**International Journal of Innovation Research in Education, Technology and Management**
Vol. 3, No. 1, February 2025, pp. 240~239
ISSN: 3046-6660, DOI: https://doi.org/10.61098/ijiretm.v3i1.269 

# Literature Review: Implementation of the Naive Bayes Algorithm for Classification in Various Fields of Data Mining

Aprilia Nurfazila[1](✉), Hetty Rohayani[2]
[1,2] universitas muhammadiyah Jambi,jambi, Indonesia
aprilianurfazila5@gmail.com , hettyrohayani@gmail.com

**Article Info**

***Article history:***

Received July 13, 2025
Revised September 1, 2025
Accepted February 16, 2026

***Keywords:***

Naive Bayes
Classification
Data Mining
Literature Review
Classification Algorithm

**ABSTRACT**

The significant increase in data volume across various sectors demands efficient, accurate, and adaptive classification methods. The Naive Bayes algorithm is one of the probabilistic classification techniques widely used in data mining due to its model simplicity and its capability to handle high-dimensional data. This study aims to systematically review the application of the Naive Bayes algorithm for data classification in various sectors in Indonesia through a Systematic Literature Review (SLR) approach. Data were obtained from scientific journals published in the last five years (2019–2024) relevant to the topic and analyzed using qualitative descriptive methods. The review results show that Naive Bayes is widely applied in the fields of health, education, social sciences, economics, and technology. Most studies report high accuracy rates, particularly in text classification and imbalanced dataset cases. However, the limitation of this algorithm lies in the assumption of attribute independence, which is often not met in real-world cases. Therefore, several studies combine Naive Bayes with other methods to improve performance. This study provides a comprehensive overview of the strengths and weaknesses of Naive Bayes and serves as a reference for selecting appropriate classification methods in future data mining applications.



***Corresponding Author:***

Aprilia Nurfazila
Universitas Muhammadiyah Jambi, Jambi, Indonesia
Email: << aprilianurfazila5@gmail.com >>

## 1. INTRODUCTION

The volume of data generated by digital systems today is growing rapidly, in the form of text, numbers, and multimedia. This massive data growth poses challenges in terms of management and utilization. One of the main approaches to address this challenge is data mining, which is a method used to process large-scale data and can be defined as a series of activities to extract large amounts of data that can be stored in databases, data warehouses, or other information storage media [1]. Data mining is the process of discovering patterns or interesting information from selected data using specific techniques or methods. The techniques, methods, or algorithms in data mining are highly varied [2]. One of the key techniques in data mining is classification, which aims to group data into specific classes based on their features. In this context, the Naive Bayes classifier is one of the most widely used algorithms due to its simplicity, efficiency, and competitive classification performance.

The main problem identified in this study is the lack of a comprehensive review of the application of the Naive Bayes algorithm across various application sectors in Indonesia, particularly in the context of research

over the last five years. It is important to understand the patterns of application of this algorithm in different fields, its effectiveness on specific data types, and its strengths and weaknesses compared to other classification algorithms. By identifying these areas, this study is expected to provide guidance for the targeted development and optimization of the Naive Bayes algorithm in the future.

Several relevant studies have been conducted in the past five years. Darmayanti and Fajri [3] applied the Naive Bayes algorithm to classify anemia diseases. Their results showed that the algorithm had high accuracy in classifying health-related data, which tends to be imbalanced. In contrast, Aulia et al. [4] combined Naive Bayes with the Decision Tree algorithm to predict machine failure in the manufacturing industry. This hybrid approach demonstrated that Naive Bayes can be integrated with other methods to enhance prediction accuracy.

In the social domain, Pamungkas and Kharisudin [5] used this algorithm to analyze public sentiment regarding the COVID-19 pandemic based on Twitter data. Their research confirmed Naive Bayes's capability in handling unstructured text data. A similar study by Putri and Kharisudin [6] analyzed user reviews on the Tokopedia marketplace application. They compared the performance of Naive Bayes with Logistic Regression and found that Naive Bayes achieved good accuracy with a lighter computational process. Meanwhile, Muttaqin and Kharisudin [7] examined the application of Naive Bayes in sentiment analysis of the Gojek application and compared it with the Support Vector Machine (SVM) algorithm. The results indicated that Naive Bayes excelled in efficiency, although it slightly lagged behind SVM in terms of accuracy.

The differences in the studies above lie in the application domains and evaluation approaches. Some studies evaluated Naive Bayes as a standalone algorithm, while others combined it with different algorithms or conducted direct performance comparisons. However, there is still a lack of research that examines its cross-domain application in a single, integrated review, making it difficult to gain a general overview of the algorithm's strengths and limitations across different application contexts.

Therefore, this study aims to conduct a systematic literature review of the application of the Naive Bayes algorithm for classification in various fields of data mining in Indonesia over the past five years. The main objective of this research is to identify application trends, effectiveness, and the development potential of the algorithm. It is hoped that this study can contribute to informed decision-making in selecting suitable classification algorithms based on data characteristics and application needs.

