SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

Jakarta Air Quality Classification Based On Air Pollutant Standard Index Using C4.5 And Naïve Bayes Algorithms

Duta Pramudya Ramadhan[1], Agung Triayudi[2*]

Informatics Study Program, Faculty of Communication and Information Technology, Universitas Nasional, Jakarta, Indonesia

Author Email: dutaprmdy281@gmail.com[1], agungtriayudi@civitas.unas.ac.id[2*]

Abstract. Increasing air pollution in DKI Jakarta has become an increasingly pressing environmental issue, which has a direct impact on public health and environmental sustainability. Therefore, it is very important to have a system that manages data-based air pollution levels. The purpose of this research is to classify air quality in DKI Jakarta through Air Pollutant Standard Index (ISPU) data. This data consists of parameters such as dust particles (PM10, PM2.5), sulfur dioxide (SO2), carbon monoxide (CO), surface ozone (O3), and nitrogen dioxide (NO2), as well as two classification algorithms used, namely C4.5 and Naïve Bayes. This research also seeks to compare the effectiveness of the two algorithms based on ISPU data collected in 15 Jakarta areas. The approach used in this research is to divide the data using three ratio scenarios, namely 70% : 30%, 80% : 20%, and 90%: 10%. In addition, performance assessment is carried out using accuracy, precision, recall and f1 score metrics. The experimental results showed better performance of C4.5, with an average accuracy of 95%, precision of 99%, recall of 94% and f1-score of 97%. In contrast, Naïve Bayes recorded an average accuracy of 81%, precision of 93%, recall of 73% and f1-score of 82%. These findings corroborate the validity of the C4. 5 algorithm is more effective in air quality classification based on ISPU, thus making it a reliable resource for air quality monitoring and management in DKI Jakarta, as well as supporting decision-making in air pollution control policies.

Keywords: Air Pollution, DKI Jakarta, Classification, C4.5, Naive Bayes.

1 Introduction

Air consists of natural gases such as nitrogen, oxygen, argon and carbon dioxide, which play an important role in the life of organisms. Even though it cannot be seen, air can be felt and clean air provides benefits, including respiratory health, reducing the risk of chronic disease, prolonging life, increasing stamina, concentration and mood [1]. Air pollution occurs when physical, chemical or biological pollutants enter the air at levels that endanger the health of humans, animals and plants. WHO reports that air pollution causes millions of deaths every year, damages the environment, especially in large urban areas such as DKI Jakarta, and triggers health problems such as respiratory and cardiovascular diseases. Traffic, industry, and waste management are the main contributors to air pollution which damages air quality, ecosystems, and threatens health in the long term [2]. Air quality is influenced by atmospheric conditions and other factors. Air pollution is measured using the Air Pollutant Standard Index (ISPU), which provides reliable information for the public and a reference for the government in managing air pollution [3].

The air contains pollutants such as dust particles (PM10 and PM2.5), sulfur dioxide (SO2), carbon monoxide (CO), surface ozone (O3), and nitrogen dioxide (NO2), which can endanger public health. Therefore, it is important to measure and assess air quality to support government policies [4]. The DKI Jakarta Provincial Environmental Service uses data mining to process big data, detecting patterns, trends and useful relationships. This method combines techniques from machine learning, signal processing, statistics, and spatial and temporal data analysis [5]. Air quality in DKI Jakarta can be measured using classification algorithms in data mining, such as C4.5 and Naïve Bayes, to organize data into certain categories [6]. The C4.5 algorithm is often used in data extraction because of its ease of understanding, ability to handle numeric and discrete parameters, and produce decision trees with rational accuracy [7]. Apart from that, this research also uses the Naïve Bayes algorithm, which is based on probability and statistics to determine the probability of an event in the future based on previous experience. Naïve Bayes separates features from one class to another [8]. The Naïve Bayes algorithm was chosen because of its ability to process large datasets, high computational efficiency, and its resistance to irrelevant attributes [9].

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

Previous studies have used various classification algorithms to assess ISPU levels in various geographic regions. For example, research in South Tangerang in 2022 used the K-Nearest Neighbor (KNN) and Naïve Bayes algorithms to predict air pollution, with KNN accuracy results of 94.44% and Naïve Bayes 86.11%. Another study in DKI Jakarta in 2023 used Naïve Bayes with accuracy results of 91.96%. In Bandung, in 2024, research examined three algorithms such as Naïve Bayes (87.50%), KNN (85%), and Support Vector Machine (SVM) (92.50%). In the Special Region of Yogyakarta, in 2024, the C4.5 algorithm shows a high accuracy of 99.94%. Another study in large Indonesian cities in the same year used KNN and Naïve Bayes, with KNN accuracy of 95.13% and Naïve Bayes 95.97%. However, these five studies have limitations, especially related to the amount and representativeness of the data used, which can affect the validity and accuracy of the classification results. Therefore, this research aims to collect data that is larger, more representative and relevant to real situations to increase accuracy and overcome these obstacles. It is hoped that the findings of this study can serve as a reference for the government, environmental agencies, and the public in addressing air pollution. With a better understanding of pollution factors, more effective and targeted policies can be formulated to reduce the impact of pollution, increase public awareness, and maintain ecosystem health and sustainability.

2 Methodology

Before presenting the research findings, a flowchart of the research method is displayed showing the procedure from data collection and processing to the use of classification algorithms for air quality analysis. Figure 1 provides a clear picture of the flow of the research conducted.

