Available online at: https://journal.upgripnk.ac.id/index.php/saintek
Jurnal Pendidikan Informatika dan Sains
15(1), 2026, 77-87
E-ISSN: 2407-1536
P-ISSN: 2089-2802

PERFORMANCE COMPARISON OF NAIVE BAYES, PSO-NAIVE BAYES, AND PCA-PSO-NAIVE BAYES FOR THE CLASSIFICATION OF ELEMENTARY SCHOOL STUDENTS' INTERESTS

Endah Herminarimawati*, Kusrini
Online Master's Program in Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
*email: endahhermi@gmail.com

Received: 2026-04-07 Accepted: 2026-06-10 Published: 2026-06-30

Abstract

Identifying students' interests at an early stage is a crucial process in education because it enables schools to direct students' potential in a structured and objective manner. However, report card data typically contain many attributes, some of which are redundant or irrelevant, and this condition can degrade the performance of classification algorithms such as Naive Bayes. This study compared the classification performance of three models, namely pure Naive Bayes, a hybrid PSO-Naive Bayes model, and a hybrid PCA-PSO-Naive Bayes model, in classifying the interests of elementary school students into four categories: Academic, Arts, Sports, and ICT. The dataset combined cognitive data in the form of report card scores in nine subjects with affective data obtained from an interest questionnaire. Particle Swarm Optimization (PSO) was applied as a feature selection technique, while Principal Component Analysis (PCA) was applied as a dimensionality reduction technique before feature optimization. Model performance was evaluated using stratified 5-fold cross validation. The results showed that the PCA-PSO-Naive Bayes model achieved the highest mean accuracy of 98.92%, compared with 97.84% for pure Naive Bayes and 97.30% for PSO-Naive Bayes. The hybrid PCA-PSO-Naive Bayes model also produced the most stable accuracy across folds, with a range of 97.3% to 100%. These findings indicate that combining PCA-based dimensionality reduction with PSO-based component selection improves both the accuracy and the stability of Naive Bayes in classifying student interests based on academic and questionnaire data.

Keywords: Naive Bayes, Principal Component Analysis, Particle Swarm Optimization, classification, student interest.

How to cite (in APA style): Herminarimawati, E., & Kusrini, K. (2026). Performance comparison of Naive Bayes, PSO-Naive Bayes, and PCA-PSO-Naive Bayes for the classification of elementary school students' interests. *Jurnal Pendidikan Informatika Dan Sains, 15*(1), 77–87. https://doi.org/10.31571/saintek.v15i1.10914

DOI: 10.31571/saintek.v15i1.10914

INTRODUCTION

The identification of students' interests in schools is still commonly carried out through conventional approaches such as subjective teacher observation. In practice, this approach is time consuming, costly, unstructured, and prone to producing inaccurate mappings of students' actual potential at an early stage. At the same time, schools hold abundant data assets in the form of report card scores and questionnaire results that have not been used optimally to support decision making

related to student interests. Interest is an essential element in determining the direction of a child's potential development from the elementary school level, so an objective and data-driven identification mechanism is needed.

Educational data mining offers a solution for extracting knowledge and patterns from historical academic data to support such decisions. Digital records of learning behavior and achievement have been shown to be a rich basis for predicting student outcomes when appropriate modeling standards are applied (Arizmendi et al., 2023; Yağcı, 2022). Through classification techniques, the mapping of student interests can be performed automatically, quickly, and objectively by integrating cognitive data in the form of report card scores with affective data obtained from questionnaires. Recent work in educational data mining has also demonstrated that combining machine learning models with metaheuristic optimization can substantially improve the quality of student performance prediction systems (Abukader et al., 2025). Behavioral and affective attributes have also been shown to add substantial predictive value beyond academic scores alone (Amrieh et al., 2016).

The Naive Bayes algorithm is widely used in educational classification tasks because it is simple, computationally efficient, and effective in handling probabilistic reasoning, and it has repeatedly outperformed more complex classifiers on academic datasets (Mueen et al., 2016). A previous study classified the potential of elementary school children based on hobbies using Naive Bayes with an undersampling technique, but it involved only 50 students, used three interest categories, and did not compare the method with any optimization approach (Karolina et al., 2025). Naive Bayes also has a fundamental weakness, namely the assumption of feature independence, in which every attribute is considered to contribute equally. The algorithm is sensitive to imbalanced data, is less suitable when features are interdependent, and its accuracy can decrease when irrelevant or redundant features are present. Most previous studies on interest classification stopped at standard classification using all available attributes without considering feature relevance, so the resulting performance was not optimal, even though feature selection has been shown to raise the accuracy of student performance classifiers considerably (Nayak et al., 2023). In the Indonesian context, Naive Bayes has also been combined with forward selection to predict students' study and career interests, with feature selection identifying the attributes that most influence interest (Kusuma et al., 2025), and machine learning has been applied to classify elementary students' interests and talents (Sartika et al., 2024).