## 2. METHOD

### 2.1 Tahapan Penelitian

This study employs a **Systematic Literature Review (SLR)** method to examine the application of the Naive Bayes algorithm in data classification across various fields of data mining in Indonesia. The research stages were designed systematically to ensure the validity and reproducibility of the review process. The stages of the research are as follows:

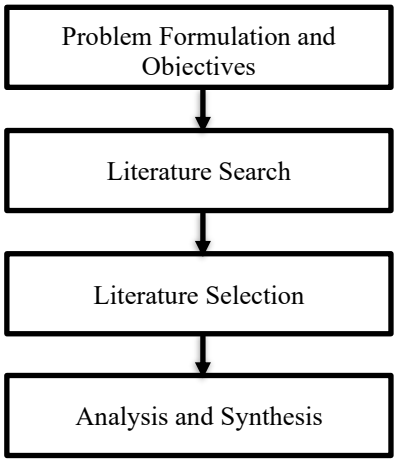


**Figure 1.** Research Workflow Diagram

**1 Problem Formulation and Objectives**

This stage includes identifying the research needs and formulating the research questions. The main focus is to understand how the Naive Bayes algorithm is applied in data classification processes across domains, as well as to evaluate its effectiveness and suitability compared to other methods.

**2 Literature Search**

A structured literature search was conducted using several scientific databases such as Google Scholar, IEEE Xplore, Garuda Ristekdikti, and DOAJ.

**3 Literature Selection**

The search was performed using keywords such as *"Naive Bayes," "classification," "data mining," "classification algorithm,"* and *"Indonesia."* The retrieved articles were filtered based on publication year (2019–2024), content relevance to the research topic, and completeness of information such as methodology, data used, and model evaluation results.

**4 Analysis and Synthesis**

Data extraction was carried out to identify key information from each publication. This included author names, publication year, application domain, type and volume of data, tools used (e.g., RapidMiner, WEKA, or Python), and evaluation metrics (e.g., accuracy, precision, recall, F1-score, and AUC). The collected data were then analyzed qualitatively to identify general trends, algorithm effectiveness, and potential future development.

### 2.2 Sources and Selection of Literature

The literature used in this study was sourced from databases that are well-recognized in the fields of computer science and informatics. The literature collection process was not carried out arbitrarily, but rather through systematic and selective stages to ensure that the analyzed data were genuinely relevant and of high scientific quality. Selected articles were required to originate from nationally or internationally indexed journals or scientific proceedings and had to present content involving the real-world application of the Naive Bayes algorithm for data classification.

The researchers applied specific inclusion criteria, such as recency (published within the last five years), direct relevance to the topic of classification using Naive Bayes, availability of model evaluation metrics, and publication in Indonesia or using data based on Indonesian contexts. Furthermore, articles that did not provide a detailed explanation of the implementation process, lacked model evaluation, or discussed Naive Bayes only from a theoretical perspective without any application context were excluded from the analysis.

Through this rigorous selection process, the researchers ensured that the data used were both credible and highly relevant, and capable of providing an accurate overview of the implementation of the Naive Bayes algorithm across various sectors, including health, education, social sciences, economics, and others.

### 2.3 Data Analysis Techniques

After the literature was collected and selected, the next step was to analyze the content of each article. The analysis technique used was descriptive-qualitative, by comparing each study based on the characteristics of the data used, the application domain, and the performance of the Naive Bayes model. The researchers organized this information into a systematic narrative with the aim of identifying general patterns and distinctive characteristics of Naive Bayes application in the context of data mining in Indonesia.

Additionally, the evaluation metrics used in each study were reviewed. Commonly used classification metrics include accuracy (overall correctness of predictions), precision (the model's ability to accurately predict positive classes), recall (the model's ability to detect all actual positives), as well as F1-score and AUC as indicators of balanced model performance. Although not all articles employed the same metrics, the available data were analyzed proportionally to gain an overall view of the Naive Bayes algorithm's performance.

In several studies, the Naive Bayes algorithm was also directly compared with other methods such as Decision Tree, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). These comparisons are important in assessing the relative strengths of Naive Bayes and identifying contexts where it may be more or less effective than other approaches. The researchers also noted instances of hybrid implementations combining Naive Bayes with other algorithms, indicating the development of adaptive methods suited to specific data types. The results of this analysis were then used as the basis for discussing trends in the use of the Naive Bayes algorithm, its optimization potential, and recommendations for future research and implementation in the field of data mining.

## 3. RESULTS AND DISCUSSION

This section presents the results of the literature review conducted, along with a discussion on the application of the Naive Bayes algorithm for classification across various domains in data mining. The study was carried out using a Systematic Literature Review (SLR) approach, aimed at identifying, evaluating, and synthesizing relevant prior research. The findings focus on the effectiveness, strengths, and challenges associated with the use of the Naive Bayes algorithm in diverse domains such as spam detection, disease prediction, sentiment analysis, and document classification. In addition, this section includes the presentation of data in the form of tables, explanations, and illustrations to support the understanding of the study results.