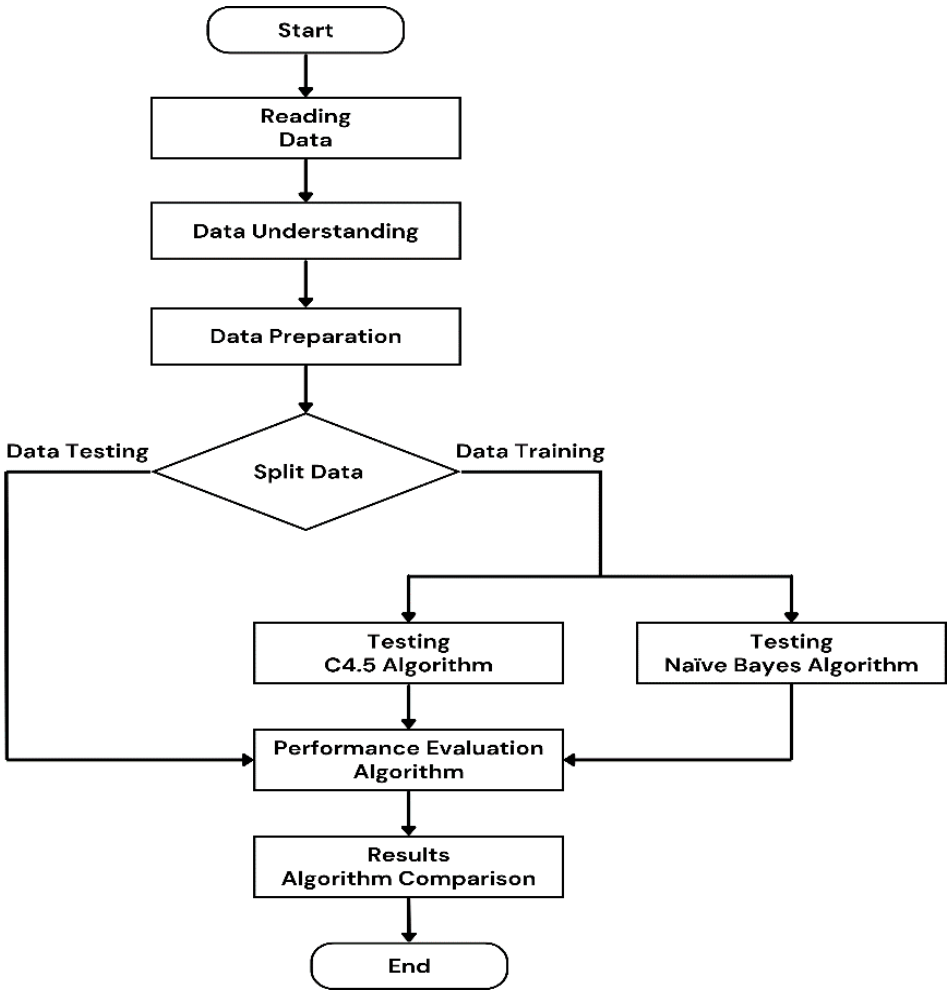

Figure 1. Research Flow

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

2.1 Reading Data

The algorithm testing design process begins by reading a dataset that has 10.800 data based on the Air Pollutant Standard Index (ISPU) Status. ISPU is a unitless index that describes air quality, taking into account its impact on health, aesthetics, and living organisms. ISPU is also the official air quality standard in Indonesia [10]. This dataset contains various similar attributes that become indicators of the process results, including location, time, dust particles (PM10), dust particles (PM2.5), sulfur dioxide (SO2), carbon monoxide (CO), surface ozone (O3), nitrogen dioxide (NO2), and ispu_status. Based on the ISPU values in the dataset, Table 1 presents information on the level of pollution and its impacts.

Table 1. ISPU Category

Range	Pollution Level	Impact
0 - 50	Good	No impact on human, animal, plant, building or aesthetic health.
51 - 100	Medium	Does not affect human or animal health, but may affect sensitive plants and aesthetics.
101 - 199	Unhealthy	Harmful to humans and sensitive animals and may have consequences for aesthetics and damage plants.
200 - 299	Very Unhealthy	Health threat to specific groups of individuals.
300 - 500	Harmful	May cause serious health impacts to the general population.

2.2 Data Understanding

To identify the type of data that can later be used to classify air quality in DKI Jakarta, the data collected will be examined and understood at this stage. Initial data collection and description, and data exploration are the two stages of data understanding.

a. Data Collection and Description
   In implementing the first stage of the data collection strategy, data was collected from the Jakarta Rendah Emisi and Udara Jakarta website. The next stage is to describe the planning data obtained previously after obtaining the initial planning data.
b. Data Exploration
   The next stage is data exploration, in which a more detailed explanation of the data content for each attribute will be given.

2.3 Data Preparation

The initial data collection process has been completed and a deeper understanding of the data has been obtained. The next stage is to examine the data using data mining procedures. The purpose of this analysis is to transform useless data into useful information so that important decision-making processes can be carried out more efficiently. This part involves accomplishing three tasks such as data selection, data cleaning, and data formatting.

a. Data Selection
   Some attributes were selected during the data collection planning process because they were thought to affect the air quality status (ispu_status). The data analysis will center on these attributes to determine their relation to air quality.
b. Data Cleaning
   In an effort to obtain the best classification, this step ensures the quality of the data used. As a step to handle data that has redundant information, missing values, and inconsistencies, existing data will be processed. One method to handle data that has missing values is to remove it from the existing data set.
c. Data Formatting
   The data formatting stage is the last action in the preparation stage. This stage creates the final data set that is prepared for processing with data mining.

2.4 Split Data

In this study, the acquired data is divided into three scenarios for evaluation purposes. The first scenario uses 70% of the data as training to assist the algorithm in identifying patterns, while the remaining 30% is used as testing data to evaluate the performance of the algorithm. In the second context, 80% of the data is used for training

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

and 20% for testing. On the other hand, in the third context, 90% of the data is focused on training and 10% of the data on testing to achieve more detailed evaluation results.

2.5 Testing C4.5 Algorithm

At this stage, the air quality classification process in DKI Jakarta is carried out by utilizing the C4.5 algorithm. The C4.5 algorithm is applied to classify data with numerical and categorical features, aiming to classify discrete attributes based on the data obtained [11]. The first step is to collect training data categorized by class, then find the root of the tree by calculating entropy and information gain. The process continues until each record has an ideal partition or stops if the same class is applied to all records, empty attributes, or empty branches [12]. The decision tree is built by calculating entropy and information gain, forming a hierarchical structure with root, internal, and leaf nodes. Entropy measures the homogeneity of the data, while information gain selects the testing attribute at each node [13]. Before calculating the information gain, use Equation 1 to calculate the entropy value:

$$Entropy\ (S) = -\sum_{i=1}^{k} P_j \times log_2 P_j \qquad (1)$$

Description:
S = The dataset to be analyzed.
A = Total number of classes in the dataset S.
Pj = Probability of each class in S.

Furthermore, Equation 2 is used to determine the information gain value, which will provide the results needed to determine which attributes are most relevant in the decision tree process:

$$Gain(S, A) = entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times entropy(S_i) \qquad (2)$$

Description:
S = The dataset to be analyzed.
A = Attributes in the data.
$|si|$ = The amount of data in A_i.
$|s|$ = Total number of data in *S*.
k = Total number of attribute values.

2.6 Testing Naïve Bayes Algorithm

The next stage is the Naïve Bayes algorithm utilized in classifying air quality in DKI Jakarta. The Naïve Bayes method is based on the assumption of conditional independence between characteristics when the class is known, making it easy to calculate the probability of each feature individually [14]. This probabilistic clustering combines frequencies and numbers from the available data [15]. Assuming that variables are independent, the Naïve Bayes approach uses Bayes' Theorem to calculate the probability of each class for each attribute [16]. The calculation procedure uses Gaussian Naïve Bayes through Equation 3:

$$P(x_i|X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma}} \qquad (3)$$

Description:
$P(x_i|X)$ = Probability of a feature x_i given a class X.
μ = Average of features x_i in the class X.
σ = Variance of features x_i in the class X.
e = Base of natural logarithm (2,71828).

2.7 Performance Evaluation Algorithm

When testing is complete, the evaluation matrix is usually displayed as a confusion matrix. In machine learning, the confusion matrix is a table that describes and assesses how well the categorization performs on test data that has labels [17]. The data is separated into two categories by the confusion matrix, namely the actual

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

classification results and the analysis results obtained by the system [18]. A more detailed explanation is given in table 2.

