E-ISSN : 3025-1311
https://ejournal.cahayailmubangsa.institute/index.php/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 11 No 5 Tahun 2026

NAÏVE BAYES ALGORITHM IMPLEMENTATION FOR MEASURING THE READING INTEREST LEVEL OF THE BENGKULU COMMUNITY TOWARD THE AKSARAYA SEMESTA LITERACY COMMUNITY PROGRAM

Vony Kurnia Safitri[1], Muhammad Husni Rifqo[2]*, Gunawan[3], Ujang Juhardi[4] Muhamad Khadafi[5]
[1 2 3]Informatics Engineering Study Program, Faculty of Engineering, Universitas Muhammadiyah Bengkulu
[4 5]Informatics Engineering Study Program, Universitas Muhammadiyah Bengkulu

Email: [1]safitrivonny700@gmail.com , [2]mhrifqo@umb.ac.id , [3]gunawan@umb.ac.id , [4]ujangjuhardi@umb.ac.id alfi132004@gmail.com

ABSTRACT

Public reading interest is an important indicator of human resource quality, yet the effectiveness of community-based literacy programs is still commonly assessed descriptively, without measurable data analysis. This study aims to implement the Naïve Bayes algorithm to classify the reading interest level of Bengkulu residents who participated in the Aksaraya Semesta Literacy Community Program into three categories, namely Low, Medium, and High, and to evaluate the performance of the resulting model. The study applies a quantitative approach using the Cross Industry Standard Process for Data Mining (CRISP-DM) framework, covering the Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment stages. Data were collected through a digital 1–5 Likert-scale questionnaire using purposive sampling, yielding 538 records after data cleaning, which were split into 376 training and 162 testing records. Ten literacy-activity features were used as input variables, while the reading-interest label was formed from a separate reading-interest indicator score to avoid data leakage. The model was built and tested in Python on the Google Colab platform using the Scikit-learn library. Results show an accuracy of 80.25%, precision of 60.21%, recall of 68.09%, and F1-score of 62.44%, while 10-fold cross validation produced an average accuracy of 80.64%, a difference of only 0.39% from the testing accuracy. The model performed best on the High category (precision 88%, recall 91%, F1-score 89%), while the Medium category was the most difficult to classify due to class imbalance. The classification results were used as a basis for strategic recommendations for the development of the Aksaraya Semesta Literacy Community Program.

Keywords: Naïve Bayes, CRISP-DM, classification, reading interest, community literacy

Article History
Received: Oktober 2026
Reviewed: Oktober 2026
Published: Oktober 2026

Plagiarism Checker
No 237.u873247.fh8374

Copyright : Author
Publish by : Kohesi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

INTRODUCTION

Reading interest is an important indicator in developing the quality of human resources because it plays a role in improving critical thinking skills, broadening insight, and supporting the advancement of education and social life. Individuals with a good reading interest tend to be more active in obtaining information and developing knowledge, whereas low reading interest can hinder the broader improvement of human resource quality [1]. Although Indonesia's literacy rate is relatively high, this condition has not been fully followed by a high level of public reading interest, so a gap still exists between the ability to read and the habit

of reading consistently. Low reading interest is influenced by internal factors such as motivation and reading habits, as well as external factors such as the availability of reading materials, family and school environments, technology use, and peer influence [2].

In the context of Bengkulu Province, low public participation in literacy activities remains a concern. Research on the development of Community Reading Gardens (Taman Bacaan Masyarakat/TBM) in the city of Bengkulu shows that the utilization and coaching of literacy activities still need to be improved, and that coaching through digital literacy can increase both reading interest and the community's digital literacy skills [3]. This condition indicates that providing literacy facilities alone is not sufficient; it must be accompanied by coaching and activities that encourage active community involvement. One community that actively carries out literacy activities in Bengkulu is the Aksaraya Semesta Literacy Community, which organizes book discussions, provides reading materials, and runs various literacy activities that directly involve the public. However, evaluation of the program's success has generally been based only on participation levels, without in-depth data analysis of participants' reading interest levels, so a data processing approach that can provide a more structured and objective picture of reading interest is needed.

An evaluation approach that relies solely on participation levels has limitations because it cannot capture the depth of participants' engagement. Two participants who attend community activities with equal regularity, for example, may have different levels of motivation, access to reading materials, and involvement in discussions, so their actual reading interest levels may also differ even though their attendance levels are the same. This condition confirms the need for a data analysis approach capable of combining several literacy-activity indicators at once to produce a more representative picture of reading interest level, which is the main focus of this study through a machine-learning-based classification approach.

