Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 142−150
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9668

Classification of School Students Lifestyle Risks Based on Smoking Behavior Using Naïve Bayes

Oktaria Dwi Cahyani*, Deltari Balka, Dinni Rezky Amelia, Rainda Cintari Aulya, Ken Ditha Tania, Allsela Meiriza, Zaqqi Yamani

Faculty Computer Science, Universitas Sriwijaya, Palembang, Indonesia
Email: [1,*]oktariadwicahyani@gmail.com, [2]deltaribalka3@gmail.com, [3]dinnirezkyameliaa@gmail.com
[4]raindacintariaviya@gmail.com, [5]kenya.tania@gmail.com, [6]allsela_meiriza@yahoo.co.id, [7]zaqqi_yamani@unsri.ac.id,
Correspondence Author Email: oktariadwicahyani@gmail.com
Submitted: **14/04/2026**; Accepted: **02/06/2026**; Published**: 05/06/2026**

Abstract−This study aims to classify students' lifestyle risks based on smoking behavior using the Naïve Bayes algorithm within a knowledge management framework. The research was conducted on students at a vocational high school within the coverage area of a local community health center. The dataset consisted of 277 valid records after undergoing data selection, cleaning, and transformation stages. The modeling process was carried out using RapidMiner software with an 80:20 data split for training (221 students) and testing (56 students). The evaluation metrics used included accuracy, precision, recall, and confusion matrix. The experimental results demonstrate that the Naïve Bayes model achieved an accuracy of 85.92%, precision of 86.12%, and recall of 92.86% for the unhealthy class. Furthermore, the classification results were integrated into a knowledge management framework to support decision-making processes in schools and community health centers. This study contributes to the application of predictive data mining in adolescent health and demonstrates how classification models can serve as effective tools for early detection, preventive interventions, and evidence-based policy formulation in educational and health settings.

Keywords: Data Mining; Knowledge Management; Naïve Bayes; RapidMiner; Students

1. INTRODUCTION

Health is an important aspect that needs to be addressed by school age because it is greatly influenced by daily lifestyle. One lifestyle factor that has a significant impact on health is smoking [1]. Smoking behavior has now transformed into a serious global health threat for various age groups, including adolescents. The World Health Organization (WHO) reports that tobacco consumption contributes to the premature deaths of millions of individuals each year through various chronic diseases such as heart attacks and strokes. Cigarettes are now considered a necessity, almost equivalent to a basic need for some. The courage to try and consume cigarettes is often seen as something to be proud of, by both men and women [2]. The high number of teenage smokers not only causes short-term physical problems such as stunted lung development but also increases the risk of cancer and long-term organ damage.

To understand and identify the level of health risks associated with smoking behavior in students in a more systematic and data-driven manner, an analytical approach capable of processing information accurately and in a structured manner is required. Data mining is the process of extracting valuable knowledge or information from a large data set complex [3]. Data Mining plays a role in extracting and processing raw data into more meaningful information, so that it can be used to discover patterns and knowledge related to the level of risk being studied. Data mining is an important part of the knowledge discovery in databases (KDD) process, which is a series of activities aimed at transforming raw data into useful information [4]. This process includes several stages, from data pre-processing to post-processing of the results of data mining by applying algorithms. Naive Bayes to classify the risks of students' unhealthy lifestyles [5]. Naïve Bayes is a classification created by British scientist Thomas Bayes that uses probability and statistical methods to predict future opportunities based on previous experience. This is known as Bayes' Theorem and combined with naïve bayes, which assumes that the conditions between attributes are mutually independent [6] Naive Bayes was chosen because it can quickly and efficiently process probability values from various variables, such as students' physical condition and cigarette smoke exposure. This method has simple calculations but still provides fairly accurate classification results, making it suitable for analyzing students' lifestyle risks [7].

The choice of the Naive Bayes algorithm in this study is supported by several previous comparative studies. Research by Samuel, Idmi, and Triyono (2025) comparing Naive Bayes and C4.5 for stroke prediction showed that the C4.5 algorithm achieved the highest accuracy of 95%, outperforming Naive Bayes [8]. Meanwhile, a study by Muttakin, Rusmana, and Ramadhani (2025) in heart disease prediction found that Decision Tree and Random Forest achieved 99% accuracy, higher than SVM (88%) and KNN (84%) [9]. On the other hand, research by Tanti et al. (2024) on the detection and classification of respiratory diseases using eight machine learning algorithms showed that Random Forest and Naïve Bayes had the best performance in terms of accuracy and class separation ability. On the other hand, research by Amritha et al. (2026) on heart disease prediction proved that Random Forest was the most effective model with an AUC of 0.9517 after going through a hyperparameter optimization process [10]. A study by Wantoro et al. (2025) that evaluated strategies for handling class imbalance on a diabetes dataset showed that Random Forest produced the best performance in detecting diabetes cases, with a significant increase in accuracy compared to other models including Naive Bayes [11]. Meanwhile, research by Husaini, Priyanto, and Martono (2026) in sentiment analysis of medical personnel services found that the SVM algorithm achieved the highest accuracy of 91.8%, outperforming Random Forest and Naïve Bayes [12]. Based on these findings, Naive Bayes remains relevant for use

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 142−150
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9668

due to its consistency and computational efficiency in classifying health data, although in some cases ensemble algorithms such as Random Forest show superior performance. The use of this classification method allows the system to predict student behavioral tendencies more objectively compared to manual analysis.