Feature selection and dimensionality reduction are established strategies for addressing these weaknesses. Feature selection methods aim to retain only the most informative attributes and have been shown to improve classification accuracy across a wide range of data conditions, including incomplete information systems (Kamalov et al., 2025; Kumar & Gupta, 2026). Wrapper approaches that embed the classifier in the search, such as hybrid PSO with local search, have proven particularly effective for selecting compact and discriminative feature subsets (Moradi & Gholampour, 2016), and self-adaptive PSO variants extend this capability to large-scale problems (Xue et al., 2019). However, the interpretation of feature selection results is not trivial, and iterative approaches are often required to obtain stable and meaningful feature subsets (Lötsch, 2026). Particle Swarm Optimization (PSO) is one of the most widely adopted metaheuristics for feature selection because of its simple mechanism and fast convergence, and its use has grown consistently across application domains over the last three decades (Salgotra et al., 2026). PSO has been applied successfully to optimization problems in various fields, such as task scheduling in fog computing systems (Monika, 2025) and multimetric route optimization in vehicular networks (Pathan & Annapurna, 2025), and hybrid PSO schemes have proven effective in complex search spaces (Fister et al., 2025). More specifically, PSO-based feature selection has repeatedly improved Naive Bayes accuracy: by about 6% in student graduation classification (Purnamasari et al., 2020), by more than 17% in sentiment analysis compared with a genetic algorithm (Rafdi et al., 2021), and in intrusion detection on high-dimensional data (Talita et al., 2021). A hybrid of correlation-based feature selection and improved binary PSO

with Naive Bayes even reached perfect accuracy on several benchmark datasets while controlling premature convergence to local optima (Jain et al., 2017).

Principal Component Analysis (PCA) complements feature selection by transforming correlated attributes into a smaller set of orthogonal components that retain most of the data variance. Comparative studies show that PCA can be used effectively either as feature extraction or as a basis for feature selection to improve classifier performance (Sari & Wijaya, 2025; Omuya et al., 2021), and comparative studies on high-dimensional data show that PCA generally outperforms alternative reduction techniques when combined with standard classifiers (Reddy et al., 2020), and PCA has also been applied to summarize complex behavioral data of school-age children (González-Devesa, 2026). Applying PCA before Naive Bayes has been reported to increase classification accuracy on benchmark data (Syafira et al., 2020). Integrating PCA to simplify the data and PSO to select the most relevant components before Naive Bayes classification is therefore expected to eliminate unnecessary variables and focus the model on the most informative information contained in report card scores and questionnaires.

This study lies in three aspects, the integration of cognitive data (report cards) and affective data (questionnaires), a multi-domain approach to student interest covering four categories (Academic, Arts, Sports, and ICT), and the sequential combination of PCA for dimensionality reduction and PSO for feature selection on top of Naive Bayes. This study aims to compare the performance of pure Naive Bayes, PSO-Naive Bayes, and PCA-PSO-Naive Bayes in classifying elementary school students' interests and to determine whether the hybrid PCA-PSO-Naive Bayes model produces higher and more stable accuracy.

METHOD

This study used a quantitative experimental method with a data mining approach to compare the effectiveness of the hybrid PSO-Naive Bayes model with the optimized PCA-PSO-Naive Bayes model in classifying the interests of elementary school students. The research procedure consisted of five stages: data collection, data preprocessing, model 1 optimization (PSO + Naive Bayes), model 2 optimization (PCA + PSO + Naive Bayes), and analysis of the results. The research workflow is presented in Figure 1.