Advances in computing technology enable more systematic data analysis through data mining approaches, particularly classification, a technique for grouping data into specific categories based on their characteristics. The Naïve Bayes algorithm is a probability-based classification method that uses Bayes' Theorem to determine the class of a data point based on its attributes. This algorithm is widely used because its computation is relatively simple and it can be applied to questionnaire-based data. The application of Naïve Bayes to Likert-scale questionnaire data has been carried out in academic evaluation research and has been shown to be usable for classification based on attributes obtained from questionnaires [4].

Several previous studies have shown that the Naïve Bayes algorithm can be applied to reading-interest problems. Utari and Ulfah applied Naïve Bayes to predict library visitors' reading interest based on age and obtained an accuracy of 80% [5]. Hafiz et al. used a Naïve Bayes Classifier to classify the reading-interest level of 911 students at MTsN 1 Payakumbuh based on a 20-item Likert-scale questionnaire into three categories and obtained a best accuracy of 98.18% on a 70:30 data split [6]. Dewi et al. compared Gaussian and Multinomial Naïve Bayes on 4,476 book-borrowing transactions and obtained a highest accuracy of 97% [7], while Murlena and Syahindra applied Naïve Bayes to classify the reading-interest patterns of regional library visitors [8]. These studies show that Naïve Bayes has been widely applied in library and formal-education contexts, but its application to communities participating in nonformal, community-based literacy programs, particularly in Bengkulu, still has room for development.

Based on the discussion above, this study was conducted to implement the Naïve Bayes algorithm to measure the reading-interest level of the Bengkulu community toward the Aksaraya Semesta Literacy Community Program. This study offers novelty compared with previous studies because it focuses on the general public participating in a nonformal, community-based literacy program, uses ten literacy-activity features as input variables, and applies a comprehensive model evaluation using confusion matrix, accuracy, precision, recall, F1-score, and 10-fold cross validation. The results are expected to provide a more objective

picture of the community's reading-interest level and to be usable as evaluation material for the development of the Aksaraya Semesta Literacy Community Program.

Problem Statement and Research Objectives

Based on the background above, this study is formulated to answer three questions: (1) what is the reading-interest level of the Bengkulu community toward the Aksaraya Semesta Literacy Community Program; (2) how is the Naïve Bayes algorithm applied to classify reading-interest level into the Low, Medium, and High categories; and (3) how does the Naïve Bayes algorithm perform based on confusion matrix, accuracy, precision, recall, F1-score, and 10-fold cross validation. The objectives of this study are to analyze the community's reading-interest level, to implement and evaluate the Naïve Bayes algorithm, and to analyze the classification results as a basis for recommendations for the development of the Aksaraya Semesta Literacy Community Program.

LITERATURE REVIEW

2.1 Reading Interest and the Role of Literacy Communities

Reading interest is a person's attraction to reading activity carried out consciously and continuously, related to a person's willingness, habits, and motivation to read [1][2]. Literacy is a person's ability to obtain, understand, and use information in daily life, which develops not only through formal education but also through social environments and community activities. Literacy communities serve as a social environment that supports the development of a reading culture by providing space for discussion, learning resources, and other literacy activities. Research on literacy villages shows that literacy communities can function as a source of learning and information for the public, with organizers acting as motivators who encourage increased reading interest [9]. Environmental factors, facilities, community, and digital media also play a role in shaping reading interest among literacy community members [10].

2.2 Data Mining and Classification

Data mining is the process of obtaining useful patterns, knowledge, or information from a set of data through the application of certain analytical techniques, used to support analysis and decision-making [11]. Classification is one of the main techniques in data mining used to group data into specific classes based on their characteristics and can be implemented using various algorithms, such as Decision Tree, Support Vector Machine, Neural Network, K-Nearest Neighbor, and Naïve Bayes [12]. The classification process generally involves training data to build the model and testing data to evaluate the model's ability to predict new data. Based on the number of classes, this study is a multiclass classification problem because reading-interest level is divided into three categories: Low, Medium, and High.