Based on a review of six previous studies, it can be identified that these studies focused on predicting degenerative diseases such as stroke, heart disease, diabetes, and respiratory diseases in the general population or hospital patients using standard medical datasets. No research has been found that specifically implements the Naive Bayes algorithm in the context of classifying unhealthy lifestyle risks, particularly smoking behavior, in a high school student population [13]. Furthermore, these studies generally stop at the stage of evaluating model accuracy without integrating the classification results into a knowledge management framework to support strategic decision-making at the community health center or educational institution level [14]. This gap is the basis for this study. This study aims not only to classify the risk level of student smoking behavior using the Naive Bayes algorithm but also to design how these prediction results can be managed as an institutional knowledge base [15]. This study offers a novelty in the form of integration between predictive data mining and knowledge management in the domain of adolescent health in the school environment.

These prediction results can then be used to manage and develop knowledge related to students' health conditions in a more targeted manner. The combination of knowledge management and technology allows student health screening data to be processed into information that can be used to aid decision-making. Through this research, it is hoped that schools can expand and strengthen their knowledge base, allowing for more comprehensive and structured mapping of student health profiles [16]. This research was conducted to assist community health center management in developing preventative and data-driven policies, thereby supporting a healthier learning environment and avoiding the negative impacts of smoking. Furthermore, this research also aims to deepen understanding of the implementation of knowledge management and data mining in classifying students' lifestyle risks based on smoking behavior using the method Naive Bayes.

2. RESEARCH METHODOLOGY

2.1 Research Stages

Based on the problems described in the introduction, the researchers searched for references from various related literatures. Furthermore, the Naïve Bayes method was determined as the algorithm to be used in solving the classification problem of students' lifestyle risks based on smoking behavior [17]. The classification of modeling of students' lifestyle risks using the Naïve Bayes algorithm was conducted using RapidMiner tools. This model aims to obtain an accurate classification of student risk levels according to each student's smoking behavior.

This research applied two main stages in the research method, namely the data collection stage and the data processing stage. The data management stage was followed by the testing stage of the Naïve Bayes algorithm implementation in the classification process using RapidMiner tools. From the testing process, the analysis results and conclusions were obtained. In general, the steps in the research methodology can be seen in Figure 1.

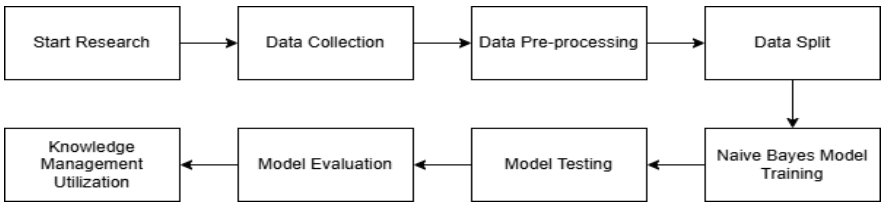

Figure 1. Research Flow

2.2 Data Collection Techniques

Primary data were collected through a structured questionnaire distributed to students at a vocational high school (SMK X) located within the coverage area of a local community health center. The questionnaire was designed to gather information on student demographics, smoking behavior (frequency and duration), and self-reported health conditions (cough, shortness of breath, chest pain). A total of 277 questionnaires were distributed to students in grades X and XI. Table 1 presents the complete list of variables captured in the questionnaire

Table 1. Questionnaire Results

Variables	Description	Data Types
Students Identity	Name, class, student number	Identity
Age	Respondent’s age (years)	Numeric
Gender	Male Female	Categorical
Smoker Status	Smoking / Non-Smoking	Categorical
Smoking Frequency	Number of cigarettes per day (if smoking)	Numeric

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 142−150
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9668

Variables	Description	Data Types
Smoking Duration	Duration of smoking (months/years)	Numeric
Family History of Smoking	Are there any family members who smoke / Are there any?	Categorical
Health Complaints	Cough, shortness of breath, chest pain, none	Categorical
Cigarette Smoke Exposure	Frequent exposure to cigarette smoke / Rarely / Never	Categorical
Health Status	Healthy / Unhealthy (based on complaints and assessment)	Categorical (Label)

2.3 Data Management Stage

Before training and testing using the Naïve Bayes method, the response data from the previous questionnaire underwent data management to select the most influential attributes in determining the classification of students' lifestyle risks. Therefore, a selection process was carried out to remove unnecessary attributes in the classification process [18]. Furthermore, data preprocessing was performed. This stage aimed to remove duplicate data, check data inconsistencies, and correct incorrect data. The selection and preprocessing processes were carried out using Microsoft Excel 2021. The available data consisted of 221 student data points from SMK X after going through the cleaning process.

2.4 Naïve Bayes Testing Stage with RapidMiner

At this stage, the classification process was carried out using the Naïve Bayes method. The equipment used was a laptop with adequate specifications. The tool used in this research was RapidMiner software. RapidMiner is one of the software often used for data mining processing [19]. Bayes' rule is a theorem in probability theory for calculating the probability of a hypothesis or event (H) based on observed evidence or data (X). Bayes' rule combines the initial probability of the hypothesis (before the evidence is observed) with the probability of the evidence under that hypothesis. The Naïve Bayes theorem uses a mathematical tool used in probability concepts to detect the possibility of a classification result; this concept is illustrated in the following formula [20]:

$$P(X) = \frac{P(H)\,.\,P(H)}{P(X)} \tag{1}$$

Bayes' theorem is a fundamental principle in probability theory and statistical inference used to update the belief of a hypothesis based on new data [21]. In this context, P(H|X) represents the posterior probability, which is the final probability that hypothesis H occurs given evidence X. P(X|H) indicates the probability that evidence X occurs will affect hypothesis H. P(H) is the prior probability, which is the initial probability of hypothesis H without considering evidence X. Furthermore, P(X) is the marginal probability, which is the total probability of evidence X occurring without considering the hypothesis. The dataset prepared in the previous stage underwent a classification process by testing the performance of the Naïve Bayes algorithm using RapidMiner tools. The dataset consisted of 221 data points with 7 attributes including the class label attribute (output attribute), namely health status [22]. Based on the available dataset, it was divided into 80% training data and 20% testing data. This study aims to obtain the accuracy results provided by the Naïve Bayes algorithm for a classification problem in predicting student lifestyle risks, whether they fall into the healthy or unhealthy class [23].

