**DigCompSci**

**Journal of Digital Computer Science and Informatics**

Vol 1 No 1 Desember 2025

ISSN: XXXX-XXXX (Print) ISSN: XXXX-XXXX (Electronic)

Open Access: https://journaldigcomsci.aon-terbitqu.com/digcompsci/index

# IMPLEMENTATION OF NAÏVE BAYES ALGORITHM BASED ON PARTICLE SWARM OPTIMIZATION FOR INTRUSION DETECTION SYSTEM (IDS)

**Nugraha [1*], Yulhan Wahyudin [2], Adrian Reza [3] (11 pt)**

[1] Universitas Nusaputra (9 points)

[2] Universitas Nusaputra (9 points)

[3] Universitas Nusaputra (9 points)

*email: nugraha@nusaputra.ac.id*

**Article Info :**

Received:

DDMMYY

Revised:

DDMMYY

Accepted:

DDMMYY

***ABSTRACT** (10 PT)*

Intrusion Detection System (IDS) is an important mechanism in detecting suspicious activity on computer networks, including intrusion attempts that have the potential to threaten data security. This study aims to analyze the application of the *Naïve Bayes algorithm* optimized with *Particle Swarm Optimization* (PSO) in the classification of network attacks using **the KDD Cup 1999 dataset** . The research stages include data collection, cleaning, attribute selection, transformation, modeling, and evaluation using RapidMiner Studio software. The *Naïve Bayes algorithm* was chosen because of its simplicity and efficiency in classifying large data, while PSO optimization was applied to improve the accuracy of the classification results. Performance evaluation was carried out with *a confusion matrix* through accuracy, precision, and *recall metrics* . The test results showed that the model produced an accuracy rate of 95.00%, *a recall* of 98.08%, and a precision of 92.73%. These findings prove that the integration of *Naïve Bayes* and PSO can improve the performance of IDS in detecting attacks with a lower error rate. The practical implication of this research is the availability of an effective computational approach to support cyber attack prevention strategies, while contributing to the development of more adaptive network security systems.

**Keywords:** Intrusion Detection System, Naïve Bayes, Particle Swarm Optimization, Data Mining, RapidMiner.



## INTRODUCTION

An Intrusion Detection System (IDS) is a software system that can be used to detect suspicious activity in a computer system or network. Data and information can be processed in a computer network, so a security system is needed in a computer network that is resistant and tolerant to network intrusions. Network intrusions are attempts to gain access. An Intrusion Detection System (IDS) is used to detect suspicious activity in a system or network and all traffic flowing into a network will be analyzed to find out if there is an attempted attack or intrusion into the network system.

Intrusion Detection System is generally located in an important network segment where the server is located or is located at the "entrance" of the network. Data and information can be processed in a computer network often there is suspicious activity or an attack on a computer or computer network, network intrusion is an attempt to gain access, Intrution there is suspicious activity in a system or network and All traffic flowing into a network will be analyzed to find out whether there is an attempted attack or intrusion into the network system Intrution Detection System is generally located in an important network segment where the server is located or is located at the "entrance" of the network.

Data mining is a form of implementation applied to find models and patterns capable of classifying data based on previous data over a specific time period. Data mining is a form of data mining used to extract knowledge from large amounts of data. One algorithm used in data mining techniques that utilizes classification theory is Naïve Bayes. Accurate and precise data is needed for classifying Intrusion Detection System attacks.

Classification is used to assess data by assigning it to a number of available classes. Classification involves building a model by training it on testing and training data to store it for use in classifying data. An intrusion detection system is the process of monitoring network traffic within a system to detect suspicious data patterns that could potentially lead to an attack.

## RESEARCH METHODS

This research uses a quantitative approach with data mining methods to explore patterns of relationships between transaction data to generate new knowledge useful for business decision-making. Data mining itself is part of the Knowledge Discovery in Database (KDD) process, which includes a series of stages ranging from data collection, cleaning, transformation, to pattern discovery (Han & Kamber, 2012). With this approach, the research focuses on the application of the FP-Growth algorithm in Market Basket Analysis (MBA) to identify association rules emerging from consumer transaction data.

### Research Design

The research design used is quantitative descriptive, which aims to explain consumer purchasing phenomena through the systematic use of transaction data. This design allows researchers to explore the relationships between products purchased simultaneously, thus providing a basis for recommendations on promotional strategies, cross-selling, and inventory management. Within the KDD framework proposed by Fayyad et al. (1996), this study emphasizes the stages of processing transaction data until valid association rules are formed.