Table 1. Article Resource

| Author(s) | Journal Title | Dataset | Method | Results |
|---|---|---|---|---|
| Achmad Ridwan (2020) | Application of the Naïve Bayes Algorithm for Diabetes Mellitus Disease Classification [8] | In this study, the data used was the *Early Stage Diabetes Risk Prediction Dataset* obtained from the UCI Machine Learning Repository. The dataset consists of 520 respondent entries with 17 attributes, covering various early symptoms of diabetes such as polyuria, sudden weight loss, blurred vision, and others. After preprocessing, 500 entries were retained for classification, comprising 314 positive and 186 negative cases | The method applied was the Naïve Bayes algorithm, a probabilistic classification technique based on the assumption that features are mutually independent. Data processing was conducted using the RapidMiner software for training, testing, and model performance evaluation | The results showed that the Naïve Bayes algorithm was able to classify the data with an accuracy of 90.20%. The precision for the positive class was 82.35%, recall was 87.50%, and the AUC value reached 0.955, which according to evaluation standards falls into the "excellent" classification. This indicates that the approach is effective for early-stage diabetes risk classification |
| Herry Derajad Wijaya, Saruni Dwiasnati (2020) | Implementation of Data Mining Using the Naïve Bayes Algorithm in Drug Sales [9] | In this study, the data used were secondary data obtained from vitamin sales records at a pharmacy in Jakarta during the period from January to May 2018. The initial dataset consisted of approximately 5,000 entries; however, for testing purposes, the researchers used a sample of 150 entries that had been classified based on variables such as age, marital status, occupation, region, and purchase time. | The method employed was the Naïve Bayes algorithm, a probabilistic classification technique based on the assumption of feature independence. The process involved three main stages: data preprocessing (removal of duplicates and inconsistencies), feature selection, and model validation using 10-fold cross-validation. Data processing was performed using RapidMiner version 8. | The results showed an accuracy rate of 88.00%, with precision reaching 92.81% and recall at 94.16%. Although the AUC value was relatively low (approximately 0.494), the model was generally considered effective in classifying vitamin products into "sold" or "unsold" categories, thus supporting better decision-making in pharmacy stock management. |
| Heliyanti Susana, et l. (2022) | Application of the Naive Bayes Classification Model on Internet Access Usage [10] | This study utilized primary data collected through a survey of students from SMA Negeri 1 Plumbon in Cirebon. A sample of 270 students from grades X, XI, and XII was selected using | The method applied was the Naive Bayes algorithm, a probabilistic classification method that assumes independence among features. | The resulting classification model achieved an accuracy of 89.83%, with prediction details: 34 true positives, 6 false positives, 0 false negatives, and 19 true negatives. When |

| Author(s) | Journal Title | Dataset | Method | Results |
|---|---|---|---|---|
| | | simple random sampling. The dataset attributes included age, gender, grade level, residence, use of mobile phones and laptops, and internet access. After data cleaning, 196 entries were retained for modeling | Data processing followed the Knowledge Discovery in Databases (KDD) process, consisting of data collection, cleaning, transformation, data mining, and evaluation stages. The implementation was carried out using RapidMiner software | applied to 59 new entries, the model predicted that 40 students would access the internet and 19 would not. The model was considered sufficiently accurate for analyzing patterns of internet access usage among students based on demographic attributes and the devices they used |
| Andhini Asri Awaliyah Arifin, et al. (2022) | Implementation of the Naive Bayes Method for Classification of Recipients of the Family Hope Program (PKH) [11] | This study used secondary data obtained from the Social Affairs Office of Asahan Regency, specifically from Sarang Helang Village, Sei Kepayang Timur District. The dataset consisted of 82 entries containing attributes such as income, home ownership status, house size, type of flooring, type of wall, type of roof, water source, and PKH recipient status | The classification approach was implemented using the Naïve Bayes algorithm. The mining process followed the CRISP-DM methodology, from business understanding to deployment. Data were split into training and testing sets (70:30 ratio), with preprocessing including data cleaning and attribute selection. The model was built under the assumption of feature independence and computed using posterior probabilities based on Bayes' Theorem. | The Naïve Bayes classification model yielded an accuracy of 88%, with precision for the positive class reaching 100%, and 77% for the negative class. The recall was 80% for the positive class and 100% for the negative class, while the F1-scores were 89% and 87% respectively. These results indicate the model's effectiveness in accurately identifying potential recipients of the PKH assistance program. |
| Dahri Yani Hakim Tanjung (2022) | Application of the Naïve Bayes Algorithm for ATM Refill Data Classification [12] | This study utilized a dataset of ATM refill records from various banks, totaling 55 entries. Each entry consisted of 6 attributes: CIB percentage, ATM location, ATM status, restock, cash out, and end cash. The data were classified into two categories: Refilled and Not Refilled | The classification method applied was the Naïve Bayes algorithm, a probabilistic approach based on the assumption of feature independence. The classification process followed the Knowledge Discovery in Databases (KDD) framework, | The model successfully classified 50 out of 55 records correctly, achieving an accuracy of 90.91%. The precision for the *Refilled* class reached 94.74%, and 88.89% for the *Not Refilled* class. The recall was 81.82% and 96.97% respectively. These results demonstrate that the Naïve Bayes algorithm is quite |