2.5 Analysis

After modeling was carried out by training on 80% of the data, the testing process was then carried out on the remaining 20% of the data, and the classification results for the testing data were obtained. To validate the classification results, a confusion matrix was used. The confusion matrix provides the performance level of the model based on correct or incorrect objects. To test the confusion matrix results displayed, manual calculations of Accuracy, Precision, and Recall were performed using the following calculations

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{4}$$

In the context of evaluating the performance of prediction or classification models, four types of results can occur, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) is the number of cases where the model predicts positive and the actual value is indeed positive. True Negative (TN) is the number of cases where the model predicts negative and the actual value is indeed negative. False Positive (FP) occurs when the model predicts positive but the actual value is negative. False Negative (FN) occurs when the model predicts negative but the actual value is positive. Accuracy is the value of how often a model makes correct predictions overall. Formula (2) calculates the proportion of all correct predictions (both positive and negative) (TP + TN) from the total overall predictions made. Precision is the value of how many of the positive predictions are truly positive. Formula (3) shows the calculation of the proportion of correct positive predictions (TP) to the total positive predictions made (TP + FP). Meanwhile, Recall is a measure of how well the model can detect all actual positive

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 142−150
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9668

cases. Formula (4) shows the calculation of the proportion of actual positive cases that are correctly detected (TP) to the total actual positive cases (TP + FN). In addition, the evaluation process employed the ROC Curve and the Area Under the Curve (AUC) to assess the model's ability to discriminate between classes accurately. The AUC value serves as a comprehensive indicator of classification performance, where higher values reflect better predictive capability. Specifically, an AUC ranging from 0.90 to 1.00 is categorized as excellent classification, indicating outstanding model performance. Values between 0.80 and 0.90 are considered good, while 0.70 to 0.80 reflects fair classification. Furthermore, an AUC of 0.60 to 0.70 is interpreted as poor, and values between 0.50 and 0.60 indicate classification failure overall.

3. RESULTS AND DISCUSSION

Based on the problem analysis at SMK X, data on the classification of students' lifestyle risks based on smoking behavior is needed to determine health interventions, with the final outcome being the grouping of students into two categories Unhealthy and Healthy. Health data were obtained from students' smoking behavior and observed physical conditions, while behavioral data were collected through a self-assessment questionnaire in which students rated their own smoking habits and perceived health conditions.

The questionnaire items were developed based on recommendations from healthcare workers at the local community health center and literature on adolescent smoking behavior. In line with the research objective, the respondents were all students of SMK X, and to facilitate data collection, the questionnaire was administered in printed form for students to fill out directly.

3.1 Research Dataset

The dataset used in this study was collected through questionnaires and interviews with students at a vocational high school within the coverage area of a local community health center. The dataset consists of 277 valid records after selection and cleaning. It includes variables such as smoking behavior, which is used to predict unhealthy lifestyle risks. The dataset was split into two parts: 80% for training and 20% for testing, ensuring that the model can be accurately tested on unseen data.

Table 2. Characteristics of the Research Dataset

Characteristics	Category	Amount	Percentage
Gender	Man	101	45.7%
	Woman	120	54.3%
Smoking Status	Smoking	41	18.55%
	No Smoking	180	81.45%
Health Status	Healthy	180	81.45%
	Not healthy	41	18.55%

3.2 Naïve Bayes Testing Using RapidMiner

At this stage, testing was conducted to determine the accuracy achieved by modeling using the Naïve Bayes method in classifying student lifestyle risk data. This modeling used RapidMiner Studio tools. The dataset was stored in Comma Separated Values (csv) file format to be readable by RapidMiner tools.

