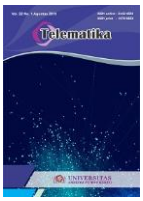

Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemenristek/BRIN, No. 85/M/KPT/2020

Performance Evaluation of Naive Bayes Algorithm for Classification of Fertilizer Types

Rastri Prathivi[1], April Firman Daru[2], Sara Sharifzadeh[3]

[1,2]Department of Informatics Engineering, Universitas Semarang
[3]Department of Computer Science, Swansea University
E-mail: vivi@usm.ac.id[1], firman@usm.ac.id[2], sara.sharifzadeh@swansea.ac.uk[3]

ARTICLE INFO

History of the article:
Received October 19, 2021
Revised December 21, 2021
Accepted February 15, 2022

Keywords:
Classification,
Naïve Bayes,
Fertilizer

Correspondece:
Telepon: +62 85950251065
E-mail: vivi@usm.ac.id

ABSTRACT

Determining the right fertilizer is very important to get optimal plant growth results. Each plant requires different nutrient requirements. Different soil types cause the soil's nutrient content and PH value to differ from one type to another. Regional conditions in a place will also cause the need for plant absorption of nutrient content to be more varied. By using the classification of the problems that have been mentioned, it can be solved by studying patterns from existing fertilizer use data into knowledge that can be used to determine decisions. In this study, modeling with the Naïve Bayes algorithm has been applied to the existing fertilizer use data where the probability value of each class has been calculated to get the highest probability value of a class. The measurement of the accuracy value of the modeling used is measured using the Split Validation method, where the training data will be divided into training data and testing data so that the accuracy value of the model is obtained. From the applied modeling, an accuracy value of 60% is obtained, which shows the level of accuracy of the model obtained from the classification results in the form of the name of the fertilizer, which is expected to help in determining the name of the fertilizer that needs to be used.

INTRODUCTION

Soil fertility is a very important part of the development of agriculture. Soil can be said to be fertile if the soil contains sufficient nutrients to support plant growth. The content of sufficient nutrients in the soil will help plant growth so that plants are able to produce products of good quality and quantity. The following are indicators of soil fertility from the characteristics of the soil's physical, chemical, and biological properties. (Walida et al., 2020). From these three parameters, efforts are needed to improve the physical properties of the soil and the chemical properties of the soil by applying fertilizer to the soil. Chemical properties for soil fertility consist of the degree of soil acidity (pH), nutrient content, and organic matter content (BO). The acidity (pH) level greatly affects the nutrient content and activity of microorganisms in the soil. Soil that is said to be fertile is soil with a pH of around 6 – 7.5 or at a neutral pH, because at that pH most nutrients are easily soluble in water, and microorganisms can thrive. In addition to the degree of acidity, the content of organic matter in the soil has a role in increasing the availability of nutrients and improving soil fertility. In other words, the absorption of nutrients is maximized because organic matter can increase the negative charge to increase the cation exchange capacity of the nutrients to be optimal (Bhatt, 2019).

Fertilizer is an essential part of the agricultural system that is useful for improving soil fertility. A good fertilizer for soil fertility must be able to apply chemical growth to the soil and not cause problems in the soil (Pahalvi, 2021). From the many types of fertilizers on the market, it is necessary to understand the types of fertilizers and their contents to determine the appropriate fertilizer to use. One solution to overcome the problem of choosing the right type of fertilizer is applying classification to select the right type of fertilizer.

Several studies related to classification include research conducted, wickramasingh proposed a system for detecting soil fertility using the principle of colorimetry and Naive Bayes classification. The result is the accuracy of the system in determining the status of the main nutrient content in the soil is verified (Wickramasingh, 2021). Jha compared the classification of soil fertility with chemical fertilizers using the Naive Bayes method, Multi-Layer Perceptron (MLP), and Logistic Regression. The result is that MLP accuracy is 60% superior to Naive Bayes, which is 54%. (Jha, 2020).

Saouza, in his book, explains that classification in Naive Bayes assumes all independent variables so that only a little training data can get accurate classification results (Souza, 2021). Because Naive Bayes has a good performance inaccuracy, this research uses the Naïve Bayes algorithm to determine the classification of fertilizers. In the problem of determining the type of fertilizer that has been described, an analysis of the sample data that has been obtained will be carried out to obtain the pattern of information that is needed.

Classification using Naïve Bayes will calculate the probability of each available attribute, whether it is a numerical or nominal attribute. Priya and friends use of Naïve Bayes algorithm to make the model very efficient in computation. The system is scalable as it can be used to test different crops. This model can be further enhanced to find the yield of every crop and for pesticide recommendations. Also, it can be modified to suggest the fertilizers and irrigation needs of crops (Priya, Ramesh, & Khosla, 2018). However, Naïve Bayes will not work well if there are many attributes whose conditional probability is zero. Therefore conditional probability with a zero value should be solved by applying Laplacian correction.