### Objects and Data Sources

The object of this research is sales transaction data from a modern retail unit that offers a variety of daily necessities. The data used is secondary data taken from the company's transaction recording system. A total of 1,684 transactions were collected, covering a wide variety of product types. Of all the products, the research focused on the 30 best-selling products that appeared most frequently in the transaction records, in accordance with the principles of data selection in data mining (Larose, 2005).

The selection of best-selling products aims to facilitate analysis and ensure that the resulting association patterns are highly relevant to consumer behavior. The primary variables analyzed are transaction ID and product name, which are then transformed into binary format as required by the FP-Growth algorithm (Witten et al., 2011).

**Research Stages**

The research stages are arranged based on the Knowledge Discovery in Database (KDD) framework, which includes the following steps:

1. Data Selection – this stage selects transaction data that is relevant to the research, namely consumer purchase transactions with a focus on the 30 best-selling products.
2. Data Cleaning – This stage cleans the data from duplication, missing data, and inconsistencies. This step is crucial to ensure analysis results are not biased by noise in the dataset.
3. Data Transformation – Transaction data, originally tabular, is converted into binary format using the One-Hot Encoding method. This transformation results in a table with a value of "1" if a product appears in a particular transaction, and "0" if it doesn't. This format serves as the primary input for the FP-Growth algorithm.
4. Data Mining – the core stage of the research, namely the application of the FP-Growth algorithm to explore association rules between products based on minimum support and minimum confidence parameters. This algorithm builds an FP-Tree structure that enables a more efficient pattern search process than the Apriori algorithm (Han & Kamber, 2012).
5. Evaluation – this stage evaluates the results of the association rules obtained using measures of support, confidence, and lift ratio. The lift ratio is used to measure the strength of the association, whether it is greater than 1 (positive), equal to 1 (neutral), or less than 1 (negative) (Larose, 2005).

**Data Analysis Techniques**

Data analysis was performed using the FP-Growth algorithm, a frequent pattern mining method designed to discover itemset patterns that frequently appear in transactions. This algorithm is considered more efficient than Apriori because it doesn't require generating a large number of candidate itemsets but instead builds a compact tree structure (FP-Tree).

The analysis process involves two main parameters, namely Minimum Support: the threshold for the frequency of occurrence of product combinations in the dataset and Minimum Confidence: the threshold for the probability of product Y appearing if product X is purchased.

By determining appropriate support and confidence values, the FP-Growth algorithm generates a series of association rules in the form of if-thens (e.g., if a consumer buys product A, then they are likely to also buy product B). These rules are then further analyzed for business relevance.

**Model Evaluation**

Evaluation of the results of the association rules is carried out using three main indicators:

1. Support – describes the proportion of transactions containing a particular combination of items to all transactions.
2. Confidence – indicates the level of reliability of the rule, namely how often item Y appears if item X appears.
3. Lift Ratio – used to evaluate the strength of a rule, whether the relationship between items is greater than random expectation. A lift value > 1 indicates a significant relationship between items, while a value < 1 indicates a weak relationship (Witten et al., 2011).

This evaluation ensures that the obtained association rules not only occur frequently, but also have practical relevance to support business strategy.

**Research Software**

The algorithm implementation process was carried out using RapidMiner Studio software, an open-source application that supports various data mining methods, including FP-Growth. RapidMiner was chosen because it has an intuitive graphical interface and allows users to build analytical models without having to manually write code. Furthermore, RapidMiner also supports automatic evaluation of association results using support, confidence, and lift parameters, thus facilitating validation of research results (Larose, 2005).

## RESULTS AND DISCUSSION

### RESULTS

After the discussion in the previous chapter, the author continues the discussion on the results of the tests that have been carried out, where to find out the results in terms of accuracy, precision and recall of the naïve Bayes algorithm based on particle swarm optimization using the rapidminer software application.