	NISN	Nama Siswa	Tanggal Lahir	nis Kelam	Merokok	enis Roko	rpapar As	erat Bada	inggi Bada	Status Giz	ngkar Peru	Sistolik	Diastolik	Status Kesehatan
2	0094000030	Siswa 31	2009-02-18	L	Ya	Rokok Kor	Ya	63	166	Normal	82	120	80	Tidak Sehat
3	0094000126	Siswa 127	2009-01-22	L	Tidak		Tidak	66	173	Normal	82	110	75	Sehat
4	0094000220	Siswa 221	2009-05-15	P	Tidak		Tidak	46	152	Normal	83	110	75	Sehat
5	0094000142	Siswa 143	2009-02-11	P	Tidak		Tidak	59	150	Normal	72	110	75	Sehat
6	0094000255	Siswa 256	2009-08-19	P	Tidak		Tidak	46	165	Normal	82	110	75	Sehat
7	0094000236	Siswa 237	2009-05-23	P	Tidak		Tidak	51	155	Normal	74	110	75	Sehat
8	0094000097	Siswa 98	2009-01-26	L	Tidak		Tidak	46	172	Normal	80	110	75	Sehat
9	0094000215	Siswa 216	2009-07-21	P	Tidak		Tidak	60	163	Normal	74	110	75	Sehat
10	0094000256	Siswa 257	2009-08-15	P	Tidak		Tidak	50	150	Normal	73	110	75	Sehat
11	0094000144	Siswa 145	2009-02-24	P	Tidak		Tidak	54	161	Normal	79	110	75	Sehat
12	0094000079	Siswa 80	2009-03-23	L	Tidak		Tidak	53	164	Normal	82	110	75	Sehat
13	0094000208	Siswa 209	2009-06-18	P	Tidak		Tidak	57	168	Normal	83	110	75	Sehat
14	0094000230	Siswa 231	2009-01-22	P	Tidak		Tidak	47	155	Normal	72	110	75	Sehat
15	0094000101	Siswa 102	2009-02-22	L	Tidak		Tidak	50	161	Normal	73	110	75	Sehat
16	0094000272	Siswa 273	2009-04-10	P	Tidak		Tidak	57	162	Normal	73	110	75	Sehat
17	0094000114	Siswa 115	2009-06-18	L	Tidak		Tidak	57	155	Normal	74	110	75	Sehat
18	0094000193	Siswa 194	2009-03-17	P	Tidak		Tidak	50	171	Normal	84	110	75	Sehat
19	0094000060	Siswa 61	2009-01-14	L	Ya	Rokok Kor	Ya	48	154	Normal	74	120	80	Tidak Sehat
20	0094000202	Siswa 203	2009-02-10	P	Tidak		Tidak	45	171	Normal	84	110	75	Sehat
21	0094000241	Siswa 242	2009-03-13	P	Tidak		Tidak	54	167	Normal	70	110	75	Sehat
22	0094000045	Siswa 46	2009-05-16	L	Ya	Rokok Kor	Ya	52	174	Normal	79	120	80	Tidak Sehat
23	0094000073	Siswa 74	2009-02-27	L	Tidak		Tidak	58	160	Normal	79	110	75	Sehat

Figure 2. Import file csv

An examination of the dataset revealed a class imbalance in the label attribute. The Healthy class contained the majority of instances with 180 records, whereas the Unhealthy class was underrepresented with only 41 records. This disparity is important to consider during model evaluation, as relying solely on accuracy may produce misleading conclusions about performance.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 142−150
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9668

Row No.	Status Kes...	Sekolah	Nama Seko...	Kelas	NIK	NISN	Nama Siswa	Tanggal La...	Jenis Kela...	Merokok	Jenis Rokok	Terpapar A...	Berat Bac
1	Tidak Sehat	Sekolah	SMKN 1 Pay	Kelas 12	161014100(	94000030	Siswa 31	2009-02-18	L	Ya	Rokok Konve	Ya	63
2	Sehat	Sekolah	SMKN 1 Pay	Kelas 12	161014100(	94000126	Siswa 127	2009-01-22	L	Tidak	?	Tidak	66
3	Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000220	Siswa 221	2009-05-15	P	Tidak	?	Tidak	46
4	Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000142	Siswa 143	2009-02-11	P	Tidak	?	Tidak	59
5	Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000255	Siswa 256	2009-08-19	P	Tidak	?	Tidak	46
6	Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000236	Siswa 237	2009-05-23	P	Tidak	?	Tidak	51
7	Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000097	Siswa 98	2009-01-26	L	Tidak	?	Tidak	46
8	Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000215	Siswa 216	2009-07-21	P	Tidak	?	Tidak	60
9	Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000256	Siswa 257	2009-08-15	P	Tidak	?	Tidak	50
10	Sehat	Sekolah	SMKN 1 Pay	Kelas 12	161014100(	94000144	Siswa 145	2009-02-24	P	Tidak	?	Tidak	54
11	Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000079	Siswa 80	2009-03-23	L	Tidak	?	Tidak	53
12	Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000208	Siswa 209	2009-06-18	P	Tidak	?	Tidak	57
13	Sehat	Sekolah	SMKN 1 Pay	Kelas 12	161014100(	94000230	Siswa 231	2009-01-22	P	Tidak	?	Tidak	47
14	Sehat	Sekolah	SMKN 1 Pay	Kelas 12	161014100(	94000101	Siswa 102	2009-02-22	L	Tidak	?	Tidak	50
15	Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000272	Siswa 273	2009-04-10	P	Tidak	?	Tidak	57
16	Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000114	Siswa 115	2009-06-18	L	Tidak	?	Tidak	57
17	Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000193	Siswa 194	2009-03-17	P	Tidak	?	Tidak	50
18	Tidak Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000060	Siswa 61	2009-01-14	L	Ya	Rokok Konve	Ya	48
19	Sehat	Sekolah	SMKN 1 Pay	Kelas 12	161014100(	94000202	Siswa 203	2009-02-10	P	Tidak	?	Tidak	45
20	Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000241	Siswa 242	2009-03-13	P	Tidak	?	Tidak	54
21	Tidak Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000045	Siswa 46	2009-05-16	L	Ya	Rokok Konve	Ya	52
22	Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000073	Siswa 74	2009-02-27	L	Tidak	?	Tidak	58
23	Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000181	Siswa 182	2009-03-10	P	Tidak	?	Tidak	49
24	Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000165	Siswa 166	2009-04-14	P	Tidak	?	Tidak	45
25	Sehat	Sekolah	SMKN 1 Pay	Kelas 11	161014100(	94000274	Siswa 275	2009-01-26	P	Tidak	?	Tidak	49
26	Sehat	Sekolah	SMKN 1 Pay	Kelas 10	161014100(	94000212	Siswa 213	2009-02-22	P	Tidak	?	Tidak	47