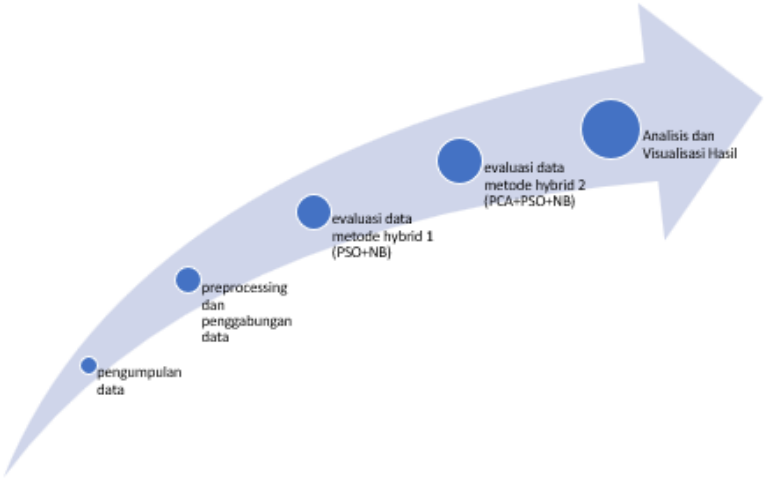

Figure 1. Research Workflow

The data were obtained from students' report cards and an interest questionnaire. The report card data consisted of numeric scores in nine subjects, namely Religious Education, Pancasila, Indonesian Language, Mathematics, Natural Sciences, Physical Education, Arts and Culture, English, and Information and Communication Technology (ICT), which describe students' cognitive abilities. The questionnaire was used to capture affective information about student interests that is not

reflected in report card scores; the qualitative responses were converted into numeric scales. The classification target consisted of four interest categories: Academic, Arts, Sports, and ICT. The baseline Naive Bayes classifier used the combined report card and questionnaire data, with the class prior probability computed using Equation (1).

$$P(Ci) = Si / s \quad (1)$$

where P(Ci) is the prior probability of class i, Si is the number of samples belonging to class i, and s is the total number of samples.

In model 1 (PSO + Naive Bayes), PSO was used to select the best combination of features. Each particle was represented by a binary array, where a value of 0 means the feature is discarded and a value of 1 means the feature is used. The training data were reduced to the features encoded as 1, a Gaussian Naive Bayes model was trained on those features, and the classification accuracy was used to compute the cost function, so that higher accuracy corresponds to lower cost. PSO then searched for the particle position with the minimum cost. The velocity and position of each particle were updated using Equation (2).

$$Vi(t) = \omega Vi(t-1) + C1A(Pbest - Xi(t-1)) + C2B(Gbest - Xi(t-1)) \quad (2)$$
$$Xi(t) = Vi(t) + Xi(t-1)$$

where Vi(t) is the newly computed velocity, ω is the inertia weight that balances exploration and exploitation, C1 is the individual acceleration coefficient with A a random number in [0, 1] forming the cognitive component, C2 is the social acceleration coefficient with B a random number in [0, 1] forming the social component, Pbest is the best position found by the particle, Gbest is the best position found by the swarm, and Xi(t) is the new position of the particle at the current iteration.

Model 2 (PCA + PSO + Naive Bayes) is similar to model 1 but adds a dimensionality reduction stage using PCA before feature optimization by PSO. The stages are as follows. First, data standardization: because PCA is highly sensitive to data scale, the data were scaled using a standard scaler so that each attribute has a mean of 0 and a variance of 1. Second, dimensionality reduction: the original features were transformed into at most eight new orthogonal principal components that retain the largest portion of the data variance. Third, component optimization by PSO: analogous to model 1, PSO selected the PCA components that contribute most to Naive Bayes accuracy. Fourth, final testing: the best PCA components selected by PSO were used to train and test the Naive Bayes model, and the results were compared with model 1.

Model performance was evaluated using stratified 5-fold cross validation so that every sample had an equal opportunity to serve as training and testing data and the class proportions were preserved in each fold. This procedure was chosen to assess how well the models generalize to independent data that they have never seen. Convergence curves of the PSO cost function were also analyzed to observe how quickly and how well the optimization found the best solution in each iteration.

RESULTS AND DISCUSSION

This study classified students' interests based on the combination of subject scores and questionnaire data using three classification models: Naive Bayes (model 1), PSO + Naive Bayes (model 2), and PCA + PSO + Naive Bayes (model 3). Model 1 used all available features, while models 2 and 3 used dominant features selected on the basis of research considerations, namely Indonesian Language, Mathematics, Natural Sciences, Physical Education, Arts, English, and ICT scores, together with the Academic, Arts, Sports, and ICT questionnaire indicators. The Religious Education and Pancasila subjects were eliminated from models 2 and 3.