| Author(s) | Journal Title | Dataset | Method | Results |
|---|---|---|---|---|
| | | | including data cleaning, transformation, and evaluation stages. Testing was conducted using RapidMiner software, with model validation performed via a cross-validation scheme | effective in predicting ATM refill status based on the available attributes |
| Muhammad Farid Satrio Wibowo, et al (2022) | Application of Data Mining and the Naïve Bayes Algorithm for Student Specialization Selection Using a Classification Method [13] | This study utilized a dataset of students from the Informatics Study Program at Universitas AMIKOM Yogyakarta, covering the academic years 2015/2016 and 2016/2017. The initial dataset consisted of 1,614 entries, but after filtering and preprocessing, only valid records (those with complete course grades and logical age values) were used. The analyzed attributes included gender, age, GPA, and grades from several subjects such as Algorithms and Programming, Data Structures, and Database Systems, among others | The classification method applied was the Naïve Bayes algorithm, a probabilistic technique based on the assumption of feature independence. The research followed standard data mining stages, including data collection, cleaning, creation of training and testing sets, and implementation of the classification model. The system was developed using PHP programming language, and Split Validation was used to evaluate the model's accuracy | The Naïve Bayes classification model achieved an accuracy of 84.27% in predicting students' chosen specializations (Programming, Multimedia, or Computer Networks). These results indicate that the model is fairly effective in helping students select a suitable area of concentration based on their academic profile. The study also emphasized that both the quantity and quality of variables significantly influence the accuracy of the model |
| Rohmat Indra Borman , Mina Wati (2020) | Application of Data Mining in Classifying Members of Kopdit Sejahtera Bandar Lampung Using the Naïve Bayes Algorithm [14] | his study utilized loan member data from the Sejahtera Credit Cooperative (Koperasi Kredit Sejahtera) in Bandar Lampung, covering the period from 2015 to 2017. The dataset consisted of 1,064 training records and 300 testing records. The analyzed attributes included gender, age, occupation, income, loan amount, and loan term | The classification method used was the Naïve Bayes algorithm, a probabilistic approach based on Bayes' Theorem. The research followed standard data mining stages, including data cleaning, integration, selection, transformation, classification, and evaluation. The | The resulting classification model achieved an accuracy of 70.33%, with a recall of 70.33% and precision of 100%. These results indicate that although the overall accuracy was not optimal, the model demonstrated a strong ability to correctly identify positive classes. The study concluded that both the quantity and quality of data significantly influence |

| Author(s) | Journal Title | Dataset | Method | Results |
|---|---|---|---|---|
| | | | implementation was conducted using two software tools: WEKA and RapidMiner | classification performance |
| Hakam Febtadianrano Putro, et al (2020) | Application of the Naïve Bayes Method for Customer Classification [15] | This study used sales transaction data of chicken products from UD. Samodro in Sukoharjo. The dataset consisted of 100 entries, divided into 75 training data and 25 test data. The analyzed attributes included purchase quantity, purchase interval, and customer location. The target classification was the customer status: *potential* or *not potential* | The classification method applied was the Naïve Bayes algorithm, a probabilistic approach based on Bayes' Theorem and the assumption of feature independence. The system was developed using PHP and MySQL, and the model was evaluated using the confusion matrix approach to measure classification performance | Out of 25 test data, the model correctly classified 23 entries, resulting in an accuracy of 92%, precision of 100%, and recall of 91%. These results demonstrate that the Naïve Bayes method is effective in identifying potential customers and can serve as a valuable tool for marketing strategy decision-making |
| Rizky Amalia (2020) | Application of Data Mining to Predict Student Graduation Outcomes Using the Naïve Bayes Method [16] | This study utilized exam score data of 6th-grade elementary school students from the *Daftar Kolektif Hasil Ujian Sekolah Berstandar Nasional (DKHUSBN)* for the 2018/2019 academic year. The data were obtained from the Department of Education in East Kotawaringin Regency, particularly from Baamang and Seranau districts. The attributes used included scores in Bahasa Indonesia, Mathematics, and Science | The classification method applied was the Naïve Bayes algorithm, a probabilistic approach based on Bayes' Theorem. The research process followed the CRISP-DM methodology, starting from data collection, preprocessing, and modeling, to evaluation. Implementation was carried out using RapidMiner, with the dataset split into training and testing sets. Evaluation was performed using the confusion matrix and ROC curve | The classification model achieved an accuracy of 82%, with precision for the *Pass* class reaching 96.67% and recall at 78.38%. The AUC value reached 0.970, indicating a very strong classification performance. These results suggest that the Naïve Bayes method is effective in predicting student graduation outcomes based on their exam scores |
| Rudika Rahman, Felix Andreas Sutanto (2023) | Data Mining to Predict Gojek Customer Satisfaction Using the Naïve Bayes | This study utilized primary data collected through questionnaires from 120 respondents who were users of Gojek's services. The data encompassed five | The classification method applied was Naïve Bayes, using a probabilistic approach based on Bayes' Theorem. The system was | The classification model achieved an accuracy of 88.9%. Out of the 36 test entries, the system successfully classified most of the data correctly. This |