2.3 The Naïve Bayes Algorithm

Naïve Bayes is a probability-based classification algorithm that uses Bayes' Theorem to determine the class of a data point based on its characteristics [13]. The name "naïve" comes from the assumption of conditional independence between features once the class is known. Although this assumption is simplistic, it makes Naïve Bayes computationally efficient and it remains widely applied to various classification problems [14]. The general form of Bayes' Theorem is written as follows:

$$P(C|X) = [P(X|C) \times P(C)] \; / \; P(X)$$

where $P(C|X)$ is the posterior probability of class C given data X, $P(X|C)$ is the likelihood, $P(C)$ is the prior probability of class C, and $P(X)$ is the evidence. The classification process is carried out by computing the prior probability of each class, computing the likelihood of each

feature for each class, combining them into a posterior probability, and then selecting the class with the highest posterior probability as the prediction result. According to Patel and Patel, the Naïve Bayes algorithm works by computing the probability of each class based on the training data and then selecting the class with the highest probability as the classification result [15]. The application of Naïve Bayes to reading-interest questionnaire data has shown that this algorithm can be used to determine reading-interest category based on attributes obtained from a questionnaire [5].

Several commonly used Naïve Bayes variants exist, including Gaussian Naïve Bayes, which assumes that feature values are normally distributed and is suitable for continuous numeric data; Multinomial Naïve Bayes, which is more appropriate for frequency or text data; and Bernoulli Naïve Bayes, which is used for binary data. This study uses Gaussian Naïve Bayes because the research features consist of responses to a 1–5 Likert-scale questionnaire treated as numeric data. The choice of the Gaussian variant is also supported by research comparing Gaussian and Multinomial Naïve Bayes on library data, which shows that Gaussian Naïve Bayes provides competitive accuracy on data with numeric characteristics [7].

2.4 Model Performance Evaluation

Classification model performance in this study was evaluated using confusion matrix, accuracy, precision, recall, F1-score, and 10-fold cross validation. The confusion matrix is used to compare the actual class with the predicted class, while accuracy indicates the proportion of correct predictions relative to all test data. Precision indicates the accuracy of a class's predictions, recall indicates the model's ability to identify all data that actually belong to a class, and F1-score is the balance value between precision and recall, which is important to use when the class distribution is imbalanced [16]. In multiclass classification, precision, recall, and F1-score values are computed using a one-versus-rest approach and then averaged using macro-averaging. Cross validation, particularly 10-fold cross validation, is used to determine the generalization ability and stability of the model by dividing the training data into ten parts that are trained and tested alternately [17].

2.5 Research Framework

The conceptual framework of this study integrates the concepts of reading interest, the role of literacy communities, and a Naïve-Bayes-based data mining approach into a single, systematic research flow. Starting from the problem of the lack of data-based evaluation of community literacy programs, this study collects participants' literacy-activity data through a questionnaire, processes that data through the CRISP-DM stages, builds a Naïve Bayes classification model, evaluates the model's performance, and finally uses the classification results as a basis for program-development recommendations. This line of thinking positions the Naïve Bayes algorithm not merely as a technical classification tool but as a component supporting data-driven decision making for literacy-community organizers, so that the research results are relevant both from the standpoint of informatics engineering scholarship and from the standpoint of practical benefit for the development of community literacy.

2.6 Related Work

Several previous studies relevant to this research are summarized in Table 1. This comparison shows that although the Naïve Bayes algorithm has been widely applied to reading-interest problems, those studies generally focus on library or formal-education contexts. This study differs from previous studies because it focuses on the general public participating in a nonformal, community-based literacy program in Bengkulu, using ten literacy-activity features as input variables and a comprehensive model evaluation.

Table 1. Summary of Related Work

Researcher (Year)	Research Focus	Main Result	Difference from This Study
Utari & Ulfah (2021)	Naïve Bayes for predicting library visitors' reading interest based on age	80% accuracy	Library setting, not a literacy community
Hafiz et al. (2026)	Reading-interest classification of MTsN 1 Payakumbuh students using a 20-item Likert questionnaire	Best accuracy 98.18% (70:30 split)	Formal-school students; different number of questionnaire items
Dewi et al. (2025)	Comparison of Gaussian and Multinomial Naïve Bayes on book-borrowing patterns	Highest accuracy 97%	Library transaction data, not a community survey
Murlena & Syahindra (2024)	Reading-interest classification of regional library visitors	Naïve Bayes effective for visit patterns	Regional library visitors as the object
This study (2026)	Reading-interest classification of participants in a nonformal literacy community in Bengkulu	Accuracy 80.25%; CV 80.64%	General public participating in a nonformal community literacy program

RESEARCH METHOD

This study uses a quantitative approach with a data mining method based on the Cross Industry Standard Process for Data Mining (CRISP-DM) framework, which consists of the Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment stages [18]. This framework was chosen because it provides systematic research stages, from understanding the objective to utilizing the model's results, making it suitable for research oriented toward questionnaire-based classification modeling.