Data preprocessing consisted of column selection, data cleaning, and merging of the report card and questionnaire datasets. Table 1 presents an example of the merged questionnaire data after

cleaning, and Table 2 presents an example of the combined feature set before the separation of features and target.

Table 1. Example of Merged Questionnaire Data after Column Selection and Cleaning

Student	Q_Acad_1	Q_Acad_2	Q_Acad_3	Q_Arts_1	Q_ICT_1
Student 1	3.0	2.0	3.0	1.0	1.0
Student 2	1.0	2.0	1.0	2.0	2.0
Student 3	3.0	3.0	3.0	1.0	1.0
Student 4	1.0	2.0	3.0	1.0	3.0
Student 5	3.0	2.0	3.0	2.0	2.0

Table 2. Example of Combined Report Card and Questionnaire Features

Student	Indonesian	Mathematics	Science	PE	ICT
Student 1	81.0	84.0	89.0	79.0	85.0
Student 2	72.0	69.0	70.0	73.0	70.0
Student 3	88.0	85.0	93.0	76.0	88.0
Student 4	93.0	90.0	95.0	81.0	92.0
Student 5	80.0	76.0	82.0	72.0	79.0

The next stage was label encoding, which converts the textual interest labels into numbers so that they can be processed by the machine learning algorithm. Table 3 presents the encoding scheme for the four target classes. Each encoded number represents the interest of one student, not a score or a rank, and these numbers form the target vector learned by Naive Bayes.

Table 3. Encoding of the Target Labels

Alphabetical Order	Label	Code
1	Academic	0
2	Sports	1
3	Arts	2
4	ICT	3

The PSO parameters used in models 2 and 3 are summarized in Table 4. These values follow common practice in binary PSO for feature selection, in which moderate acceleration coefficients and a relatively large inertia weight encourage broad exploration of the search space (Salgotra et al., 2026).

Table 4. PSO Parameters

Parameter	Value
Number of iterations	50
Cognitive coefficient (c1)	0.5
Social coefficient (c2)	0.5

Parameter	Value
Inertia weight (w)	0.9
Number of neighbors (k)	30
Minkowski p-norm (p)	2

The best feature combination found by PSO across the five folds was represented by the binary vector [1 1 0 0 1 1 0 1 1 1 0 1 1 1 1 1 1 0 1], with a training accuracy of 100%. The accuracy of each model in every fold of the stratified 5-fold cross validation is presented in Table 5.

Table 5. Accuracy of Each Model per Fold

Fold	Naive Bayes (%)	PSO + NB (%)	PCA + PSO + NB (%)
Fold 1	100.0	100.0	100.0
Fold 2	97.3	100.0	100.0
Fold 3	100.0	100.0	100.0
Fold 4	97.3	91.9	97.3
Fold 5	94.6	94.6	97.3

Based on these results, all models achieved high accuracy in almost all folds. The mean accuracy of each model, computed as the average of the five fold accuracies, is presented in Table 6.

Table 6. Mean Accuracy of the Three Models

Model	Mean Accuracy (%)
Naive Bayes	97.84
PSO + Naive Bayes	97.30
PCA + PSO + Naive Bayes	98.92

Model 3 produced the highest mean accuracy of 98.92%, while model 2 produced the lowest mean accuracy of 97.30%. Each model is analyzed in turn below.

The pure Naive Bayes model used all features in the dataset, including the Religious Education and Pancasila scores, and achieved a mean accuracy of 97.84%. In folds 1 and 3 the model obtained perfect accuracy of 100%, which indicates that the class probability distributions could be learned well by the algorithm. However, the accuracy decreased slightly in folds 4 and 5. This condition shows that using all features does not always produce the best performance, because some features may be less relevant to the identification of student interests. This finding is consistent with the literature on feature selection, which emphasizes that irrelevant or redundant attributes can degrade classifier performance (Kamalov et al., 2025).