Table 2. Confusion Matrix

Actual	Classification Negative	Classification Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

The following is an explanation of Table 2.3 *confusion matrix*, which provides a deeper understanding of the classification performance evaluation applied in this study: [19]

a. True positive, refers to data that is truly positive and the model also recognizes the data as positive.
b. True negative, refers to data that is part of the negative class, and which the model can classify as negative.
c. False positive, refers to data that is actually in the negative class, but the model mistakenly recognizes it as positive.
d. False negative, refers to data that should be classified as positive, but the model mistakenly identifies it as negative.

The next step is to continue evaluating the performance of the algorithm, where there is a relationship between evaluation measures such as accuracy, precision, recall, and f1-score with the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. Each statistic is calculated differently: [20]

a. Accuracy describes how precisely the system can classify determined by Equation 4:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (4)$$

b. Precision is the proportion of positive category data that can be accurately classified by the system determined by Equation 5:

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (5)$$

c. Recall is the percentage of positive category data that is accurately classified by the system, compared to all true positive data determined by Equation 6:

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (6)$$

d. F1-score is a matrix with the harmonic mean between precision and recall determined by Equation 7:

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \times 100\% \quad (7)$$

2.8 Results Algorithm Comparison

The results of air quality analysis in DKI Jakarta are displayed through the application of algorithms. The test data utilized includes various variables related to air pollution attributes. The accuracy analysis of the two algorithms will be described in detail, including a comparison of their performance in producing accurate classifications.

3 Results and Discussion

In this section, the results of Jakarta air quality classification based on Air Pollutant Standard Index using C4.5 And Naïve Bayes algorithms will be presented. The purpose of this study is to compare the ability of the two algorithms to categorize air quality in Jakarta based on the categories set in the ISPU.

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

3.1 Reading Data

Data collection on air quality in DKI Jakarta, obtained from Jakarta Rendah Emisi and Udara Jakarta website, resulted in a total of 10.800 data points. The data contains a number of shared attributes, including location, time, pm10, pm2.5, so2, co, o3, no2, and ispu_status. Table 3 presents 10 examples of the overall data collected.

Table 3. Air Quality Data

location	time	pm10	pm2.5	so2	co	o3	no2	ispu_status
dki1	9/1/2024 0:00	-	-	-	-	-	-	-
dki1	9/1/2024 1:00	-	-	-	-	-	-	-
dki1	9/1/2024 1:59	-	-	-	-	-	-	-
dki1	9/1/2024 3:00	55	71	14	23	28	37	Medium
dki1	9/1/2024 4:00	54	71	14	23	28	37	Medium
dki2	9/1/2024 3:00	58	75	58	7	33	18	Medium
dki2	9/1/2024 4:00	58	75	58	7	33	17	Medium
dki3	9/18/2024 6:00	48	45	38	7	34	36	Good
dki4	9/11/2024 8:00	74	82	55	22	26	17	Medium
dki5	9/5/2024 14:00	36	102	35	20	40	14	Unhealthy

Table 3 shows the locations of the 15 air quality monitoring points in DKI Jakarta, which are divided into five zones namely dki1, dki2, dki3, dki4, and dki5. A more comprehensive analysis for each region is presented below:

a. dki1 includes the areas of Street Bundaran HI, Street Cempaka Putih, and Gambir Station representing the Central Jakarta area.
b. dki2 covers the areas of Street Kelapa Gading, Penjaringan Flats, and Jakarta International Stadium representing North Jakarta.
c. dki3 covers the areas of Street Jagakarsa, Street Pasar Minggu, and Street Fatmawati representing South Jakarta.
d. dki4 includes the areas of Street Lubang Buaya, GOR Ciracas, and Delonix Park representing East Jakarta.
e. dki5 covers the areas of Street Kebon Jeruk, Kota Tua, and Kalideres Terminal representing West Jakarta.

3.2 Data Understanding

Data on air quality in DKI Jakarta was collected from the Jakarta Low Emission and Jakarta Air website, which includes 10,800 data with 9 attributes, such as location, time, pm10, pm2.5, so2, co, o3, no2, and ispu_status, over a 24-hour time period in September 2024. The first step in the data understanding process is data collection and description to understand the properties of the dataset before classification. Table 4.2 describes the attributes collected.

Table 4. Data Attributes

Attributes	Type	Description
location	Object	Geographical location where the air quality assessment was conducted
time	Object	Date and time of air quality data collection
pm10	Object	Concentration of dust particles less than 10 micrometers in size
pm2.5	Object	Concentration of dust particles less than 2.5 micrometers in size
so2	Object	Airborne sulfur dioxide levels
co	Object	Airborne carbon monoxide levels
o3	Object	Airborne surface ozone level
no2	Object	Nitrogen dioxide levels in the air
ispu_status	Object	ISPU status indicates air quality based on pollutant concentrations

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

Conducting additional data investigation was done next after completing the data collection and description stage. After that, a more thorough explanation of the content of each attribute or exploration in the acquired data will be given in Table 5.

Table 5. Data Exploration

Attributes	Description	Data Exploration
location	Geographical locations include dki1, dki2, dki3, dki4, and dki5.	Total amount of 10.800 data
time	Data collected from September 1 to 30, 2024	Total amount of 10.800 data
pm10	Dust particle concentration less than 10 micrometers	Total amount of 10.800 data
pm2.5	Dust particle concentration less than 2.5 micrometers	Total amount of 10.800 data
so2	Airborne sulfur dioxide levels	Total amount of 10.800 data
co	Airborne carbon monoxide levels	Total amount of 10.800 data
o3	Airborne surface ozone levels	Total amount of 10.800 data
no2	Airborne nitrogen dioxide levels	Total amount of 10.800 data
ispu_status	ISPU status indicates air quality based on pollutant concentrations: a. Good b. Medium c. Unhealthy d. Very Unhealthy e. Harmful	ispu_status air quality: a. Value "-" = 3.615 b. Good = 650 c. Medium = 3.768 d. Unhealthy = 2.767 e. Very Unhealthy = 0 f. Harmful = 0 Total amount of 10.800 data

3.3 Data Preparation

After collecting and understanding the data, the next step is to prepare it for the data extraction procedure, which involves data selection, cleaning and formatting. Attributes that are expected to affect air quality are selected, while time features are removed, as they are only used for data identification. The C4.5 and Naïve Bayes algorithms will be used to handle attributes such as location, pm10, pm2.5, so2, co, o3, no2, and ispu_status. Classification will be based on air quality data from DKI Jakarta during September 2024, with three values for ispu_status, namely Good, Medium, and Unhealthy, are selected from the original dataset, as shown in Table 4.