3.1 Population, Sample, and Research Instrument

The population of this study is the entire Bengkulu community that has participated in Aksaraya Semesta Literacy Community activities. The sample was taken using purposive sampling with the criteria of residing in the Bengkulu area, having participated in community activities, being willing to fill out the questionnaire, and completing the questionnaire in full. Data were collected through a digital questionnaire based on Google Form with a 1–5 Likert scale (1 = Strongly Disagree to 5 = Strongly Agree). The features (variable X) used as inputs to the classification process include intensity of program participation, frequency of attendance, motivation for joining the community, access to reading materials, and involvement in discussion activities, while the label (variable Y) is the reading-interest level categorized as Low, Medium, and High. The label was formed from the total score of five reading-interest indicators kept separate from the predictor features in order to avoid data leakage.

3.2 Data Processing Stages

The Data Understanding stage was carried out to check the amount of data, the structure of the attributes, the distribution of responses, missing values, and data duplication. The Data Preparation stage included data cleaning, encoding of questionnaire responses into numeric form, formation of the reading-interest-level label based on a score interval (Interval = (Maximum Score − Minimum Score) / 3), and splitting the dataset into training and testing data with a 70:30 ratio. In the Modeling stage, the Gaussian Naïve Bayes algorithm was used to build the classification model using the Python programming language on the Google Colab platform with the Scikit-learn library. In the Evaluation stage, the model was tested using confusion matrix, accuracy, precision, recall, F1-score, and 10-fold cross validation with a macro-average approach for multiclass classification. The Deployment stage used the classification results as a basis for interpretation and for formulating strategic recommendations for the development of the Aksaraya Semesta Literacy Community Program.

The multiclass model evaluation formulas used are as follows: Accuracy = (Number of correct predictions / Total test data) × 100%; Precision = TP / (TP + FP); Recall = TP / (TP + FN); and F1-score = 2 × (Precision × Recall) / (Precision + Recall), with TP, FP, and FN computed using a one-versus-rest approach for each class.

RESULTS AND DISCUSSION

4.1 Data Characteristics and Data Preparation Results

The object of this study is the Aksaraya Semesta Literacy Community located in the city of Bengkulu, a community focused on improving reading culture through shared-reading sessions, book discussions, and the sharing of reading materials. The characteristics of the respondents who answered the research questionnaire varied in terms of gender, age, education level, and occupation, so they reasonably represent the general public participating in a nonformal, community-based literacy program in the Bengkulu area. This diversity of characteristics is important to ensure that the reading-interest classification results reflect not just a particular group but the overall picture of program participants.

Research data were obtained from 541 Bengkulu community respondents who had participated in Aksaraya Semesta Literacy Community activities through a digital 1–5 Likert-scale questionnaire. After data cleaning and missing-value handling, 538 clean records were obtained (3 rows, or about 0.55% of the data, were removed for being incomplete), which were then split into 376 training records (70%) and 162 testing records (30%). The total reading-interest indicator score ranged from 7 to 25, with the class interval computed using the formula (Maximum Score − Minimum Score) / 3, yielding the categories Low (score 7 – <13), Medium (score 13 – <19), and High (score ≥19).

4.2 Prior Probability

Prior probability was computed based on the proportion of data in each class relative to the entire training dataset. The calculation results are presented in Table 2.

Table 2. Prior Probability of Each Class

Reading Interest Category	Training Data Count	Prior Probability
Low	9	0.0239
Medium	73	0.1941
High	294	0.7819
Total	376	1.0000

Table 2 shows that the High class has the largest prior probability (0.7819) because it is dominated by 294 of 376 training records, while the Medium class has a probability of 0.1941 and the Low class only 0.0239. This imbalance in distribution is one of the factors affecting the model's ability to recognize minority classes during evaluation.

4.3 Naïve Bayes Implementation and Classification Results

The Naïve Bayes classification process computes the likelihood probability of each feature value for each class, then combines it with the prior probability to obtain the posterior probability. The class with the highest posterior probability is selected as the prediction. As an illustration, for the intensity-of-program-participation feature (X1), the High class has a likelihood of 0.494983 for value 4 and 0.391304 for value 5, while the Medium class has a likelihood of 0.474359 for value 4, showing that the distribution of feature values differs across categories and contributes to the model's classification decision.

