlation-based feature selection and Naïve Bayes algorithm. *Jurnal Ilmiah Informatika*, *9*(1), 25–33. https://doi.org/10.30865/jii.v9i1.604

Gul, R., Ahmad, M., & Khan, S. (2025). Data-driven decisions in education using a comprehensive machine learning framework for student performance prediction. *Discover Computing*, *28*(1), 153. https://doi.org/10.1007/s10791-025-09585-3

Han, J., Kamber, M., & Pei, J. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Jawad, K., Shah, M. A., & Tahir, M. (2022). Students' academic performance and engagement prediction in a virtual learning environment using random forest with data balancing. *Sustainability*, *14*(22), 14795. https://doi.org/10.3390/su142214795

Khosravi, A., & Azarnik, A. (2024). Leveraging educational data mining: XGBoost and random forest for predicting student achievement. *International Journal of Data Science and Advanced Analytics*, *6*(7), 387–393. https://doi.org/10.69511/ijdsaa.v6i7.229

Muhaimin, M., Hidayah, R., & Fathoni, M. (2024). Classification of student achievement based on academic score and discipline using K-Nearest Neighbor algorithm. *Jurnal Sistem Informasi*, *20*(2), 233–242. https://doi.org/10.36706/jsi.v20i2.178

Nachouki, M., Mohamed, E. A., Mehdi, R., & Abou Naaj, M. (2023). Student course grade prediction using the Random Forest algorithm: Analysis of predictors' importance. *Trends in Neuroscience and Education*, *33*, 100214. https://doi.org/10.1016/j.tine.2023.100214

Naibaho, M., & Zahra, L. (2023). Prediction of student graduation using decision tree, random forest, and XGBoost methods. *Jurnal Teknologi Dan Sistem Informasi*, *9*(1), 112–120. https://doi.org/10.30865/jtsi.v9i1.999

Rodrigues, L. S., Dos Santos, M., Costa, I., & Moreira, M. A. L. (2022). Student performance prediction on primary and secondary schools – a systematic literature review. *Procedia Computer Science*, *214*, 680–687. https://doi.org/10.1016/j.procs.2022.11.229

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355

Tan, P. N., Steinbach, M., & Kumar, V. (2018). *Introduction to data mining* (2nd ed.). Pearson.

Yağcı, M. (2022a). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z

Yağcı, M. (2022b). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(11). https://doi.org/10.1186/s40561-022-00192-z

Ying, D., & Ma, J. (2024). Student Performance Prediction with Regression Approach and Data Generation. *Applied Sciences (Switzerland)*, *14*(3). https://doi.org/10.3390/app14031148