| Author(s) | Journal Title | Dataset | Method | Results |
|---|---|---|---|---|
| | Algorithm [17] | key attributes: app quality, punctuality, ride comfort, driver friendliness, and pricing. Each attribute was rated on a four-level satisfaction scale, ranging from "strongly disagree" to "strongly agree" | developed using the waterfall model and implemented as a web-based application. The dataset was split into 70% training data (84 entries) and 30% testing data (36 entries), with evaluation conducted using a confusion matrix | research demonstrates that the Naïve Bayes method is quite effective in predicting customer satisfaction with Gojek's services and can serve as a useful tool for evaluating service quality |
| Putu Sainanda Cahyani Moonallika, et al (2020) | Applying Data Mining to Predict Student Graduation Using the Naïve Bayes Classifier *(Case Study: STMIK Primakara)* [18] | This research utilized two data sources: alumni data and active student data from STMIK Primakara. The analyzed attributes included GPA from semesters 1 to 4, gender, school background, place of origin, and employment status. The data was divided into three groups: 23 alumni records for training, 10 alumni records for validation, and 10 active student records for testing | The classification method used was the Naïve Bayes Classifier, employing a probabilistic approach based on Bayes' Theorem. The study used a confusion matrix to compare predicted results with actual data. The implementation process involved assigning weights to GPA values, calculating class probabilities for all attributes, and conducting prediction tests using WEKA and Excel software | The model achieved an accuracy, recall, and precision of 80%. Out of 10 test records from active students, the system successfully predicted graduation on time or delay with fairly high accuracy. It was found that 8 out of 10 predictions matched the actual outcomes, indicating that this approach is suitable for supporting academic evaluation and graduation prediction |
| Fauzan Alghifari, Didi Juardi (2021) | Applying Data Mining to Food and Beverage Sales Using the Naïve Bayes Algorithm [19] | This study utilized food and beverage sales data from the restaurant "Makan Barbeque Sepuasnya" covering the period from February to December. The data was collected through direct interviews with the restaurant owner and includes monthly sales figures for food and beverages. After preprocessing, the numerical data was classified into "Profit" and "Loss" categories | The classification method applied was Naïve Bayes, using a probabilistic approach based on Bayes' Theorem. The research followed the stages of Knowledge Discovery in Database (KDD): starting from data collection, preprocessing, data transformation using discretization, algorithm implementation | The classification model achieved its best performance in a 3-fold cross-validation scenario, with an accuracy of 88.73%, precision of 64.42%, and recall of 45.41%. The highest kappa value reached 0.451, which is considered moderate. The model successfully grouped food and beverage transactions into profit or loss categories and can be used to support more targeted |

| Author(s) | Journal Title | Dataset | Method | Results |
|---|---|---|---|---|
| | | based on specific thresholds (e.g., food sales exceeding IDR 15 million were considered profitable) | with RapidMiner, and model performance evaluation. Validation was performed using cross-validation and a confusion matrix | restaurant marketing strategies |
| Nurul Alfiah (2021) | Classification of Social Assistance Recipients in the Family Hope Program Using the Naive Bayes Method [20] | This study utilized data on impoverished citizens in Cimanggu District, Cilacap Regency, obtained from the Social Services Office as of January 2019. The initial dataset consisted of 14,832 entries, which were refined through the removal of missing values, resulting in a final dataset of 9,455 records. The analyzed attributes covered 17 variables, including household size (ART), land status, type of flooring, roof type, drinking water source, immovable assets, ownership of boats, refrigerators, telephones, and laptops | The study applied the Naive Bayes Classification method using the CRISP-DM approach. Classification was performed using Weka version 3.9.3, with a percentage split validation technique. Data was divided into various proportions, and the highest accuracy was achieved at a 60% split configuration. The model was tested using a confusion matrix, and performance evaluation included precision, recall, and overall accuracy | The classification model produced its highest accuracy at 84.48%, with a precision of 85% and recall of 98%. Based on the results, the system successfully mapped that of the 3,782 test data entries, 613 citizens were eligible and 3,169 were ineligible to receive PKH assistance. The most influential attributes in the classification process were ownership of boats, laptops, motorboats, and the type of basic household facilities. The emerging pattern indicated that citizens without assets and with a household size of three or more (ART $\geq$ 3) were more likely to qualify for PKH assistance |
| Enggar Novianto, et al (2023) | Classification Using K-Nearest Neighbor, Naive Bayes, and Decision Tree Algorithms for Predicting Undergraduate Graduation Status [21] | This study used academic data from undergraduate law students in the Faculty of Law at Universitas Sebelas Maret, specifically from graduates in 2018 and 2019. The dataset consisted of 300 student entries with attributes such as gender, GPA from semesters 1 to 7, final cumulative GPA, and graduation status (on time or delayed). The data was retrieved from the academic system and formatted in Excel for analysis using RapidMiner | Three classification algorithms were applied: K-Nearest Neighbor (K-NN), Naïve Bayes, and Decision Tree. Implementation followed a 10-fold cross-validation scheme using RapidMiner. Each method was tested using a confusion matrix to measure accuracy, precision, and recall in predicting student graduation outcomes | Evaluation results showed that the K-NN method had the highest performance with an accuracy of 96.67%, followed by Decision Tree at 94.00%, and Naïve Bayes at 77.33%. K-NN also achieved the highest precision and recall scores. The study concluded that K-NN is the most effective algorithm for classifying on-time graduation status and recommended further exploration using clustering techniques for comparison |