Figure 3. Dataset After Transformation

In the modeling process employing the Naïve Bayes algorithm, the dataset was first divided at an 80:20 ratio, resulting in 221 records designated for training and 56 for testing, with this split performed randomly to avoid any potential bias. From there, the training data was fed into the Naïve Bayes operator, where the algorithm learned from the examples, and the model produced from that learning stage was subsequently passed to the Apply Model operator. Meanwhile, the testing data was directed straight into the same Apply Model operator so that the model's predictions could be generated on unseen instances. After the model had run successfully, the complete output, containing those predictions was forwarded to the Performance operator, which then calculated various metrics to assess how well the model performed in classifying the data.

xampleSet (221 examples, 1 special attribute, 17 regular attributes) View Filter (221 / 221): all

Sekolah	SMKN 1 Pa...	Kelas 10	161014551...	0094382958	Anggi Safitri	2009-11-15	P	Tidak	J	K	60.0	166.0	Normal	85.0	70.0	103.0
Sekolah	SMKN 1 Pay	Kelas 10	1610160906	108439039	ARJUNA PU	2010-06-09	L	Ya	Rokok Konve	?	43	159	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610167010	107113505	Asti Asyifa Al	2010-10-30	P	Tidak	?	?	39	148	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610167003	118984751	Asyfah	2011-03-30	P	Tidak	?	?	48	156	Normal	85	70	103
Sekolah	SMKN 1 Pay	Kelas 10	1610163105	3105467632	CHOIRUL U	2010-05-31	L	Tidak	?	?	41	156	Normal	85	70	103
Sekolah	SMKN 1 Pay	Kelas 10	1610165601	107146535	DINA AULIA	2010-01-16	P	Tidak	?	?	56	165	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610164612	83488581	Firda Firdiya	2008-12-06	P	Tidak	?	?	56	165	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610165009	101992051	Fitri Anggrain	2010-09-10	P	Tidak	?	?	56	165	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610115207	108020310	Hani Ammar	2010-07-12	P	Tidak	?	?	53	162	Normal	85	70	126
Sekolah	SMKN 1 Pay	Kelas 10	1610146107	95714857	JULIANA	2009-07-21	P	Tidak	?	?	43	156	Normal	85	70	99
Sekolah	SMKN 1 Pay	Kelas 10	1603142211	91010949	M. DAYU SU	2009-11-22	L	Ya	Rokok Konve	Ya	69	157	Normal	85	70	107
Sekolah	SMKN 1 Pay	Kelas 10	1610162305	3105999371	M. FAHRIAN	2010-05-23	L	Tidak	?	?	69	163	Kurang	85	70	108
Sekolah	SMKN 1 Pay	Kelas 10	1610166204	91340799	Mar atussoli	2009-04-22	P	Tidak	?	?	46	154	Normal	85	70	112
Sekolah	SMKN 1 Pay	Kelas 10	1610161302	104491008	Rahmat Soli	2010-02-13	L	Ya	Rokok Konve	Ya	44	162	Normal	85	70	124
Sekolah	SMKN 1 Pay	Kelas 10	1610161302	102471960	Rahmat Soli	2010-02-13	L	Tidak	?	?	47	160	Normal	85	70	99
Sekolah	SMKN 1 Pay	Kelas 10	1603140102	101395389	Refky Andyka	2010-02-01	L	Tidak	?	?	56	165	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610146810	103547592	SASTIA	2010-10-28	P	Tidak	?	?	39	153	Normal	85	70	99
Sekolah	SMKN 1 Pay	Kelas 10	1603144712	91954722	TIARA	2009-12-07	P	Tidak	?	?	52	155	Normal	85	70	98
Sekolah	SMKN 1 Pay	Kelas 10	1610145812	99986353	ULVY RIYAN	2009-12-18	P	Tidak	?	?	51	153	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1603141705	102554557	AHMEID FUF	2010-05-17	L	Ya	Rokok Konve	Ya	40	157	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610052809	84211215	Aldi	2008-09-28	L	Tidak	?	?	43	166	Normal	85	70	129
Sekolah	SMKN 1 Pay	Kelas 10	1603143010	92126752	ALPIANSYAH	2009-10-30	L	Ya	Rokok Konve	?	56	165	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610144204	103954732	ANNUR IZZA	2010-04-02	P	Tidak	?	?	49	156	Normal	85	70	141
Sekolah	SMKN 1 Pay	Kelas 10	1610164411	107258481	Azzillia Nurul	2010-11-04	P	Tidak	?	?	64	165	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1603146012	95348863	CAHYA PRA	2009-12-20	P	Tidak	?	?	48	146	Normal	85	70	119
Sekolah	SMKN 1 Pay	Kelas 10	1603146206	101413975	ELCIA MITRI	2010-06-22	P	Tidak	?	?	44	153	Normal	85	70	106
Sekolah	SMKN 1 Pay	Kelas 10	1602181508	103585246	Encen	2010-08-15	L	Tidak	?	?	65	169	Normal	85	70	96
Sekolah	SMKN 1 Pay	Kelas 10	1671064512	3115244679	Fatimah Azz	2010-12-05	P	Tidak	?	?	46	156	Normal	85	70	116
Sekolah	SMKN 1 Pay	Kelas 10	1610165107	109703246	INTAN NUR	2010-07-11	P	Tidak	?	?	40	153	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610140708	92478090	JUNDRI KUI	2009-08-07	L	Ya	Rokok Konve	Ya	56	165	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1603036410	3109148386	LIVI SAGITA	2010-10-24	P	Tidak	?	?	55	159	Normal	85	70	101
Sekolah	SMKN 1 Pay	Kelas 10	1610166206	101022459	MELINDA	2010-06-22	P	Tidak	?	?	56	165	Normal	85	70	110
Sekolah	SMKN 1 Pay	Kelas 10	1610162707	109114303	MUHAMMAD	2010-07-27	L	Ya	Rokok Konve	?	44	163	Normal	85	70	88
Sekolah	SMKN 1 Pay	Kelas 10	1603142311	86156058	MUHAMMAD	2008-11-23	L	Tidak	?	?	54	166	Normal	85	70	110