Table 6. Temporary Attributes

No	Atribut
1	location
2	time
3	pm10
4	pm2.5
5	so2
6	co
7	o3
8	no2
9	ispu_status

The second stage, data cleaning, aims to ensure that only high-quality data is used for classification. Some missing values were detected in the collected data, as illustrated in Figure 2.

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

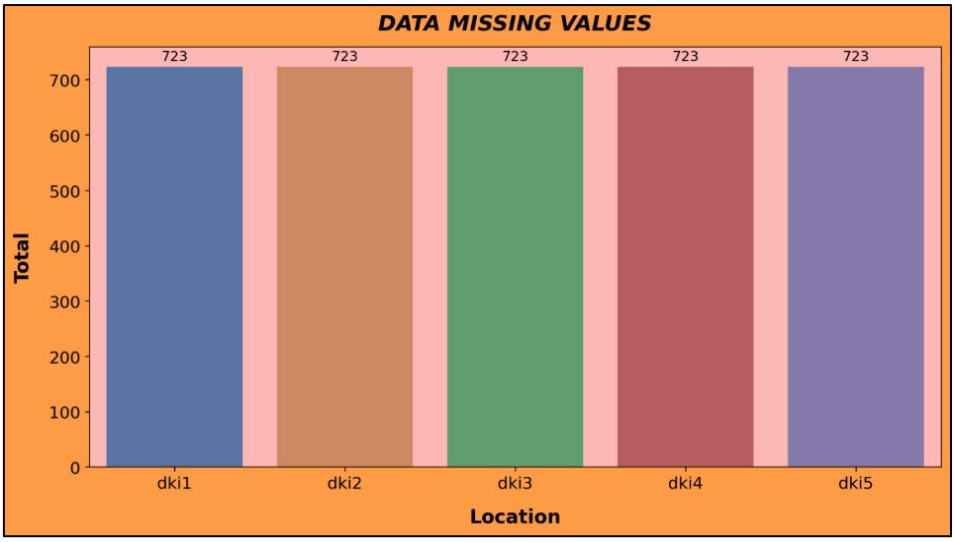

Figure 2. Data Missing Values

As a result, the data will be processed to handle the missing values. The data cleaning process will identify 3.615 "-" values and convert them to NaN, as shown in Table 7.

Table 7. Data Change

location	time	pm10	pm2.5	so2	co	o3	no2	ispu_status
dki1	9/1/2024 0:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dki1	9/1/2024 1:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dki1	9/1/2024 1:59	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dki1	9/1/2024 3:00	55	71	14	23	28	37	Medium
dki1	9/1/2024 4:00	54	71	14	23	28	37	Medium

Row deletion occurs after the data changes to NaN. Table 8 shows the air quality data in DKI Jakarta for the September 2024 period after cleaning, which is 7.185 data.

Table 8. Data Cleaning

s	time	pm10	pm2.5	so2	co	o3	no2	ispu_status
dki1	9/1/2024 3:00	55	71	14	23	28	37	Medium
dki1	9/1/2024 4:00	54	71	14	23	28	37	Medium
dki2	9/1/2024 3:00	58	75	58	7	33	18	Medium
dki2	9/1/2024 4:00	58	75	58	7	33	17	Medium
dki3	9/18/2024 6:00	48	45	38	7	34	36	Good
dki4	9/11/2024 8:00	74	82	55	22	26	17	Medium
dki5	9/5/2024 14:00	36	102	35	20	40	14	Unhealthy

The final step in preparation is data formatting. During this phase, the data is organized and ready to be processed with data mining methods. LabelEncoder changes the location value from "object" to "integer". Table 9 shows the data and attributes used to classify air quality in DKI Jakarta.

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

Table 9. Attributes and Data

Attributes	Type	Data Exploration
lokasi	Integer	The geographic location contains: 1. dki1 changes to 0 2. dki2 changes to 1 3. dki3 changes to 2 4. dki4 changes to 3 5. dki5 changes to 4 Total amount of 7.185 data
pm10	Object	Total amount of 7.185 data
pm2.5	Object	Total amount of 7.185 data
so2	Object	Total amount of 7.185 data
co	Object	Total amount of 7.185 data
o3	Object	Total amount of 7.185 data
no2	Object	Total amount of 7.185 data
ispu_status	Object	ispu_status air quality includes: 1. Good = 650 2. Medium = 3.768 3. Unhealthy = 2.767 Total amount of 7.185 data

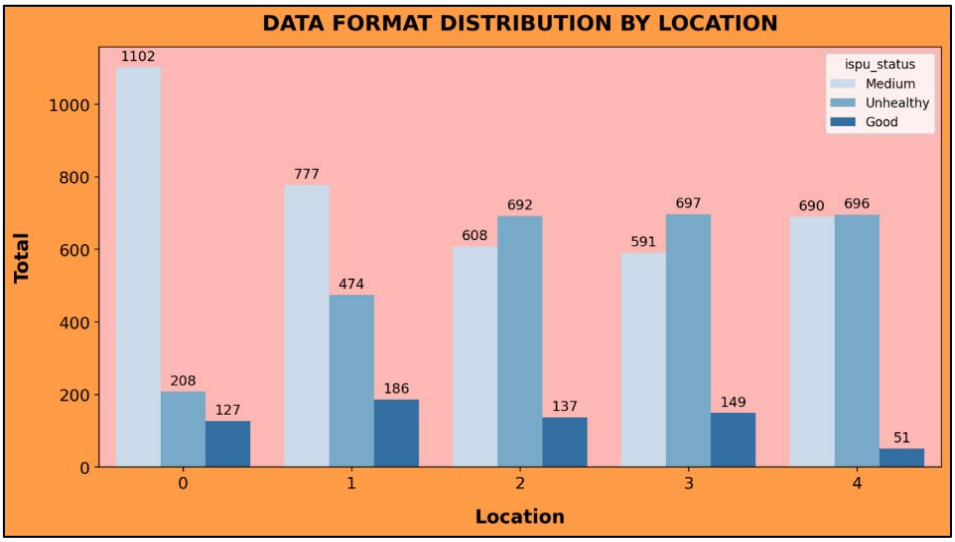

Figure 3. Data Format Distribution

The illustration in Figure 3 shows the distribution of data formats based on location that have gone through the encode process into numbers 0 to 4, which represent each region with information including:

a. 0 (dki1), Medium (1,102 data), Unhealthy (208 data), and Good (127 data).
b. 1 (dki2), Medium (777 data), Unhealthy (474 data), and Good (186 data).
c. 2 (dki3), Unhealthy (692 data), Medium (608 data), and Good (137 data).
d. 3 (dki4), Unhealthy (697 data), Medium (591 data), and Good (149 data).
e. 4 (dki5), Unhealthy (696 data), Medium (690 data), and Good (51 data).

3.4 Split Data

With a total net data of 7.185, the data was divided into three scenarios for training and testing. The first scenario used 80% of the data for training and 20% of the data for testing, resulting in 5.748 training data and 1.437 testing data. The second scenario segmented the data into 70% for training and 30% for testing, resulting in

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

5.030 training data and 2.155 testing data. In the third scenario, 90% of the data is used for training and 10% of the data for testing, resulting in 6.466 training data and 719 testing data.