| Author(s) | Journal Title | Dataset | Method | Results |
|---|---|---|---|---|
| Ayu Entini Lumban Raja, Koko Handoko (2023) | Implementation of Data Mining Using the Naive Bayes Algorithm for Classifying Eligibility of Basic Food Assistance Recipients [22] | This study utilized data on basic food assistance recipients from Batu Aji Subdistrict, Batam City. Data was collected through interviews and field observations of low-income families. The attributes used included head of household, housing condition, income level, and home ownership status. The dataset was divided into 25 records for training and 5 records for testing | The approach applied was the Naïve Bayes algorithm, which classified data into two categories: eligible and ineligible for assistance. The process included data selection, probability modeling of each attribute, and testing using RapidMiner software. Evaluation was conducted using a confusion matrix and an accuracy performance vector | The model achieved an accuracy of 80%, with a precision of 100% for predicting the "eligible" class. Probability analysis revealed that low income and lack of home ownership were the most dominant factors influencing eligibility classification. The system is considered helpful for ensuring that food assistance is targeted appropriately |
| Adittia Fathah, Christina Juliane (2025) | Gender Classification Using Facial Data with Naïve Bayes and K-Nearest Neighbors Algorithms [23] | This study used facial measurement data from 5,001 individuals not photographs, but numerical observations recorded directly. The attributes analyzed included forehead width and height, nose width and length, lip thickness, and the distance from nose to lips. The attribute "hair length" was excluded as it doesn't belong to facial anatomy | The research compared two classification algorithms: Naïve Bayes and K-Nearest Neighbor (K-NN). The study followed the Knowledge Discovery in Database (KDD) framework, starting with literature review, attribute selection, data transformation into numeric form (binary and continuous), modeling, and evaluation. Data was split using an 80:20 ratio for training and testing. Accuracy evaluation was performed using confusion matrix and Area Under Curve (AUC) via cross-validation | The Naïve Bayes model demonstrated superior performance, achieving a stable AUC of 0.996 across multiple fold configurations, while K-NN reached a maximum of 0.992 at k = 11. Naïve Bayes was considered more consistent and accurate due to the independence between attributes. The study concluded that gender classification based on facial measurements yields high accuracy and is more reliable than digital photo-based classification, which is susceptible to modification |
| Fitriana Harahap, et al (2023) | Implementation of Data Mining to | The study utilized air conditioner (AC) sales data from Amanah | The Naïve Bayes classifier was employed, using a | The model showed moderately strong performance with 75% |

| Author(s) | Journal Title | Dataset | Method | Results |
|---|---|---|---|---|
| | Predict Best-Selling AC Products for Sales Optimization Using the Naïve Bayes Method [24] | Elektronik store. The dataset consisted of 32 entries with four key attributes: product name, capacity (PK), electricity consumption (Watt), and price. Each entry was labeled as either "best-selling" or "not best-selling" based on sales performance categories | probabilistic approach grounded in Bayes' Theorem. The process included data collection, analysis of class and attribute distributions (calculating prior and likelihood values), and implementation using RapidMiner software. Validation was carried out using a confusion matrix to assess classification accuracy and predictive performance | accuracy, 66.67% precision, and 66.67% recall. The classification results revealed that LG-branded products were most frequently identified as "best-selling." These findings suggest that the Naïve Bayes method can support management decisions regarding inventory planning and sales strategy by analyzing product trends |
| Ajif Yunizar Pratama Yusuf, Rafika Sari (2022) | Implementation of the Naïve Bayes Algorithm for Classifying Students' Understanding of the MBKM Program [25] | The dataset used in this study was sourced from an official 2021 survey by the Ministry of Education and Culture (Kemendikbud), targeting students of Universitas Bhayangkara Jakarta Raya. It contained 403 instances with five numerical attributes covering study duration and exam performance and one class attribute representing the level of knowledge (very low, low, middle, high). The class distribution was imbalanced, which posed challenges for classification accuracy | To address this, the study employed the Naïve Bayes algorithm combined with an ensemble learning approach, specifically Adaptive Boosting. The classification process involved data preprocessing (feature selection, integration, and transformation), followed by 10-fold cross-validation, Likert scaling of survey responses, and evaluation using a confusion matrix along with metrics such as accuracy, sensitivity, specificity, and G-Mean | The classification model achieved an average accuracy of 86.16%, with an error rate of 13.84%. Testing results indicated that this method effectively categorized students' understanding of the MBKM program. The study also recommended centralized socialization efforts to improve both awareness and participation in the program |
| Reza Alfaresy Chaerudin, et al (2022) | Implementasi Algoritma Naïve Bayes Untuk Analisis Klasifikasi Survei Kesehatan Mental (Studi Kasus: Open Sourcing | Penelitian ini menggunakan dataset hasil survei kesehatan mental dari organisasi *Open Source Mental Illness (OSMI)* terhadap pekerja di industri teknologi. Dataset awal berjumlah 1.259 entri dan setelah *data cleaning* tersisa 1.254 | Klasifikasi dilakukan dengan algoritma *Naïve Bayes*, diproses menggunakan Python dan Jupyter Notebook. Model dibangun dengan pembagian data 70% untuk pelatihan dan 30% | Model mencapai akurasi tertinggi sebesar 72%, dengan *precision* 73% dan *recall* 92% pada skenario data latih 70% dan data uji 30%. Meskipun *specificity*-nya relatif rendah (26%), penerapan *SMOTE* secara |