### Dataset Description

In this study, calculations will be carried out using the Naïve Bayes Algorithm method for data processing. 1000 clean datasets will later be divided into 90% training data and 10% testing data. The initial stage carried out in this study is preparing the data, the data to be processed is IDS data. As a calculation of the Naïve Bayes algorithm with the classification method, the author takes the data to be processed as an example, namely 1000 data records and takes 90% training data (900 Data) and 100 testing data, as shown in the table below:

**Table 1 Description of Transaction Data**

| **NAÏVE BAY'S** | | | | | |
|---|---|---|---|---|---|
| **P(CI)** | | | | | |
| **Step** | | **Number of Cases** | **No** | **Yes** | **Amount** |
| | Total | 100 | | | |
| | Classification | | 52 | | 0.52 |
| | Classification | | | 48 | 0.48 |
| | | | | | |
| Training: Calculate P(X\|Ci) for each class | | | | | |
| Protocol Type | | | | | |
| | icmp | | 42 | | 0.807692 |
| | icmp | | | 38 | 0.791667 |
| | | | | | |
| | udp | | 2 | | 0.038462 |
| | udp | | | 10 | 0.208333 |
| | | | | | |
| | TCP | | 4 | | 0.076923 |
| | TCP | | | 4 | 0.083333 |
| | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| **Dst_Host_Srv_Count** | | | | | |
| | 255 | | 44 | | 0.846154 |
| | 255 | | | 26 | 0.541667 |
| | | | | | |
| | 1 | | 4 | | 0.076923 |
| | 1 | | | 26 | 0.541667 |
| Flag | | | | | |
| | 2 | | 8 | | 0.153846 |
| | 2 | | | 50 | 1.041667 |
| | | | | | |
| | 4 | | 40 | | 0.769231 |
| | 4 | | | 2 | 0.041667 |
| | | | | | |
| Attack | | | | | |
| | Normal | | 52 | | 1 |
| | Normal | | | 0 | 0 |
| | | | | | |
| | Dos | | 0 | | 0 |
| | Dos | | | 42 | 0.875 |
| | | | | | |
| | Probe | | 0 | | 0 |
| | Probe | | | 6 | 0.125 |
| | | | | | |
| dst_host_count | | | | | |
| | 255 | | 46 | | 0.884615 |
| | 255 | | | 28 | 0.583333 |
| | | | | | |
| | 4 | | 2 | | 0.038462 |
| | 4 | | | 24 | 0.5 |

**Evaluation and Testing of the Rapid Miner Tool**

In this process, the classification method uses the Naïve algorithm. Bayes applied to *Intrusion prediction Detection System* . In this study, the author used calculation testing with *tools* Rapid Miner , the test results obtained using the Rapid Miner *tool* are as follows:

1. Import the data required for processing in the Rapid Miner tool . In the Rapid Miner application , select Read. excel and click Import Data, then select the data that will be used and then determine the attributes and labels that will be used.
2. Click Cross Validation , in the process view, add the Naïve algorithm Bayes to the process display screen.

3. Next, in the Apply Model menu, data modeling will be carried out from the dataset used in this process.
4. Then select Performance to apply the Naïve algorithm. Bayes on the accuracy , precision and recall processes that will be carried out.
5. Connect all these commands so that the process display screen displays the following flow:

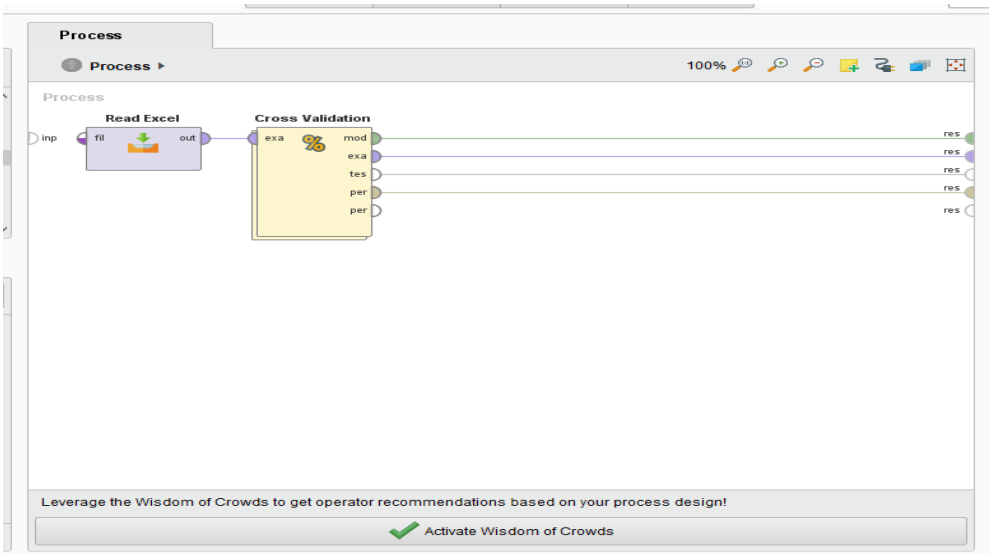


**Figure 1 Rapid Miner Process**

6. Double click on Cross Validation of the visualization form using Performance can also be seen in the following image :