3.5 Testing C4.5 Algorithm

The C4.5 algorithm uses the training data to create a decision tree with the target attribute ispu_status, which is categorized into Good, Medium, and Unhealthy. The decision tree is built by selecting the attribute with the highest information gain at each node, and this procedure is performed iteratively until all data is categorized or the stopping criterion is reached. To prevent overfitting, the parameters *max_depth=6* were used to limit the depth of the tree, and *min_samples_split=5* to set the minimum number of samples at the split nodes, thus ensuring that the model remains generalized to newly acquired data. Figure 4, 5, and 6 show the details of data splitting and attribute distribution.

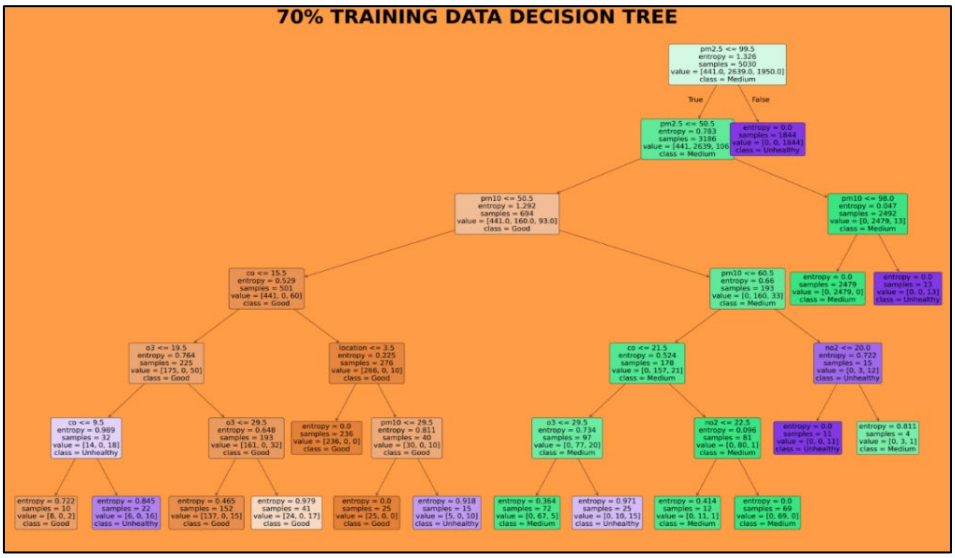

Figure 4. 70% Training Data Decision Tree

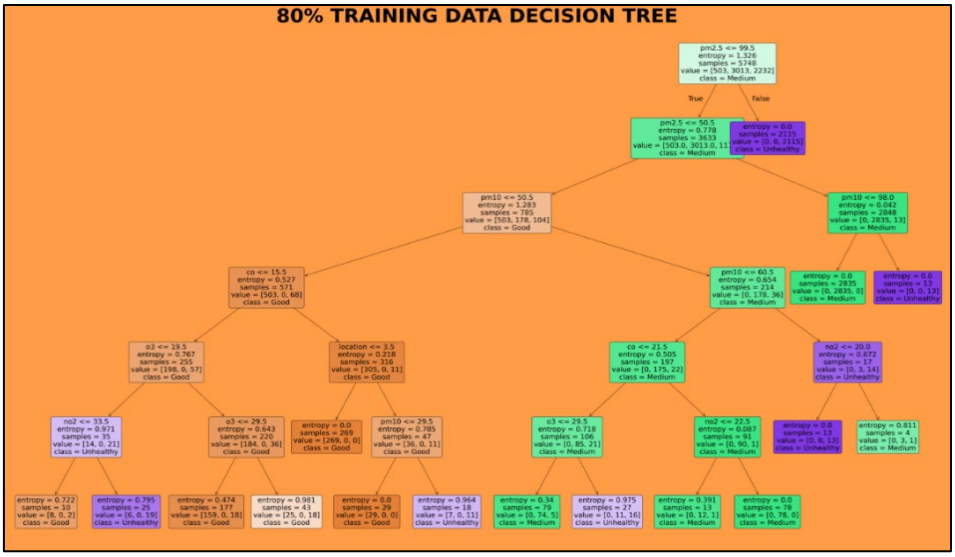

Figure 5. 80% Training Data Decision Tree

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

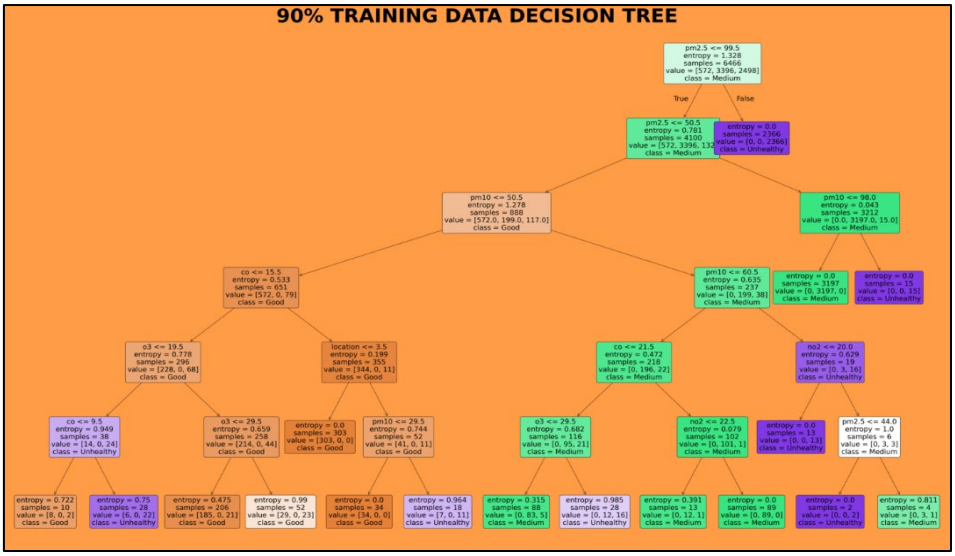

Figure 6. 90% Training Data Decision Tree

Furthermore, the test results carried out on the air quality classification method in DKI Jakarta using the C4 algorithm. 5 algorithm are shown in tables 4.10, 4.11 and 4.12. These tables present a confusion matrix that describes the comparison between the model classification and the actual value of the test data, so that it can be used to assess the classification performance generated by the C4.5 algorithm.