| Author(s) | Journal Title | Dataset | Method | Results |
|---|---|---|---|---|
| | Mental Illness) [26] | entri. Data mencakup variabel-variabel seperti umur, gender, riwayat keluarga, pengalaman intervensi, dan persepsi responden terhadap kesehatan mental di lingkungan kerja. | untuk pengujian, serta menerapkan teknik resampling *SMOTE* untuk menyeimbangkan distribusi data antar kelas. Evaluasi kinerja model dilakukan melalui *confusion matrix* serta perhitungan *precision*, *recall*, *specificity*, dan *accuracy* | signifikan membantu meningkatkan prediksi pada kelas minoritas. Model juga diimplementasikan ke dalam sistem prediksi berbasis web menggunakan framework *Flask*. |
| Ahmad Baidowi, Sutisna (2024) | Implementation of the Naïve Bayes Algorithm for Mental Health Survey Classification Analysis (Case Study: Open Sourcing Mental Illness) [27] | This study utilized data from a mental health survey conducted by the Open Sourcing Mental Illness (OSMI) organization, targeting employees in the tech industry. The initial dataset contained 1,259 entries, which were cleaned down to 1,254 entries. Key variables included age, gender, family history, experience with mental health interventions, and respondents' perceptions of mental health in the workplace | Classification was performed using the Naïve Bayes algorithm, implemented with Python and Jupyter Notebook. The data was split into 70% for training and 30% for testing, and the SMOTE resampling technique was used to balance class distribution. Model performance was evaluated using a confusion matrix and metrics including precision, recall, specificity, and accuracy | The classification model achieved a peak accuracy of 72%, with a precision of 73% and recall of 92% under the 70/30 split scenario. Specificity was relatively low (26%), but SMOTE significantly improved predictions for minority classes. The model was further implemented into a web-based prediction system using the Flask framework |

## 4. CONCLUSION

This study aims to evaluate the application of the Naive Bayes algorithm as a classification method across various data mining domains through a systematic literature review approach. Based on an analysis of several scientific articles published in the past five years, it can be concluded that the Naive Bayes algorithm is one of the most widely used classification methods, applied extensively in various sectors such as healthcare, education, social sciences, economics, and technology.

Most of the analyzed studies indicate that Naive Bayes has advantages in terms of computational efficiency and ease of implementation, especially when applied to small- to medium-sized datasets. Furthermore, in the context of text-based or unstructured data, this algorithm is still capable of delivering accurate classification results. Common performance evaluation metrics such as accuracy, precision, recall, F1-score, and AUC show that Naive Bayes is quite competitive compared to other algorithms, particularly in terms of speed and initial classification accuracy.

However, some limitations were also identified, particularly regarding the assumption of feature independence that underlies this algorithm. In cases where the data exhibit strong correlations between attributes, the performance of Naive Bayes tends to decline. To address this, several studies have proposed hybrid approaches or combinations with other algorithms such as Decision Tree and K-Nearest Neighbor. Other studies also show that Naive Bayes can be further optimized through appropriate data preprocessing techniques, such as feature selection and data balancing.

Overall, this review concludes that the Naive Bayes algorithm remains a relevant and effective method for various classification cases in data mining. The choice of this algorithm should be aligned with the characteristics of the data, application needs, and classification objectives. This literature review also provides a comprehensive overview of the trends and patterns of Naive Bayes utilization in Indonesia, which can serve as a foundation for future research development in the field of classification and data mining in general.

**ACKNOWLEDGEMENTS**

Author thanks. In most cases, sponsor and financial support acknowledgments.

**REFERENCES**

[1] S. Amandha, H. Rohayani, and K. Kurniawansyah, "Implementation of Data Mining for Predicting Student Graduation Using the K-Nearest Neighbor Algorithm at Jambi Muhammadiyah University," *Indones. J. Artif. Intell. Data Min.*, vol. 7, no. 1, p. 134, 2024, doi: 10.24014/ijaidm.v7i1.26150.

[2] H. Rohayani and M. C. Umam, "Prediksi Penentuan Program Studi Berdasarkan Nilai Siswa dengan Algoritma Backpropagation," *J. Inf. Syst. Res.*, vol. 3, no. 4, pp. 651–657, 2022, doi: 10.47065/josh.v3i4.1935.

[3]. P. Darmayanti dan I. N. Fajri, "Klasifikasi Penyakit Anemia Menggunakan Algoritma Naive Bayes," *Sustain.*, vol. 11, no. 1, pp. 1–14, 2024, [Online]. Available: http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI

[4] dan D. A. I. R. Aulia, A. Hermawan, "Pendekatan Hybrid: Naïve Bayes dan Decision Tree untuk Prediksi Kerusakan Mesin pada Industri Manufaktur PT X," *Sustain.*, vol. 11, no. 1, pp. 1–14, 2025, [Online]. Available: http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI

[5] F. S. Pamungkas and I. Kharisudin, "Analisis Sentimen dengan SVM, NAIVE BAYES dan KNNuntuk Studi Tanggapan Masyarakat Indonesia TerhadapPandemi Covid-19 pada Media Sosial Twitter," *Pros. Semin. Nas. Mat.*, vol. 4, pp. 1–7, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/prisma/article/view/45038

[6] M. I. Putri and I. Kharisudin, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Analisis Sentimen Data Review Pengguna Aplikasi Marketplace Tokopedia," *Prism. Pros. Semin. Nas. Mat.*, vol. 5, pp. 759–766, 2022, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[7] M. N. Muttaqin and I. Kharisudin, "Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine dan K Nearest Neighbor," *UNNES J. Math.*, vol. 10, no. 2, pp. 22–27, 2021, [Online]. Available: http://journal.unnes.ac.id/sju/index.php/ujm

[8] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020, doi: 10.47970/siskom-kb.v4i1.169.

[9] H. D. Wijaya and S. Dwiasnati, "Implementasi Data Mining dengan Algoritma Naïve Bayes pada Penjualan Obat," *J. Inform.*, vol. 7, no. 1, pp. 1–7, 2020, doi: 10.31311/ji.v7i1.6203.

[10] Heliyanti Susana, "Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet," *J. Ris. Sist. Inf. dan Teknol. Inf.*, vol. 4, no. 1, pp. 1–8, 2022, doi: 10.52005/jursistekni.v4i1.96.