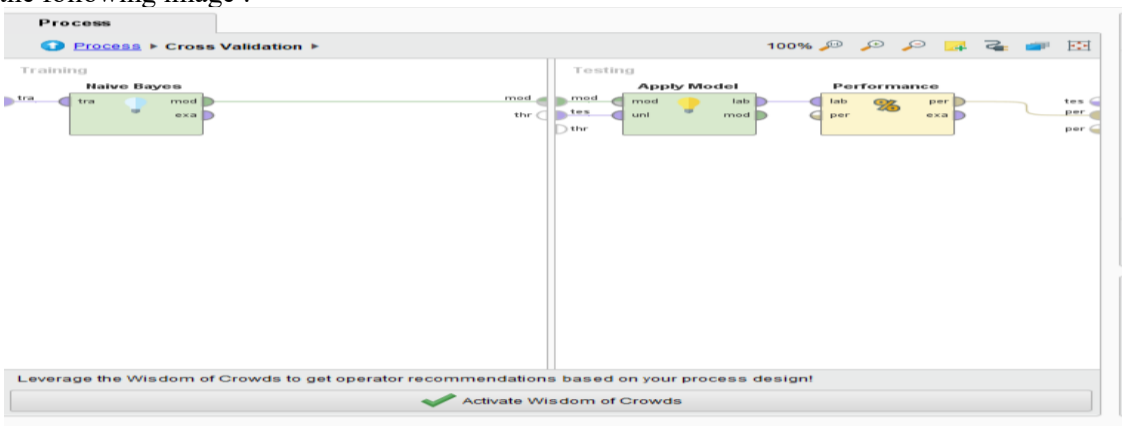


**Figure 2 Rapid Miner Process**

7. After *running The process* on the Rapid Miner *tool* , obtained the results of accuracy , precision and recall of the processed data and can be seen in the following image:

accuracy: 95.00% +/- 5.00% (micro average: 95.00%)

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 51 | 4 | 92.73% |
| pred. Yes | 1 | 44 | 97.78% |
| class recall | 98.08% | 91.67% | |

## DISCUSSION

Method testing is carried out to determine the results of the analyzed calculations and to measure whether the method and algorithm used are functioning properly. The testing process uses tools. *rapidmener* and see whether the data matches the results obtained through the tool . Meanwhile, validation of methods and algorithms *Naive Bayes* is carried out by measuring the results of *accuracy , precision and recall* and can be calculated using *the Confusion Matrix* as follows:

1. Naive Bayes

   accuracy value is calculated by adding up the correct data that has a positive value ( True Positive ) plus the Negative value ( True Negative) divided by the number of correct data that has a positive value ( True Negative) Positive ), Negative ( True Negative) and added with false data that has a positive value ( False Positive), Negative ( False Negative).

**Table 2** ***Confusion Matrix Accuracy Calculation ( Naïve Bayes )***

| | | True Value | |
|---|---|---|---|
| | | *TRUE* | *FALSE* |
| Prediction Value | *TRUE* | TP<br>51 | FP<br>4 |
| | *FALSE* | FN<br>1 | TN<br>44 |

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100\%$$

$$= \frac{51 + 44}{51 + 44 + 4 + 1} * 100\%$$

$$= \frac{95}{100} * 100\%$$

$$= 0{,}95 * 100\%$$

$$= 95.00\%$$

Table 3 ***Confusion Matrix for Precision Calculation ( Naive) Bayes )***

| | | **True Value** | |
|---|---|---|---|
| | | *TRUE* | *FALSE* |
| Prediction Value | *TRUE* | TP<br>51 | FP<br>4 |
| | *FALSE* | FN<br>1 | TN<br>44 |

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} * 100\%$$

$$= \frac{51}{51 + 4} * 100\%$$

$$= \frac{51}{55} * 100\%$$

$$= 0.9272727273 * 100\%$$

$$= 92.73\%$$

**Table 5 Confusion Matrix Recall Calculation ( Naive Bayes )**

| | | **True Value** | |
|---|---|---|---|
| | | *TRUE* | *FALSE* |
| Prediction Value | *TRUE* | TP<br>51 | FP<br>4 |
| | *FALSE* | FN<br>1 | TN<br>44 |