Table 10. Confusion Matrix C4.5 30% Testing Data

Actual	Good Classification	Medium Classification	Unhealthy Classification
Good Actual	209	0	0
Medium Actual	64	1.065	0
Unhealthy Actual	32	4	781

Table 11. Confusion Matrix C4.5 20% Testing Data

Actual	Good Classification	Medium Classification	Unhealthy Classification
Good Actual	147	0	0
Medium Actual	46	709	0
Unhealthy Actual	21	4	510

Table 12. Confusion Matrix C4.5 10% Testing Data

Actual	Good Classification	Medium Classification	Unhealthy Classification
Good Actual	78	0	0
Medium Actual	25	347	0
Unhealthy Actual	8	2	259

Based on Table 4.10, 4.11, and 4.12, the evaluation of the C4.5 algorithm can be done by calculating accuracy, precision, recall, and f1-score. The results of the calculations performed to determine these values are presented below:

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

a. 30% Testing Data

$$\text{Accuracy} = \frac{2.055}{2.155} \times 100\% = 95\%$$

$$\text{Precision} = \frac{1.065}{1.065 + 4} \times 100\% = 100\%$$

$$\text{Recall} = \frac{1.065}{1.065 + 64} \times 100\% = 94\%$$

$$\text{F1} - \text{score} = 2 \times \frac{1 \times 0{,}94}{1 + 0{,}94} \times 100\% = 97\%$$

b. 20% Testing Data

$$\text{Accuracy} = \frac{1.366}{1.437} \times 100\% = 95\%$$

$$\text{Precision} = \frac{709}{709 + (4 + 0)} \times 100\% = 99\%$$

$$\text{Recall} = \frac{709}{709 + (46 + 0)} \times 100\% = 94\%$$

$$\text{F1} - \text{score} = 2 \times \frac{0{,}99 \times 0{,}94}{0{,}99 + 0{,}94} \times 100\% = 97\%$$

c. 10% Testing Data

$$\text{Accuracy} = \frac{684}{719} \times 100\% = 95\%$$

$$\text{Precision} = \frac{347}{347 + 2} \times 100\% = 99\%$$

$$\text{Recall} = \frac{347}{347 + 25} \times 100\% = 93\%$$

$$\text{F1} - \text{score} = 2 \times \frac{0{,}99 \times 0{,}93}{0{,}99 + 0{,}93} \times 100\% = 96\%$$

3.6 Testing Naïve Bayes Algorithm

In Naive Bayes air quality classification, training data is used to build a model with the target attribute ispu_status classified into Good, Medium, and Unhealthy. This algorithm applies Bayes' Theorem with the assumption that the features in the dataset have a Gaussian (normal) distribution, thus calculating posterior probabilities to determine the likelihood of the data falling into one of the categories. The Gaussian Naïve Bayes classification graphs in Figure 7, 8, and 9 show the classification results based on the probabilities of the three categories. The Medium classification dominates with probabilities often close to 1, while Good has a smaller distribution but is still significant at certain points. In addition, Unhealthy, although it has some high probability peaks, is generally less dominant than Medium, indicating a more concentrated pattern of data distribution.