For a more detailed picture, Table 3 presents the likelihood distribution for the intensity-of-program-participation feature (X1) for each class, before it is combined with the prior probability in Table 2 to obtain the posterior probability of each class.

Table 3. Likelihood Distribution of Feature X1 for Each Class

X1 Feature Value	Likelihood | Low	Likelihood | Medium	Likelihood | High
1	0.0714	0.0256	0.0068
2	0.3571	0.1026	0.0204
3	0.2143	0.2308	0.0885
4	0.2143	0.4744	0.4950
5	0.1429	0.1667	0.3913

Table 3 shows that in the Low class, response value 2 (Disagree) dominates with a likelihood of 0.3571, while in the High class, response values 4 (Agree) and 5 (Strongly Agree) dominate with likelihoods of 0.4950 and 0.3913, respectively. This differing distribution pattern is the basis on which the model distinguishes response characteristics across classes: respondents who consistently give high responses on the intensity indicator and other features will obtain a higher posterior probability for the High class, and vice versa.

As an illustration of the classification process, suppose there is a test record with an intensity feature (X1) value of 4 and an attendance-frequency feature (X2) value of 4. The unnormalized posterior probability for the High class is obtained by multiplying the prior probability of the High class (0.7819) by the likelihood of X1=4 for the High class (0.4950) and the likelihood of X2=4 for the High class, while for the Medium class it is obtained by multiplying the prior probability of the Medium class (0.1941) by the likelihoods of X1=4 and X2=4 for the Medium class. Because the likelihood of X1=4 for the High class (0.4950) is greater than for the Medium class (0.4744) and the prior probability of the High class is much larger, the resulting posterior product for the High class tends to dominate, making it more likely that this test record is classified as High. This computational mechanism underlies the entire classification process applied to the 162 testing records in this study.

After all 162 testing records were processed through this mechanism, the confusion matrix presented in Table 4 was obtained.

Table 4. Confusion Matrix of Naïve Bayes Classification Results

Actual \ Predicted	Low	Medium	High
Low	3	1	0
Medium	3	12	16
High	1	11	115

Based on Table 4, the model correctly classified 3 of 4 Low-class records, 12 of 31 Medium-class records, and 115 of 127 High-class records, so the total number of correct predictions is 130 out of 162 testing records (32 records were misclassified). The values on the confusion-matrix diagonal show that the model's performance is best on the High class, consistent with the dominance of the High class in the training data. In the Medium class, the most common misclassification was 16 records predicted as High, whereas in the High class only 12 records were incorrectly predicted as Low or Medium.

4.4 Model Performance Evaluation

The classification-report results and manual calculations based on the confusion matrix in Table 4 are presented in Table 5.

Table 5. Model Evaluation Results per Class and Macro Average

Class	Precision	Recall	F1-Score	Support
Low	43%	75%	55%	4
Medium	50%	39%	44%	31
High	88%	91%	89%	127
Macro Average	60.21%	68.09%	62.44%	162

Overall, the Gaussian Naïve Bayes model achieved an accuracy of 80.25% (130/162 × 100%), a macro-average precision of 60.21%, a macro-average recall of 68.09%, and a macro-average F1-score of 62.44%. The manual accuracy calculation, using the diagonal values of the confusion matrix (3 + 12 + 115 = 130) divided by the total test data (162), yields the same value of 80.25%, so the program results and the manual calculation are mutually consistent. The macro-average precision calculation is obtained from the average precision of the three classes: (0.43 + 0.50 + 0.88) / 3 ≈ 0.6021, while the recall for each class is computed as 3/4 = 0.75 for Low, 12/31 ≈ 0.39 for Medium, and 115/127 ≈ 0.91 for High, with an average of 0.6809.

To test the model's stability, 10-fold cross validation was performed on the training data. The accuracy score of each fold is presented in Table 6.

Table 6. 10-Fold Cross Validation Accuracy Results

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8
87.04%	83.33%	81.48%	83.33%	83.33%	81.48%	83.33%	77.78%
Fold 9	**Fold 10**	**Average**					
73.58%	71.70%	80.64%					

The average accuracy across the ten folds is 80.64% [(87.04 + 83.33 + 81.48 + 83.33 + 83.33 + 81.48 + 83.33 + 77.78 + 73.58 + 71.70) / 10]. The difference between the cross-validation accuracy (80.64%) and the testing-data accuracy (80.25%) is only 0.39%, indicating that the model's performance is relatively consistent and does not show significant overfitting

to any particular data split. The variation in scores across folds, ranging from 71.70% to 87.04%, also shows that the model's performance can fluctuate somewhat depending on the composition of minority classes within each fold, though it remains within a reasonable range for data with an imbalanced class distribution.