[11] A. A. A. Arifin, W. Handoko, and Z. Efendi, "Implementasi Metode Naive Bayes Untuk Klasifikasi Penerima Program Keluarga Harapan," *J-Com (Journal Comput.*, vol. 2, no. 1, pp. 21–26, 2022, doi: 10.33330/j-com.v2i1.1577.

[12] D. Y. Hakim Tanjung, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Data Pengisian ATM," *Infosys (Information Syst. J.*, vol. 7, no. 1, p. 12, 2022, doi: 10.22303/infosys.7.1.2022.12-24.

[13] M. F. S. Wibowo, N. F. Puspitasari, and B. Satya, "Penerapan Data Mining Dan Algoritma Naïve Bayes Untuk Pemilihan Konsentrasi Mahasiswa Menggunakan Metode Klasifikasi," *J. Inf. Syst. Manag.*, vol. 3, no. 2, pp. 39–45, 2022, doi: 10.24076/joism.2022v3i2.680.

[14] R. I. Borman and M. Wati, "Penerapan Data Maining Dalam Klasifikasi Data Anggota Kopdit Sejahtera Bandarlampung Dengan Algoritma Naïve Bayes," *J. Ilm. Fak. Ilmu Komput.*, vol. 09, no. 01, pp. 25–34, 2020.

[15] H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *J. Teknol. Inf. dan Komun.*, vol. 8, no. 2, 2020, doi: 10.30646/tikomsin.v8i2.500.

[16] R. Amalia, "Penerapan Data Mining Untuk Memprediksi Hasil Kelulusan Siswa menggunakan Metode Naïve Bayes," *J. Inform. dan Sist. Inf.*, vol. 6, no. 1, pp. 33–42, 2020.

[17] R. Rahman and F. A. Sutanto, "Data Mining Untuk Memprediksi Tingkat Kepuasan Konsumen Gojek Menggunakan Algoritma Naive Bayes," *J. Interkom J. Publ. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 18, no. 1, pp. 8–18, 2023, doi: 10.35969/interkom.v18i1.280.

[18] P. Sainanda Cahyani Moonallika, K. Queena Fredlina, and I. Kresna Sudiatmika, "Penerapan Data Mining Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes Classifier (Studi Kasus STMIK Primakara)," *Progresif J. Ilm. Komput.*, vol. 6, no. 1, pp. 47–56, 2020.

[19] F. Alghifari and D. Juardi, "Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," *J. Ilm. Inform.*, vol. 9, no. 02, pp. 75–81, 2021, doi: 10.33884/jif.v9i02.3755.

[20] N. Alfiah, "Klasifikasi Penerima Bantuan Sosial Program Keluarga Harapan Menggunakan Metode Naive Bayes,"

*Respati*, vol. 16, no. 1, p. 32, 2021, doi: 10.35842/jtir.v16i1.386.

[21] E. Novianto, A. Hermawan, and D. Avianto, "Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, Decision Tree Untuk Prediksi Status Kelulusan Mahasiswa S1," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 8, no. 2, pp. 146–154, 2023, doi: 10.36341/rabit.v8i2.3434.

[22] A. L. R. Entini and K. Handoko, "Jurnal Comasie IMPLEMENTASI DATA MINING DENGAN ALGORITMA NAIVE BAYES," *J. Comaise*, vol. 03, pp. 343–351, 2023.

[23] S. Tinggi, M. Informatika, and S. Likmi, "KLASIFIKASI GENDER MENGGUNAKAN DATA WAJAH DENGAN ALGORITMA GENDER CLASSIFICATION USING FACE DATA WITH THE NAÏVE BAYES ALGORITHM AND K-NEAREST NEIGHBORS ALGORITHM," vol. 12, no. 1, pp. 99–110, 2025, doi: 10.25126/jtiik.2025128724.

[24] F. Harahap, W. Fahrozi, R. Adawiyah, E. T. Siregar, and A. Y. N. Harahap, "Implementasi Data Mining dalam Memprediksi Produk AC Terlaris untuk Meningkatkan Penjualan Menggunakan Metode Naive Bayes," *J. Unitek*, vol. 16, no. 1, pp. 41–51, 2023, doi: 10.52072/unitek.v16i1.541.

[25] A. Y. P. Yusuf and R. Sari, "Implementasi Algoritma Naïve Bayes Untuk Klasifikasi Pemahaman Program MBKM Bagi Mahasiswa," *J. Inform. Inf. Secur.*, vol. 3, no. 2, pp. 171–180, 2022, doi: 10.31599/jiforty.v3i2.1713.

[26] R. Alfarezy, E. Ermatita, and R. M. B. Wadu, "Implementasi Algoritma Naïve Bayes Untuk Analisis Klasifikasi Survei Kesehatan Mental (Studi Kasus: Open Sourcing Mental Illness)," *Inform. J. Ilmu Komput.*, vol. 19, no. 1, pp. 1–10, 2023, doi: 10.52958/iftk.v19i1.4696.

[27] A. Baidowi, "Implementasi Data Mining Klasifikasi Fuel Surcharge Menggunakan Algoritma Naive Bayes Studi Kasus PT Pelabuhan Indonesia ( Persero ) Regional 2 Tanjung Priok," vol. 5, no. 3, pp. 2854–2863, 2024.