In model 2, the Religious Education and Pancasila features were eliminated, and PSO was then applied to select the best features from the set of dominant features. Theoretically, PSO aims to find the most optimal feature combination to improve classification ability. Nevertheless, the results show that the mean accuracy of model 2 was 97.30%, slightly lower than pure Naive Bayes, and the decline was most visible in fold 4, which only reached 91.9%. This indicates that PSO-based feature selection does not always produce a better feature combination than using all features directly. The condition may be caused by the relatively limited amount of data, by an initial feature set that was already sufficiently representative, or by PSO becoming trapped in a local optimum, a known risk of swarm-

based metaheuristics in small search spaces (Fister et al., 2025; Salgotra et al., 2026). This result contrasts with studies in which PSO-based feature selection improved Naive Bayes performance on larger datasets (Jain et al., 2017; Purnamasari et al., 2020; Talita et al., 2021), which suggests that the benefit of wrapper-based selection depends on the dimensionality of the data and the amount of redundancy among features (Moradi & Gholampour, 2016). In addition, because the PSO fitness function was evaluated on the training data, the selected subset can be optimistically biased toward the training folds, which is a common interpretation challenge in feature selection studies (Lötsch, 2026).

Model 3 is a hybrid model that combines PCA, PSO, and Naive Bayes. The stages consisted of eliminating the Religious Education and Pancasila features, selecting the dominant features, standardizing the data, reducing the dimensionality with PCA, selecting the best components with PSO, and classifying with Naive Bayes. This model produced the highest mean accuracy of 98.92%. In folds 1, 2, and 3 the model achieved perfect accuracy of 100%, while in folds 4 and 5 the accuracy remained at 97.3%. These results show that the combination of PCA and PSO improved the quality of the data representation so that Naive Bayes could classify more accurately. Because the principal components produced by PCA are orthogonal, this transformation also mitigates the violation of the feature independence assumption that underlies Naive Bayes, which explains why the PCA-based model outperformed the model that applied PSO directly to the original correlated features (Sari & Wijaya, 2025; Syafira et al., 2020). Similar gains from combining PCA-based reduction with Naive Bayes have been reported across domains, from benchmark datasets to text and medical data (Omuya et al., 2021; Reddy et al., 2020).

The elimination of the Religious Education and Pancasila subjects in models 2 and 3 was based on the assumption that the identification of student interests is more strongly influenced by subjects that are directly related to the interest domains, namely Indonesian Language, Mathematics, Natural Sciences, Physical Education, Arts, English, and ICT. The results show that even though several features were eliminated, model performance remained high and even improved in model 3. This confirms that using more relevant features can increase classification effectiveness (Kumar & Gupta, 2026).

The stability of the models can be assessed from the range of accuracies across folds, as presented in Table 7.

Table 7. Stability of Model Accuracy

Model	Accuracy Range (%)
Naive Bayes	94.6 to 100
PSO + NB	91.9 to 100
PCA + PSO + NB	97.3 to 100

Based on these ranges, model 3 has the smallest variation in accuracy. A small variation indicates that the model has better generalization ability across different subsets of the data. The fold-by-fold accuracy of the three models is visualized in Figure 2.

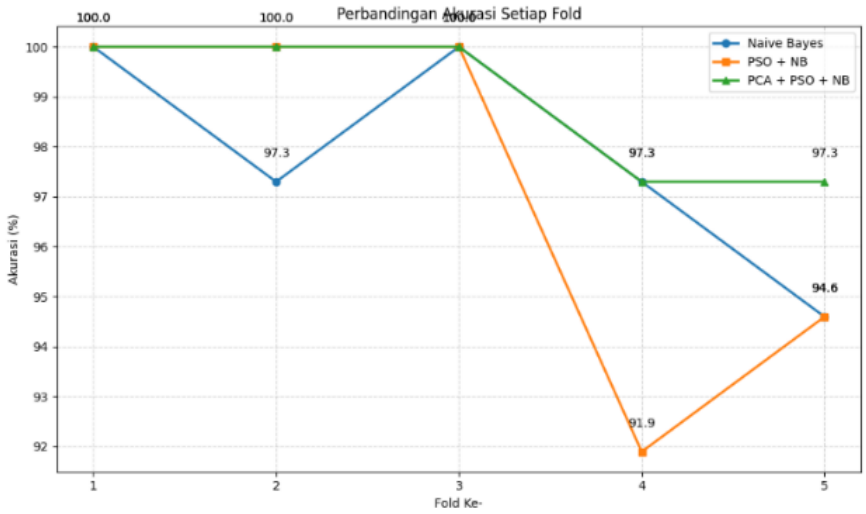

Figure 2. Accuracy of Each Model per Fold

Figure 2 shows that pure Naive Bayes (circles) fluctuated across folds, PSO + NB (squares) was the least stable of the three methods, and PCA + PSO + NB (triangles) was the most stable method with the highest overall performance. The comparison of the mean accuracy of the three models is summarized in Figure 3.