Discussion

The evaluation results show that the model performs best on the High category, with precision of 88%, recall of 91%, and F1-score of 89%, while performance on the Medium category is much lower (precision 50%, recall 39%, F1-score 44%) and the Low category yields low precision (43%) despite a relatively high recall (75%). This pattern is consistent with the prior-probability results in Table 2: the dominance of the High class (294 of 376 training records) gives the model more patterns from which to learn that class's characteristics, while the far smaller Medium and Low classes make it harder for the model to distinguish them from the High class. Of the 31 actual Medium-category records, 16 were instead predicted as High, indicating overlapping characteristics between the Medium and High categories in the research data.

This finding is consistent with the nature of the Naïve Bayes algorithm, which uses prior probability as one of the components determining posterior probability, so that classes with a larger data proportion tend to be predicted correctly more easily, while minority classes are prone to misclassification. Compared with Hafiz et al., who obtained accuracy of up to 98.18% in classifying students' reading interest using a Likert questionnaire [6], the 80.25% accuracy of this study is relatively lower but comparable to Utari and Ulfah, who obtained 80% accuracy in a library context [5]. This difference in results is likely influenced by differences in the research object, the number and proportion of classes, and the number of features used as input variables.

The consistency between the testing-data accuracy (80.25%) and the average 10-fold cross-validation accuracy (80.64%) strengthens the model's validity, as the small difference (0.39%) shows that the model not only performs well on one particular data split but is also stable when tested across various training and validation splits. This confirms that the Naïve Bayes algorithm can be used to classify the reading-interest level of the Bengkulu community toward the Aksaraya Semesta Literacy Community Program, although performance on the minority classes (Medium and Low) still needs to be improved in future research, for example through data-balancing techniques such as oversampling or by increasing the number of respondents in the minority classes.

A comparison with Table 1 also shows that studies with a larger and more balanced number and proportion of classes, such as Dewi et al., who obtained 97% accuracy [7], and Hafiz et al., who obtained up to 98.18% accuracy [6], tend to produce higher performance than this study. This indicates that, in addition to the characteristics of the Naïve Bayes algorithm itself, the quality and balance of data distribution have a major influence on the final performance of a classification model. In other words, the 80.25% accuracy obtained in this study does not merely reflect a limitation of the Naïve Bayes algorithm, but is also a consequence of field-data characteristics that are naturally dominated by participants with a High reading-interest level, given that respondents are active literacy-program participants who voluntarily joined the community activities.

From a methodological standpoint, separating the predictor features (intensity, frequency, motivation, access to reading materials, and discussion involvement) from the indicators forming the reading-interest label is an important step that distinguishes this study from some other reading-interest classification studies that tend to use the total questionnaire score directly as both a feature and the basis for labeling. This separation reduces the risk of data leakage, although the consequence is that the relationship between features and label becomes indirect, making classification performance on minority classes more challenging compared with schemes in which the features and label come from the same source.

4.6 Research Limitations

This study has several limitations that should be considered when interpreting the results or planning future research. First, the class distribution in both the training and testing data is highly imbalanced, with the High class making up more than 78% of the training data, so the model tends to be biased toward the majority class and performance on the Low and Medium classes is suboptimal. Second, the amount of data in the Low class is relatively very small (9 training records and 4 testing records), so the precision and recall estimates for that class are less stable and sensitive to small changes in data composition. Third, this study used only one classification algorithm, Naïve Bayes, without comparing it with other algorithms such as Decision Tree, Random Forest, SVM, or K-Nearest Neighbor, so it cannot yet be confirmed whether Naïve Bayes is the most optimal algorithm for this problem. Fourth, the research data are cross-sectional, collected during a single time period, and therefore do not capture longitudinal changes in participants' reading interest over time.

From the perspective of community-development practice, the results of this study also underscore the importance of a differentiated activity strategy based on participant characteristics. Rather than applying a single, uniform type of literacy activity to all participants, community organizers can use the classification results to tailor the form of intervention to the needs of each group. Participants in the Low category, for instance, are likely to need a more personal and intensive approach, such as individual or small-group mentoring, compared with participants in the High category, who tend to already have well-established reading habits and instead need more room to develop as literacy advocates. This classification-based approach is in line with the principle of data-driven program evaluation, which is increasingly applied across social and educational programs to improve the effectiveness of resource allocation.