Figure 4. Prediction results obtained with the naive Bayes algorithm

Table 3. Naive Bayes algorithm performance

	true Healthy	true UnHealthy	class precision
pred. Healthy	173	5	97,19%

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 142−150
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9668

	true Healthy	true UnHealthy	class precision
pred. UnHealthy	34	65	65,66%
class recall	83,57%	92,86%	

Looking at the confusion matrix, the model performed very well at identifying healthy students, correctly predicting 173 out of the healthy cases with a high precision of 97.19%. However, its recall for the healthy class was lower at 83.57%, meaning a fair number of healthy students were mistakenly labeled as unhealthy. For the unhealthy class, the model captured 65 out of the actual unhealthy cases, achieving a strong recall of 92.86%, but its precision dropped to 65.66%, indicating that many students predicted as unhealthy were actually healthy. This trade-off highlights the impact of the original class imbalance since healthy cases outnumbered unhealthy ones, the model tended to over-predict the minority class. Relying solely on accuracy would therefore mask these differences, making precision and recall essential for a proper evaluation.

3.3 Classification Results and Model Evaluation

From the modeling results, the Naïve Bayes algorithm achieved excellent performance in performing binary classification on the student lifestyle risk dataset. The resulting confusion matrix is presented in Table 3.

Table 4. Data Cleaning Results

	Prediction: Healthy	Prediction: Unhealthy
Actual: Healthy	173	5
Actual: Unhealthy	34	65

The dataset, consisting of 277 students, was divided into training and testing subsets to facilitate model development and evaluation. Specifically, 221 students, representing 80% of the total data, were allocated to the training set, while the remaining 56 students, or 20%, formed the test set. This division was performed randomly to avoid any potential bias in the subset composition. Class distribution for each of these subsets is detailed in Table 4.

Table 5. Class Composition in Training Data and Test Data

Class	Training Data (n=221)	Test Data (n=56)	Total (n=277)
Healthy	166 (75.1%)	41(73.2%)	207(74.7%)
Not healthy	55 (24.9%)	15 (26.8%)	70 (25.3%)

The trained model was then tested using test data (56 students). The predicted results were compared with the actual values and presented in the confusion matrix in Table 5.

Table 6. Confusion Matrix of Naïve Bayes Classification Results

	Prediction: Healthy	Prediction: Unhealthy
Actual: Healthy	173 (TN)	34 (FP)
Actual: Unhealthy	5 (FN)	65 (TP)

Based on Table 5, the implemented Naïve Bayes classification model produced the following prediction distribution. Out of a total of 277 test data samples, there were 173 data points with an actual *Healthy* category that were correctly predicted as *Healthy* (True Negative, TN), and 65 data points with an actual Unhealthy category that were correctly predicted as *Unhealthy* (True Positive, TP). Meanwhile, the model made incorrect predictions on 34 samples where actually healthy students were predicted as Unhealthy (False Positive, FP), and 5 samples where actually unhealthy students were predicted as *Healthy* (False Negative, FN).

This, the Naïve Bayes model demonstrates good performance in identifying the Unhealthy class, with a relatively high True Positive value (65) compared to a very low False Negative value (5). However, the considerably high number of False Positives (34) indicates that the model tends to be overly sensitive in predicting the Unhealthy class, resulting in several healthy students being misclassified as unhealthy. This condition needs to be further considered depending on the application objectives of the system, whether it is more important to minimize false negatives (the risk of undetected unhealthy conditions) or to reduce false positives (avoiding incorrect labeling of healthy students). Based on the values inconfusion matrix, the evaluation metrics are calculated as follows:

a. Accuracy

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

$$Accuracy = \frac{65+173}{65+173+34+5} x\ 100\%$$

Accuracy = 0,8592= 85.92%

From these calculations, the accuracy value of the classification test results using the algorithm Naive Bayes is 85.92%.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 142−150
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9668

b. Precision

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

$$Precision = \frac{65}{65+34}\ x\ 100\%$$

Precision= 0,6565 = 65.66%

From these calculations, the precision value of the classification test results of the Algorithm Naive Bayes has a success rate of 65.66%.

c. Recall

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

$$Recall = \frac{65}{65+5}\ x\ 100\%$$

Recall= 0,9285= 92.86%

From this calculation, the value recall Algorithm classification test results of Naive Bayes have a success rate of 92.86%.

d. Performance Vector

From the evaluation of classification using the algorithm Naïve Bayes produces accuracy values that can be seen in Performance Vector below this.