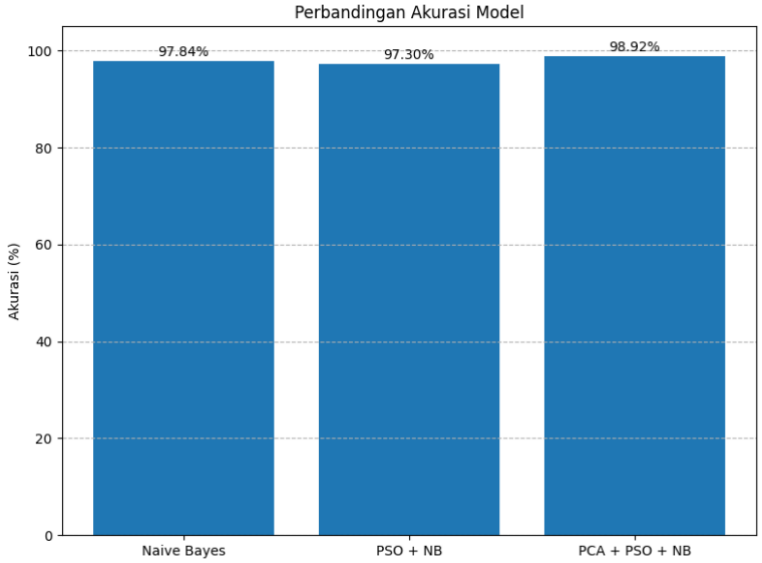

Figure 3. Comparison of Naive Bayes, PSO + NB, and PCA + PSO + NB

Compared with previous work, this study extends the Naive Bayes-based classification of elementary students' potential reported by Karolina et al. (2025), which used only 50 students, three interest categories, and no optimization, by adding a fourth category (ICT), integrating cognitive and affective data sources, and comparing two optimization pipelines. The finding that a sequential PCA and metaheuristic pipeline improves classification performance is also in line with recent educational data mining studies that combine metaheuristic optimization with machine learning models (Abukader et al., 2025) and with feature-selection-based approaches to predicting students' study interests (Kusuma et al., 2025; Nayak et al., 2023).

This study has several limitations. The dataset came from a single school with a limited number of students, so the very high accuracies obtained should be interpreted with caution and validated on larger and more heterogeneous datasets. The evaluation was limited to accuracy; future work should

report precision, recall, and F1-score per class, especially if the class distribution is imbalanced. Finally, the PSO fitness evaluation on training data could be replaced by an internal validation scheme to reduce the risk of overfitting during feature selection.

CONCLUSION

Based on the evaluation using stratified 5-fold cross validation, all three models showed very good classification ability in identifying elementary school students' interests. The PCA + PSO + Naive Bayes model provided the best performance with a mean accuracy of 98.92% and showed better stability than the other models, with fold accuracies ranging from 97.3% to 100%. These results indicate that the combination of PCA-based dimensionality reduction, PSO-based component selection, and Naive Bayes classification improves the effectiveness of student interest identification based on academic and questionnaire data. The contribution of this study lies in the integration of cognitive (report card) and affective (questionnaire) data, the multi-domain treatment of student interest in four categories (Academic, Arts, Sports, and ICT), and the sequential PCA + PSO + Naive Bayes pipeline, which extends previous work that used only pure Naive Bayes with fewer categories and no optimization (Karolina et al., 2025). Future research should validate the proposed model on larger multi-school datasets, evaluate additional performance metrics, and compare the PCA + PSO pipeline with other feature selection metaheuristics.

REFERENCES

Abukader, A., Alzubi, A., & Adegboye, O. R. (2025). Intelligent system for student performance prediction: An educational data mining approach using metaheuristic-optimized LightGBM with SHAP-based learning analytics. 1-25.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application, 9*(8), 119-136.

Arizmendi, C. J., Bernacki, M. L., Raković, M., Plumley, R. D., Urban, C. J., Panter, A. T., Greene, J. A., & Gates, K. M. (2023). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods, 55*(6), 3026-3054. https://doi.org/10.3758/s13428-022-01939-9

Fister, I., Deb, S., Fister, I., & Andre, J. (2025). Hybrid GA-PSO method with local search and image clustering for automatic IFS image reconstruction of fractal colored images. *Neural Computing and Applications, 7*, 11635-11661. https://doi.org/10.1007/s00521-023-08954-7

González-Devesa, D. (2026). Principal component analysis applied to in-school inertial measurement unit-derived data during physical activity: A systematic review highlighting children's behavioral patterns. *Sensors, 26*(8), 2542.