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\%$$

$$= \frac{51}{51 + 1} * 100\%$$

$$= \frac{51}{52} * 100\%$$

$$= 0{,}9807692308 * 100\%$$

$$= 98.08\%$$

From the test data, the results of the data indicate the level of *accuracy , recall* and *precision* of the *Naive algorithm. Bayes* . The following are the results of the *accuracy , recall* and *persicion*

**Table 6 Results of Accuracy , Recall and Perception Values**

| No | Algorithm | Accuracy | Precision | Recall |
|---|---|---|---|---|
| 1 | Naive Bayes | 95.00% | 92.73% | 98.08% |

**Analysis of Results**

Based on the results obtained in this study, the test results of the Naïve algorithm Bayes produces an accuracy rate of 95.00%, Recall 98.08 and Precision 92.73% because the attributes ( attack ) are normal, dos, probe and r2l, normal ( no attack), each data based on the results of determining the label or attribute gets good results in a data or all attributes in increasing the results of *accuracy , recall* and *precision* . And the test results in classifying IDS data are included in the No class which means in the prediction of intrusion data detection The system ensures that there are no attacks on a network. This increased accuracy can facilitate decision-making and preventative measures for each algorithm. This is one factor contributing to high accuracy because each attribute and class or label has an impact on the algorithm.

**CONCLUSION**

From the analysis and results of the Intrusion Detection System data processing, the following conclusions can be drawn:

1. This increase in accuracy can facilitate decision making and prevention efforts for each algorithm is one of the factors that causes high accuracy values because each attribute and class or label has an influence on the algorithm.
2. Based on the results obtained in this study, the test results of the Naïve Bayes algorithm produced an accuracy level of 95.00%, Recall 98.08 and Precision 92.73% because the attributes (attack) were normal, dos, probe and r2l, normal (no attack).
3. The classification method is processed using the Naïve Bayes algorithm, the results of which also show new information, namely the probability value of Class No (0.520) distributions Class Yes (0.480).

**REFERENCES**

M. Khudadad and Z. Huang, “Intrusion Detection with Tree-Based Data Mining Classification Techniques by Using KDD,” Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST, vol. 227 LNICST, no. 6, pp. 294–303, 2018, doi: 10.1007/978-3-319-73447-7_33.

INT Wirawan and I. Eksistyanto, “Application of Naive Bayes in Intrusion Detection System with Variable Discretization,” JUTI J. Ilm. Teknol. Inf., vol. 13, no. 2, p. 182, 2015, doi: 10.12962/j24068535.v13i2.a487.

L. Dhanabal and SP Shantharajah, “A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 6, pp. 446–452, 2015, doi: 10.17148/IJARCCE.2015.4696.

Jupriyadi, “Implementation of Feature Selection Using the FVBRM Algorithm for Attack Classification in Intrusion Detection Systems (IDs),” Semin. Nas. Teknol. Inf., vol. 17, no. January 2018, pp. 1–6, 2018.

N. Rosli et al., “Jurnal Teknologi,” vol. 1, pp. 1–6, 2015.

G. Wang et al., “No Title 大学生の職業未決定の研究,” Appl. Catal. A Gen., vol. 58, no. 2, pp. 15–22, 2013, doi: 10.1179/1743280412Y.0000000001.

H. Tiaield, "Data Mining Based Cyber-Attack Detection," Univ. Glas., vol. 13, no. 2, pp. 90–104, 2017, doi: 10.1063/1.2975179.

M. Surahman et al., "APPLICATION OF SVM-BASED MACHINE LEARNING METHOD FOR," pp. 196–206, 2020.

I. Rahmadani, HS Tambunan, and IS Damanik, "Application of Data Mining in Cities Based on Provinces Responsive to Narcotics Threats Using K-Medoids," vol. 2, pp. 93–99, 2020.

G. Widi N. Dicky Nofriansyah, Data Mining Algorithms and Testing. Yogyakarta: CV Budi Utama, 2015. Suyanto, Data

a Mining. Yogyakarta: Informatika, 2017.

O. Villacampa, "(Weka - Thesis) Feature Selection and Classification Methods for Decision Making: A Comparative Analysis," ProQuest Diss. Theses, no. 63, p. 188, 2015.

Retno Tri Vulandari, Data Mining. Yogyakarta: Gava Media, 2017.

Y. Silalahi, Kristiani Silalahi., Murfi, Hendri., Satria, "Comparative Study of Feature Selection for Support Vector Machine in Credit Risk Assessment Classification," vol. 1, no. 2, pp. 119–136, 2017.