PerformanceVector

```
PerformanceVector:
accuracy: 86.07% +/- 23.98% (mikro: 85.92%)
ConfusionMatrix:
True:   Sehat   Tidak Sehat
Sehat:  173     5
Tidak Sehat:    34      65
AUC (optimistic): 1.000 +/- 0.000 (mikro: 1.000) (positive class: Tidak Sehat)
AUC: 0.500 +/- 0.000 (mikro: 0.500) (positive class: Tidak Sehat)
AUC (pessimistic): 0.763 +/- 0.364 (mikro: 0.763) (positive class: Tidak Sehat)
precision: 86.12% +/- 27.77% (mikro: 65.66%) (positive class: Tidak Sehat)
ConfusionMatrix:
True:   Sehat   Tidak Sehat
Sehat:  173     5
Tidak Sehat:    34      65
recall: 92.86% +/- 21.43% (mikro: 92.86%) (positive class: Tidak Sehat)
ConfusionMatrix:
True:   Sehat   Tidak Sehat
Sehat:  173     5
Tidak Sehat:    34      65
```

Figure 5. Performance Vector

3.4 Utilization of Knowledge

The classification results obtained were used within a knowledge management framework to support decision-making at schools and community health centers. By utilizing the model results, schools can design more targeted prevention programs for students at risk of smoking. Additionally, these results assist community health centers in planning data-driven interventions. Overall, using these classification results enhances the effectiveness of evidence-based policies and prevention programs.

Table 7. Knowledge Management Integration Framework

Dimension	Input from Classification	Knowledge Output	Action
Health Sector	List of Unhealthy students	Risk profile database	Design early intervention: smoking cessation counseling, scheduled health check-ups
Education Sector	Behavioral patterns by class/grade	Insight on smoking prevalence	Formulate smoke-free area policies; conduct targeted outreach
Information Systems	Classification results (updated periodically)	Institutional knowledge base	Continuous KDD cycle: collect → process → classify → update

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 142−150
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9668

Dimension	Input from Classification	Knowledge Output	Action
Organizational Knowledge	Interpretation by health workers, teachers, researchers	Collective understanding of adolescent health factors	Evidence-based policy formulation at school and community health center level

This framework demonstrates that the value of classification results extends beyond technical performance metrics. By systematically channeling prediction outputs into institutional processes, schools and community health centers can translate data-driven insights into concrete preventive actions. Students classified as Unhealthy are flagged for follow-up interventions, while aggregate trends inform policy decisions on a broader level. The KDD cycle embedded in this framework ensures that the knowledge base remains current: as new student cohorts are enrolled and new data is collected, the model can be retrained and the institutional knowledge base updated accordingly. This creates a sustainable, evidence-based health management system within the school environment.

4. CONCLUSION

This study implemented the Naïve Bayes algorithm to classify students' lifestyle risks based on smoking behavior at SMK X using the Knowledge Discovery in Databases (KDD) framework, including data collection, pre-processing, attribute selection, data splitting, modeling, testing, and evaluation. The dataset consisted of 277 student records, divided into training data of 221 students (80%) and test data of 56 students (20%). Using RapidMiner software, the model achieved an accuracy of 85.92%, precision of 86.12%, and recall of 92.86% for the unhealthy class. These results indicate that the Naïve Bayes algorithm is reasonably effective in classifying students' health status based on smoking behavior and related health complaints. Beyond technical performance, this study integrates predictive data mining with knowledge management to support decision-making in both educational and health sectors. The classification results enable early detection of at-risk students, facilitating targeted interventions such as smoking cessation counseling, regular health check-ups, and the formulation of smoke-free area policies within schools. Future research should expand to larger and more diverse datasets from multiple schools, implement k-fold cross-validation for more robust evaluation, and compare multiple algorithms such as Decision Tree, Random Forest, and Support Vector Machine (SVM) to determine the most optimal model for lifestyle risk classification.