Jain, I., Jain, V. K., & Jain, R. (2017). Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Applied Soft Computing, 62*, 203-215. https://doi.org/10.1016/j.asoc.2017.09.038

Kamalov, F., Sulieman, H., Alzaatreh, A., Emarly, M., Chamlal, H., & Safaraliev, M. (2025). Mathematical methods in feature selection: A review. 1-29.

Karolina, U., Wijaya, N., & Alamsyah, D. (2025). Classification of elementary school children's potential based on hobbies using the Naive Bayes method with undersampling technique. 4(3), 340-350.

Kumar, S., & Gupta, P. (2026). High accuracy data classification and feature selection for incomplete information systems using extended limited tolerance relation and conditional entropy approach. *IEEE Access*, 14, 1-15.

Kusuma, A. C., Susanto, A., & Rachmawanto, E. H. (2025). Penerapan metode klasifikasi dan seleksi fitur untuk memprediksi minat studi dan karir siswa. *Jurnal Informatika dan Teknologi Komputer (JITEK), 5*(1).

Lötsch, J. (2026). Resolving interpretation challenges in machine learning feature selection with an iterative approach in biomedical pain data. *European Journal of Pain*. https://doi.org/10.1002/ejp.70221

Monika, S. H. (2025). Multi-parametric and priority driven particle swarm (MPPPSO) optimized task scheduling approach for improving performance of fog computing system. *Progress in Artificial Intelligence*, 1-18.

Moradi, P., & Gholampour, M. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing, 43*, 117-130. https://doi.org/10.1016/j.asoc.2016.01.044

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science, 8*(11), 36-42. https://doi.org/10.5815/ijmecs.2016.11.05

Nayak, P., Vaheed, S., Gupta, S., & Mohan, N. (2023). Predicting students' academic performance by mining the educational data through machine learning-based classification model. *Education and Information Technologies, 28*, 14611-14637. https://doi.org/10.1007/s10639-023-11706-8

Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications, 174*, 114765. https://doi.org/10.1016/j.eswa.2021.114765

Pathan, M., & Annapurna, K. (2025). Particle swarm optimization-based multimetric route optimization towards quality of service enhancement in vehicular ad hoc networks. *International Journal of Communication Systems, 38*(16), e70266.

Purnamasari, E., Rini, D. P., & Sukemi, S. (2020). Feature selection using particle swarm optimization algorithm in student graduation classification with naive Bayes method. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 4*(3), 490-495.

Rafdi, A., Saepudin, S., & Sihabudin, S. (2021). Sentiment analysis using naive Bayes algorithm with feature selection particle swarm optimization (PSO) and genetic algorithm. *International Journal of Advances in Data and Information Systems, 2*(2), 89-97.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, 8, 54776-54788. https://doi.org/10.1109/ACCESS.2020.2980942

Salgotra, R., Chaudhary, R., Verma, P., & Mirjalili, S. (2026). The rise of particle swarm optimization: A systematic bibliometric and thematic review (1995-2025). *Archives of Computational Methods in Engineering*. Advance online publication.

Sari, D. P., & Wijaya, K. (2025). New feature selection using principal component analysis: A comparative study of performance as feature extraction versus selection. *Journal of Data Science and Artificial Intelligence, 4*(1), 45-58.

Sartika, D., Sudarsono, B. G., & Purwanti, E. (2024). Support vector machine analysis for interest and talent classification with Python library. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi), 10*(2).

Syafira, D., Suwilo, S., & Zarlis, M. (2020). Analysis of attribute reduction effectiveness on the naive Bayes classifier method. *Journal of Physics: Conference Series, 1566*, 012062.

Talita, A. S., Nataza, O. S., & Rustam, Z. (2021). Naive Bayes classifier and particle swarm optimization feature selection method for classifying intrusion detection system dataset. *Journal of Physics: Conference Series, 1752*, 012021. https://doi.org/10.1088/1742-6596/1752/1/012021

Xue, Y., Xue, B., & Zhang, M. (2019). Self-adaptive particle swarm optimization for large-scale feature selection in classification. *ACM Transactions on Knowledge Discovery from Data, 13*(5), 1-27. https://doi.org/10.1145/3340848

Yağcı, M. (2022). 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