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

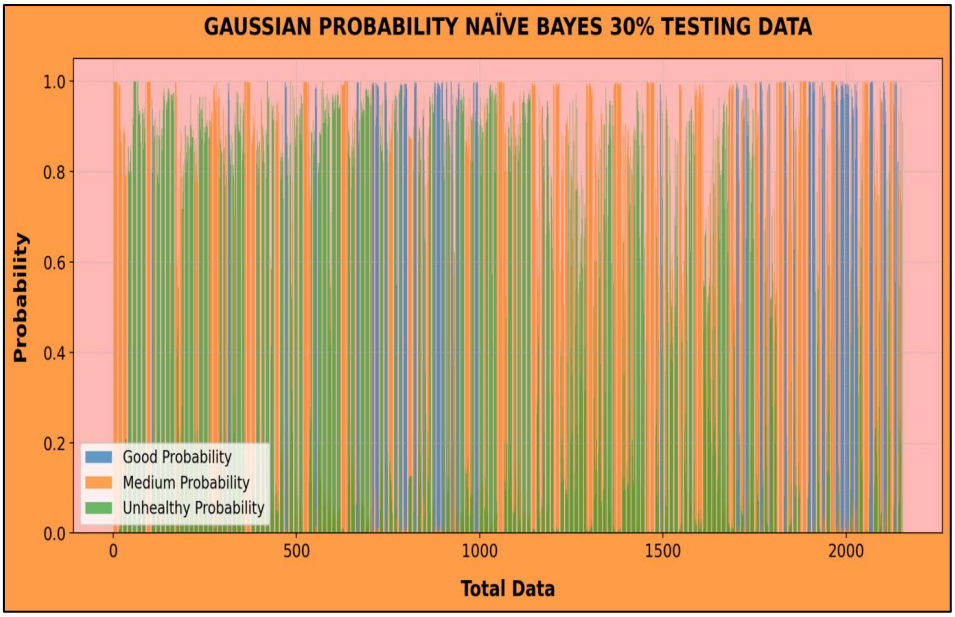

Figure 7. Probability 30% Testing Data

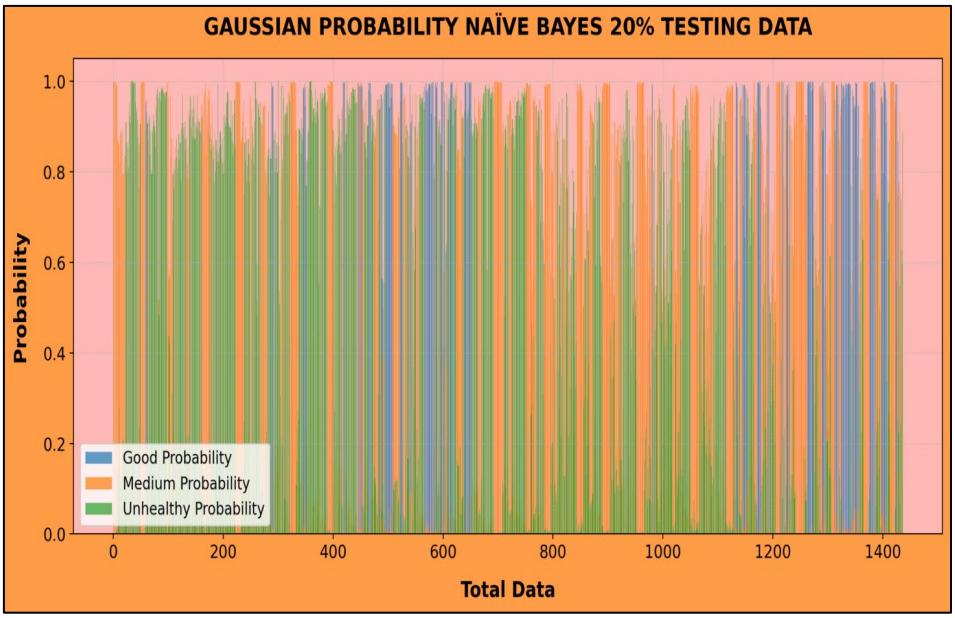

Figure 8. Probability 20% Testing Data

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

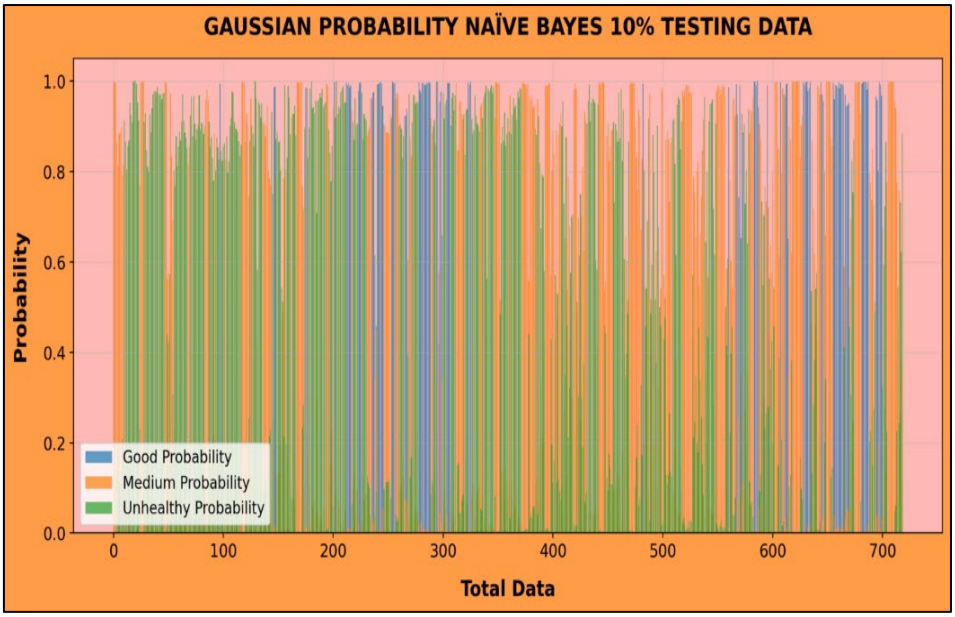

Figure 9. Probability 10% Testing Data

Table 13, 14, and 15 show the results of tests conducted on the classification of air quality in DKI Jakarta assessed using the Naïve Bayes algorithm.

Table 13. Confusion Matrix Naïve Bayes 30% Testing Data

Actual	Good Classification	Medium Classification	Unhealthy Classification
Good Actual	205	4	0
Medium Actual	62	841	226
Unhealthy Actual	29	74	714

Table 14. Confusion Matrix Naïve Bayes 20% Testing Data

Actual	Good Classification	Medium Classification	Unhealthy Classification
Good Actual	145	2	0
Medium Actual	44	549	162
Unhealthy Actual	20	41	474

Table 15. Confusion Matrix Naïve Bayes 10% Testing Data

Actual	Good Classification	Medium Classification	Unhealthy Classification
Good Actual	77	1	0
Medium Actual	24	266	82
Unhealthy Actual	8	19	242

Based on Table 13, 14, and 15 in evaluating the performance of the Naïve Bayes algorithm, the accuracy, precision, recall, and f1-score values can be calculated. The results of the calculations performed to determine the values are shown below:

SAGA: Journal of Technology and Information Systems
Vol 2, Issue 4, November 2024, Page 311-327
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i4.395

a. 30% Testing Data

$$\text{Accuracy} = \frac{1.760}{2.155} \times 100\% = 82\%$$

$$\text{Precision} = \frac{841}{841 + 78} \times 100\% = 92\%$$

$$\text{Recall} = \frac{841}{841 + 288} \times 100\% = 74\%$$

$$\text{F1} - \text{score} = 2 \times \frac{0{,}92 \times 0{,}74}{0{,}92 + 0{,}74} \times 100\% = 82\%$$

b. 20% Testing Data

$$\text{Accuracy} = \frac{1.168}{1.437} \times 100\% = 81\%$$

$$\text{Precision} = \frac{549}{549 + 43} \times 100\% = 93\%$$

$$\text{Recall} = \frac{549}{549 + 206} \times 100\% = 73\%$$

$$\text{F1} - \text{score} = 2 \times \frac{0{,}93 \times 0{,}73}{0{,}93 + 0{,}73} \times 100\% = 82\%$$

c. 10% Testing Data

$$Accuracy = \frac{585}{719} \times 100\% = 81\%$$

$$Precision = \frac{266}{266 + 20} \times 100\% = 93\%$$

$$Recall = \frac{266}{266 + 106} \times 100\% = 72\%$$

$$F1 - score = 2 \times \frac{0{,}93 \times 0{,}72}{0{,}93 + 0{,}72} \times 100\% = 81\%$$

3.7 Results Algorithm Comparison

The objective of this study is to compare the performance of the C4.5 and Naïve Bayes algorithms in classifying air quality levels in DKI Jakarta. Table 16 visualizes the objective by showing the comparison of classification results between the C4.5 and Naïve Bayes algorithms:

Table 16. Algorithm Comparison

Algorithm	Ratio	*Accuracy*	*Precision*	*Recall*	*F1-score*
C4.5	70% : 30%	95%	100%	94%	97%
	80% : 20%	95%	99%	94%	97%
	90% : 10%	95%	99%	93%	96%
Naïve Bayes	70% : 30%	82%	92%	74%	82%
	80% : 20%	81%	93%	73%	82%
	90% : 10%	81%	93%	72%	81%

Table 16 compares the performance of air quality classification in DKI Jakarta in terms of accuracy, precision, recall, and f1-score, using a confusion matrix to assess the effectiveness of each algorithm. The findings from the evaluation show that the C4. 5 algorithm has the best performance in all situations. The C4.5 algorithm, with ratios of 70% : 30%, 80% : 20% and 90% : 10%, consistently achieved 95% accuracy, 99%-100% precision, 93%-94% recall and 96%-97% f1-score, indicating highly accurate classification rates. In contrast, Naive Bayes showed lower performance, with accuracy ranging from 81% to 82%, precision 92-93%, recall 72%-74% and f1 score 81%-82%. These findings corroborate the superiority of C4.5, which managed to consistently maintain 95% accuracy, 99%-100% accuracy indicating positive classification accuracy, 93%-94% recall indicating good detection coverage, and 96%-97% f1 score, indicating a balance between accuracy and recall in correctly classifying the data.

4 Conclusion

Based on this study, air quality in DKI Jakarta can be effectively classified using the Air Pollutant Standard Index (ISPU) from 15 representative areas including Central Jakarta, North Jakarta, South Jakarta, East Jakarta and West Jakarta. The ISPU data used includes relevant attributes such as dust particles (PM10, PM2.5), pollutant gases (SO2, CO, O3, NO2), location, time, and air quality status (ispu_status), which are key parameters in air quality classification. This research reveals that the C4.5 algorithm performs better than Naïve Bayes in air quality classification. In the context of data ratios of 70% : 30%, 80% : 20%, and 90% : 10%, the C4.5 algorithm recorded an average accuracy of 95%, precision of 99%, recall of 94%, and f1-score of 97%, indicating a high degree of accuracy and consistency in classification. On the other hand, Naïve Bayes achieved lower results, with an average accuracy of 81%, precision of 93%, recall of 73% and f1-score of 82%, indicating that this algorithm is less suitable for handling datasets with correlated attributes.