4.7 Implications for Program Development

The classification results were used as a basis for formulating strategic recommendations for the organizers of the Aksaraya Semesta Literacy Community. For participants in the Low reading-interest category, it is recommended to increase the intensity of routine reading activities, provide more engaging reading materials, and offer ongoing mentoring and motivation. For the Medium category, it is recommended to sustain participation through discussion activities and book reviews and to add variety to literacy activities. For the High category, participants can be involved as mentors or volunteers in community activities to extend the program's impact to other participants. These recommendations show that a data-mining-based approach can provide an objective picture of the condition of literacy-program participants, enabling organizers to design more targeted program-development strategies than an evaluation based solely on participation levels.

Overall, the contribution of this study can be seen from two perspectives. From a scholarly perspective, this study extends the application of the Naïve Bayes algorithm from the library and formal-education contexts widely studied in previous research to the context of the general public participating in a nonformal, community-based literacy program, while also demonstrating the importance of separating features and labels to avoid data leakage in questionnaire-based classification research. From a practical perspective, this study produces a classification model that can be used directly by the organizers of the Aksaraya Semesta Literacy Community as a program-evaluation tool, so that decisions about developing literacy activities are based not only on intuition or participation levels but are also supported by quantitative evidence from data analysis.

CONCLUSION

The Naïve Bayes algorithm was successfully implemented to classify the reading-interest level of the Bengkulu community toward the Aksaraya Semesta Literacy Community Program into three categories—Low, Medium, and High—using 538 records split into 376 training records and 162 testing records. The model achieved an accuracy of 80.25%, a precision of 60.21%, a recall of 68.09%, and an F1-score of 62.44%, with the best performance on the High category (precision 88%, recall 91%, F1-score 89%). The results of 10-fold cross validation produced an average accuracy of 80.64%, differing by only 0.39% from the testing-data accuracy, indicating that the model's performance is relatively consistent and stable. The High category was the most dominant category in the training data (prior probability 0.7819), which contributed to the best classification performance for that category, while the Medium category was the most difficult to classify (recall 39%, F1-score 44%) due to the imbalance in the amount of data across classes. Overall, the Naïve Bayes algorithm has proven usable as a data-based evaluation tool to support the formulation of strategic recommendations for developing the Aksaraya Semesta Literacy Community Program, although the class imbalance in the data needs attention in future research. Future research is recommended to use a larger and more balanced amount of data, to compare Naïve Bayes with other classification algorithms such as Decision Tree, Random Forest, SVM, or KNN, and to develop the model into a web- or mobile-based application so that the reading-interest classification process can be used directly by program organizers.

Recommendations

Based on the results and limitations of this study, several recommendations can be offered. For the organizers of the Aksaraya Semesta Literacy Community, the classification results can be used as routine evaluation material to design more targeted literacy programs according to participants' reading-interest categories, for example by conducting periodic data collection using the same questionnaire instrument. For future researchers, it is recommended to use a larger and more balanced amount of data across classes, for example through oversampling techniques such as SMOTE or by increasing the target number of respondents in the Low and Medium categories, so that the model achieves better generalization on minority classes. Future research can also compare the performance of Naïve Bayes with other classification algorithms, such as Decision Tree, Random Forest, Support Vector Machine, or K-Nearest Neighbor, to determine which algorithm offers the best performance for measuring reading-interest level in the context of nonformal literacy communities. In addition, the model that has been built has the potential to be developed into a web- or mobile-based application, so that the reading-interest classification process can be carried out automatically and used directly by program organizers in data-driven decision-making.

ACKNOWLEDGMENT

The authors would like to thank the Informatics Engineering Study Program, Faculty of Engineering, Universitas Muhammadiyah Bengkulu, for its support throughout this research. The authors also thank the organizers and all participants of the Aksaraya Semesta Literacy Community who were willing to take part in the research questionnaire, as well as the supervising lecturer and examiners who provided guidance, feedback, and suggestions for improvement from the proposal stage through the completion of this study.

REFERENCES

Anidi & Anlianna, "Systematic Literature Review of the Low Interest in Reading in Indonesia with a Narrative Inquiry Approach," *Arus Jurnal Sosial Dan Humaniora*, vol. 4, no. 1, pp. 254–264, 2024, doi: 10.57250/ajsh.v4i1.377.