REFERENCES

[1] F. Ramadhan, D. Herlambang, A. P. Dipta, U. Bina, and S. Informatika, “Prediction of Health Status Based on Lifestyle Using Decision Tree and Feature Importance," RIGGS: Journal of Artificial Intelligence and Digital Business, vol. 4, no. 4, pp. 9616–9623, 2026, doi : 10.31004/riggs.v4i4.5246.
[2] B. Artini, Intiyaswati, and M. N. Pandeirot, "Perilaku Merokok pada Remaja di SMK Kota Surabaya," Jurnal Pengabdian Masyarakat, vol. 5, no. 1, pp. 27–32, 2023, doi: 10.47560/pengabmas.v5i1.609.
[3] I. M. D. Maysanjaya, Buku Ajar. Data Mining. Undiksha Press, 2022.
[4] A. F. Riany, G. Testiana, S. S. Informasi, “Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes,” Jurnal Teknologi Informasi, vol. 9, pp. 42–54, 2023, Accessed: 12 March 2026. [Online]. Available: https://jurnal.univpgri-palembang.ac.id/index.php/JurnalTeknologiInformasi.
[5] T. S. Kumar, *Introduction to Data Mining 1st ed*. Pearson Education, 2006, Accessed: 12 March 2026. [Online]. Available: https://www.amazon.ca/dp/9332571406/
[6] A. Pratama, “Implementasi Algoritma Naive Bayes Untuk Memprediksi Cuaca,” Jurnal Informatika dan Teknologi, vol. 8, no. 2, pp. 1637–1642, 2024, doi: 10.36040/jati.v8i2.8967.
[7] F. Sirait *et al.*, “Penerapan Naive Bayes untuk Identifikasi Keterlambatan Perkembangan Anak Berdasarkan Data Kesehatan pada Program Studi Kebidanan” Jurnal Media Informatika (JUMIN), vol. 6, no. 2, pp. 739–745, 2024, doi :10.62027/sevaka.v2i4.525.
[8] M. Samuel. Idmi, and Triyono, “Analisis perbandingan model naïve bayes dan c4.5 untuk prediksi stroke berdasarkan riwayat data medis dengan pendekatan matriks korelasi,” Jurnal Ilmiah Informatika, vol. 10, no. 4, pp. 3749–3759, 2025, doi:10.29100/jipi.v10i4.8653.
[9] A. Muttakin, Rusmana, and Ramadhani, "Komparasi Algoritma Decision Tree, Random Forest, SVM, dan KNN untuk Prediksi Penyakit Jantung," Jurnal Informatika, vol. 1, no. 2, pp. 35 42, 2025, doi: 10.15294/eduel.v13i1.22163.
[10] Y. D. Amritha, N. Luh, P. Ika, and W. P. Dananjaya, “Model Machine Learning yang Dioptimalkan untuk Prediksi Penyakit Jantung Menggunakan R Shiny,” Jurnal Komputer dan Sains Terapan, vol. 8, no. 01, pp. 1–10, 2026, doi: 10.53863/kst.v8i01.1994
[11] A. Wantoro et al., “Analisis Komparatif Strategi Penanganan Imbalanced Data pada Klasifikasi Penyakit Diabetes Menggunakan Data Mining," Jurnal Simpul Inovasi, no. 2, 2025, doi: 10.20884/1.jsi.2025.2.1.16198
[12] M. Husaini, Priyanto, and Martono, "Analisis Sentimen Kinerja Tenaga Medis Indonesia Menggunakan Modeling RoBERTa dan Metode Machine Learning," Jurnal Edukasi Elektro, vol. 13, no. 1, pp. 1–8, 2026. accessed: 13 march 2026. [Online]. Available: https://journal.unnes.ac.id/journals/eduel/index.
[13] F. Itsnani *et al.*, “Klasifikasi Risiko Kesehatan Berbasis Data Perilaku Remaja,” Jurnal Riset Teknik Komputer, vol. 2, no. 4, pp. 55–63, 2025, doi : 10.69714/q0zapc82.
[14] P. Widodo, “Analisis Kinerja Algoritma Naive Bayes dalam Klasifikasi Data pada Pasien Tuberkulosis Berbasis Data Mining,” Jurnal Online Gita Berbasis Teknologi dan Cara, vol. 5, no. 1, pp. 75–81, 2025, doi: 10.47065/jogtc.v5i1.8999.
[15] W. Fadri, “Klasifikasi Penyakit Hati dengan Menggunakan Metode Naive Bayes,” Jurnal Informasi dan Teknologi, vol. 5, no. 1, pp. 32–36, 2023, doi: 10.37034/jidt.v5i1.230.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 142−150
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9668

[16] F. A. Sumantri and Y. H. Chrisnanto, "Prediksi Risiko Kesehatan Mental Mahasiswa Menggunakan Klasifikasi Naive Bayes," Jurnal Ilmiah Komputasi, vol. 12, no. 3, pp. 383–393, 2025, doi: 10.30865/jurikom.v12i3.8648.

[17] I. T. Monowati and R. Setyadi, "Penerapan Algoritma Naïve Bayes Dalam Memprediksi Pengusulan Penghapusan Peralatan dan Mesin Kantor," Journal of Software Engineering, Information and Communication Technology, vol. 4, no. 2, pp. 483–491, 2023, doi: 10.47065/josh.v4i2.2674.

[18] P. Rahmawati and A. Larasati, "Pengembangan Model Persetujuan Kredit Nasabah Bank Dengan Algoritma Klasifikasi Naïve Bayes , Decision Tree , Dan Artificial Neural Network," Jurnal Sistem Informasi, vol. 17, no. 1, pp. 1–12, 2022, doi: 10.14710/jati.1.1.1-12.

[19] D. Florencia, "Prediksi Jenis Kesehatan Kejiwaan Berdasarkan Usia Menggunakan Metode Naïve Bayes Berbasis Website," Seminar Nasional Teknologi Informasi, vol. 8, pp. 15030–15040, 2024, Accessed: 12 March 2026. [Online]. Available: https://paperity.org/p/358147184. v

[20] D. R. Andriyani, M. Afdal, and S. Monalisa, "Analisis Sentimen Masyarakat Terhadap Penghapusan Honorer Berdasarkan Opini Dari Twitter Menggunakan Naïve Bayes Classifier," Building of Informatics, Technology and Science (BITS), vol. 5, no. 1, pp. 49–58, 2023, doi: 10.47065/bits.v5i1.3541.

[21] J. P. Tanjung, F. C. Tampubolon, A. W. Panggabean, and M. Anjas, "Customer Classification Using Naive Bayes Classifier With Genetic Algorithm Feature Selection," Jurnal dan Penelitian Teknik Informatika, vol. 7, no. 1, pp. 584–589, 2023, doi: 10.33395/sinkron.v8i1.12182.

[22] S. T. Utami, S. Lestari, and H. W. Nugroho, "Prediction Of Anemia Using The Particle Swarm Optimization (PSO) And Naïve Bayes Algorithm," Computer Engineering and Informatics Journal, vol. 3321, no. X, pp. 1–8, 2024, doi: 10.24014/coreit.v10i1.28428.

[23] S. Andriyanto and M. S. Hasibuan, "Application of Nave Bayes Algorithm for SMS Spam Classification Using Orange," International Journal of Artificial Intelligence and Software Computing Applications, vol. 1, 2022, doi: 10.47679/ijasca.v1i1.3.