This research supports air quality monitoring in DKI Jakarta, contributes to increasing environmental awareness among the public and supports policy makers' decisions regarding air pollution regulations. The results show that the C4.5 algorithm outperforms Naïve Bayes in classifying air quality based on the Air Pollutant Standard Index (ISPU). For further development, this research can be expanded with a larger dataset, both in terms of area coverage and time period, to increase the model's coverage. In addition, the application of other classification algorithms can improve the analysis and performance comparison of different methods. Incorporating additional attributes, such as meteorological factors or emission sources, may also improve the accuracy and reliability of air quality classification in the future.

References

[1] A. A. H. Kirono, I. Asror, and Y. F. A. Wibowo, “Klasifikasi Tingkat Kualitas Udara Dki Jakarta Menggunakan Algoritma Naïve Bayes,” *e-Proceeding Eng.*, vol. 9, no. 3, p. 1962, 2022.

[2] A. D. Wiranata, S. Soleman, I. Irwansyah, I. K. Sudaryana, and Rizal, “Klasifikasi Data Mining Untuk Menentukan Kualitas Udara Di Provinsi Dki Jakarta Menggunakan Algoritma K-Nearest Neighbors (K-Nn),” *Infotech J. Technol. Inf.*, vol. 9, no. 1, pp. 95–100, 2023, doi: 10.37365/jti.v9i1.164.

[3] A. Budianita, N. Iman, F. M. Hana, and C. B. Hakim, “Komparasi Algoritma K-Nearest Neighbor dan Naive Bayes pada Klasifikasi Tingkat Kualitas Udara Kota Tangerang Selatan,” vol. 6, no. 1, pp. 320–327, 2024.

[4] A. A. Nababan, M. Jannah, M. Aulina, and D. Andrian, “Prediksi Kualitas Udara Menggunakan Xgboost Dengan Synthetic Minority Oversampling Technique (Smote) Berdasarkan Indeks Standar Pencemaran Udara (Ispu),” *JTIK (Jurnal Tek. Inform. Kaputama)*, vol. 7, no. 1, pp. 214–219, 2023, doi: 10.59697/jtik.v7i1.66.

[5] R. Setiawan and A. Triayudi, “Klasifikasi Status Gizi Balita Menggunakan Naïve Bayes dan K-Nearest Neighbor Berbasis Web,” *J. Media Inform. Budidarma*, vol. 6, no. 2, p. 777, 2022, doi: 10.30865/mib.v6i2.3566.

[6] B. Valentino Jayadi, T. Handhayani, and M. Dolok Lauro, “Perbandingan Knn Dan Svm Untuk Klasifikasi Kualitas Udara Di Jakarta,” *J. Ilmu Komput. dan Sist. Inf.*, vol. 11, no. 2, 2023, doi: 10.24912/jiksi.v11i2.26006.

[7] I. Romli and A. T. Zy, “Penentuan Jadwal Overtime Dengan Klasifikasi Data Karyawan Menggunakan Algoritma C4.5,” *J. Sains Komput. Inform. (J-SAKTI*, vol. 4, no. 2, pp. 694–702, 2020.

[8] Y. Apridiansyah, N. D. M. Veronika, and E. D. Putra, “Prediksi Kelulusan Mahasiswa Fakultas Teknik Informatika,” *JSAI J. Sci. Appl. Informatics*, vol. 4, no. 2, pp. 236–247, 2021.

[9] P. S. Zakaria, R. Julianto, and R. S. Bernada, “Implementasi Naive Bayes Menggunakan Python dalam Klasifikasi Data,” *BIIKMA Bul. Ilm. Ilmu Komput. dan Multimed.*, vol. 1, no. 2, pp. 126–131, 2023.

[10] A. Nugroho, I. Asror, and Y. F. A. Wibowo, "Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan Open Government Data Menggunakan Algoritma Random Forest," *eProceedings Eng.*, vol. 10, No. 2, no. 2, pp. 1824–1834, 2023, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20030%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20030/19395

[11] T. H. Hasibuan and D. Mahdiana, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma C4.5 Pada Uin Syarif Hidayatullah Jakarta," *Skanika*, vol. 6, no. 1, pp. 61–74, 2023, doi: 10.36080/skanika.v6i1.2976.

[12] L. Y. Lumban Gaol, M. Safii, and D. Suhendro, "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Dengan Menggunakan Algoritma C4.5," *Brahmana J. Penerapan Kecerdasan Buatan*, vol. 2, no. 2, pp. 97–106, 2021, doi: 10.30645/brahmana.v2i2.71.

[13] M. Bagriacik and F. E. B. Otero, "Multiple fairness criteria in decision tree learning," *Appl. Soft Comput.*, vol. 167, no. PA, p. 112313, 2024, doi: 10.1016/j.asoc.2024.112313.

[14] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Comput. Oper. Res.*, vol. 135, p. 105456, 2021, doi: 10.1016/j.cor.2021.105456.

[15] 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.

[16] M. M. Arif, H. Setiawan, and A. S. Fitrani, "Penggunaan Datamining Untuk Memprediksi Masa Studi Mahasiswa di Universitas Muhammadiyah Sidoarjo Dengan Algoritma Naive Bayes," *... dan Manajemen)*, vol. 4, no. 3, pp. 622–629, 2023, [Online]. Available: https://pkm.tunasbangsa.ac.id/index.php/kesatria/article/view/210%0Ahttps://pkm.tunasbangsa.ac.id/index.php/kesatria/article/download/210/209

[17] M. Fahmy Amin, "Confusion Matrix in Three-class Classification Problems: A Step-by-Step Tutorial," *J. Eng. Res.*, vol. 7, no. 1, pp. 0–0, 2023, doi: 10.21608/erjeng.2023.296718.

[18] Hozairi, Anwari, and S. Alim, "Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes," *Netw. Eng. Res. Oper.*, vol. 6, no. 2, p. 133, 2021, doi: 10.21107/nero.v6i2.237.

[19] D. Nurnaningsih, D. Alamsyah, A. Herdiansah, and A. A. J. Sinlae, "Identifikasi Citra Tanaman Obat Jenis Rimpang dengan Euclidean Distance Berdasarkan Ciri Bentuk dan Tekstur," *Build. Informatics, Technol. Sci.*, vol. 3, no. 3, pp. 171–178, 2021, doi: 10.47065/bits.v3i3.1019.

[20] R. R. Adhitya, Wina Witanti, and Rezki Yuniarti, "Perbandingan Metode Cart Dan Naïve Bayes Untuk Klasifikasi Customer Churn," *INFOTECH J.*, vol. 9, no. 2, pp. 307–318, 2023, doi: 10.31949/infotech.v9i2.5641.