F. Y. Yuniar & Pamujo, "Analysis of Factors Causing Low Reading Interest Among Upper-Grade Students at SD Negeri 1 Banjarsari Kulon," *Jurnal Sekolah*, vol. 8, no. 2, pp. 273–280, 2024.

S. Syahril & Y. Astria, "Development of Community Reading Garden (TBM) Communities Through Digital Literacy to Increase Public Reading Interest in Bengkulu City," *AL Maktabah*, vol. 10, no. 2, p. 147, 2025, doi: 10.29300/mkt.v10i2.8970.

M. F. Erinsyah, G. W. Sasmito, D. S. Wibowo, & V. K. Bakti, "Evaluation System for Academic Applications Using the Likert Scale Method and the Naïve Bayes Algorithm," *Komputa: Jurnal Ilmiah Komputer Dan Informatika*, vol. 13, no. 1, pp. 74–82, 2024, doi: 10.34010/komputa.v13i1.10940.

L. Utari & Y. Ulfah, "Application of the Naïve Bayes Method for Predicting Reading Interest Based on Age," *TeknoIS: Jurnal Ilmiah Teknologi Informasi Dan Sains*, vol. 11, no. 1, pp. 67–74, 2021, doi: 10.36350/jbs.v11i1.104.

M. Hafiz, E. Budianita, A. Nazir, S. K. Gusti, S. Syarif, & K. Riau, "Classification of Reading Interest Using the Naïve Bayes," *Costing Journal of Economic, Business and Accounting*, vol. 9, no. 1, pp. 211–224, 2026.

F. R. Dewi, A. Eviyanti, A. S. Fitrani, I. Ratna, & I. Astutik, "Classification of Book Borrower Patterns Based on Profession Using the Naïve Bayes Algorithm," *UMSIDA Preprints Server*, 2025, doi: 10.21070/ups.7020.

M. Murlena & W. Syahindra, "Application of the Naïve Bayes Algorithm in Classifying the Reading Interests of Regional Library Visitors," *Knowbase: International Journal of Knowledge in Database*, vol. 4, no. 1, pp. 94–105, 2024, doi: 10.30983/knowbase.v4i1.8680.

W. Herwina & D. Nurul, "Increasing Reading Interest of Learners Through Literacy Villages at PKBM Al-Hidayah Tasikmalaya," *JIV-Jurnal Ilmiah Visi*, vol. 16, no. 2, pp. 1–10, 2022, doi: 10.21009/JIV.1602.1.

A. Islam, R. Rusli, & R. V. Zwagery, "Identification of Factors Shaping Reading Interest Among Literacy Community Members Using Omnilegent," *Jurnal Ecopsy*, vol. 11, no. 1, p. 28, 2024, doi: 10.20527/ecopsy.2024.04.003.

A. Srirahayu & L. S. Pribadie, "Review Paper: Data Mining Classification," *Jurnal Ilmiah Informatika Global*, vol. 14, no. 1, pp. 1–10, 2023, doi: 10.36982/jiig.v14i1.2981.

N. A. Ramadhani & H. A. Rosyid, "Review: Data Mining Algorithms for Data Classification," *Jurnal Inovasi Teknologi Dan Edukasi Teknik*, vol. 2, no. 12, pp. 550–556, 2022, doi: 10.17977/um068v2i122022p550-556.

T. N. Firyal, H. Hamdani, & A. Septiarini, "Application of the Naïve Bayes Method for Classifying Public Satisfaction with Service Quality," *Jurnal Rekayasa Teknologi Informasi (JURTI)*, vol. 8, no. 2, p. 133, 2024, doi: 10.30872/jurti.v8i2.11301.

J. A. Rieuwpassa, S. Sugito, & T. Widiharih, "Implementation of the Naive Bayes Classifier Method for Sentiment Classification of Netflix App Reviews on Google Play," *Jurnal Gaussian*, vol. 12, no. 3, pp. 362–371, 2024, doi: 10.14710/j.gauss.12.3.362-371.

N. Patel & K. Patel, "A Study on Naïve Bayes Classification Algorithm and Its Applications," *International Journal of Computer Science and Engineering*, 2022.

R. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *Journal of Machine Learning Technologies*, 2011.

R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995.

N. C. Sastya & I. Nugraha, "Application of the CRISP-DM Method in Data Analysis to Determine Customer Behavior at MeatSolution," *Unistek*, vol. 10, no. 2, pp. 103–115, 2023, doi: 10.33592/unistek.v10i2.3079.