There are about 13 to 20 chemical fertilizers on the market with various brands. The following are the types of fertilizers that are widely used that contain chemicals, namely urea fertilizer, ZA (Zwavelzure Ammonium), SP-36 (super phosphate), KCl (Potassium Chloride), NPK Phonska (Nitrogen Phosphate Potassium), Dolomite (Lime Carbonate), ZK (Zwavelzure Kali). The number of fertilizers available in the market and the need to choose the right fertilizer to produce optimal crop yields requires the application of a classification method to facilitate the selection of fertilizers to obtain optimal crop yields.

By applying the Naïve Bayes algorithm to the fertilizer use data obtained, a model can be generated that can be used to classify new data so that classification results will be obtained based on patterns from old data that have been modeled with the Naïve Bayes algorithm. In this case, modeling will be carried out using the Naïve Bayes algorithm to obtain an accuracy value that can be used as a benchmark for the classification results obtained through the new data tested.

RESEARCH METHODS

The research method consists of several stages: data analysis, fertilizer grouping, preprocessing data, classification process, and calculation of accuracy.

1. Data analysis

The data used in this research is a public dataset published on the Kaggle data sharing site under the name "Fertilizer Prediction," which can be accessed using the link: kaggle.com/gdabhishek/fertilizer-prediction. The Data contains data from Temperature, Humidity, Moisture, Soil Type, Crop Type, Nitrogen, Potassium, Phosphorus, and Fertilizer names, which have 99 (ninety-nine) data records, which are presented in Table 1.

Table 1. Dataset Fertilizer Prediction (source: kaggle.com/gdabhishek/fertilizer-prediction)

No	*Temperature*	*Humidity*	*Moisture*	*SoilType*	*CropType*	*Nitrogen*	*Potassium*	*Phosphorus*	*Fertilizer Name*
1	26	52	38	Sandy	Maize	37	0	0	Urea
2	29	52	45	Loamy	Sugarcane	12	0	36	DAP
3	34	65	62	Black	Cotton	7	9	30	14-35-14
4	32	62	34	Red	Tobacco	22	0	20	28-28
5	28	54	46	Clayey	Paddy	35	0	0	Urea
6	26	52	35	Sandy	Barley	12	10	13	17-17-17
…	…	…	…	…	…	…	…	…	…
96	30	60	27	Red	Tobacco	4	17	17	10-26-26
97	38	72	51	Loamy	Wheat	39	0	0	Urea
98	36	60	43	Sandy	Millets	15	0	41	DAP
99	29	58	57	Black	Sugarcane	12	0	10	20-20

2. Fertilizer Grouping

Fertilizer (Roy, 2017) is a substance that gives plants the nutrients they need for plant growth. In addition to needing water and nutrients, carbon dioxide from the air, and energy from sunlight, plants also need other nutrients. Other nutrients are divided into primary nutrients (nitrogen, phosphorus, and potassium/potassium) and secondary nutrients (calcium, magnesium, and sulfur). In addition, plants also need other nutrients in much smaller amounts, and they are referred to as micronutrients (boron, chlorine, copper, iron, manganese, molybdenum, and zinc). To prevent land degradation, nutrients that plants have taken up must be replaced through fertilizers in the soil.

From the dataset used in this study, there are 7 (seven) groups or classes that are names of fertilizers, as shown in Table 2.

Table 2. Type of Fertilizer in The Dataset Used

Fertilizer Name	*Fertilizer Type*	*Suitable land*
10-26-26	NPK compound fertilizer (N:10%, P:26%, K:26%)	All types of soil
14-35-35	NPK compound fertilizer (N:14%, P:35%, K:35%)	All types of soil
17-17-17	NPK compound fertilizer (N:17%, P:17%, K:17%)	All types of soil
20-20	NPK compound fertilizer (N:20%, P:20%, K:0%)	All types of soil
28-28	NPK compound fertilizer (N:28%, P:28%, K:0%)	All types of soil
DAP	DAP compound fertilizer (N:18%, P:46%)	All types of soil
Urea	Single fertilizer (N:46%)	All types of soil

3. Preprocessing Data

Raw Data is very vulnerable to missing values, noise, and inconsistencies, as well as data quality that affects the results of data mining. To improve data quality, raw Data needs to be processed to improve quality and eliminate errors that will occur. Stages that can be applied to perform preprocessing include Data Cleaning, Data Integration, Data Transformation, and Data Reduction (Bhatia, 2019).

4. Classification Process With Naïve Bayes

Naïve Bayes is an algorithm that can be used to predict the probability of membership of a class, such as the probability that a particular tuple belongs to a particular class (Han, Kamber, & Pei, 2012).

Classification using Naïve Bayes is different from a classification using a Decision Tree. It is based on the hypothesis that the given Data belongs to a certain class. In this theorem, probabilities are calculated for the hypothesis to be true.

In classification problems on categorical/nominal data attributes, we need to determine the posterior probability P(H|X) using the P(H), P(X), and P(X|H) probabilities (Suntoro, 2019).

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Where :

$P(H|X)$ = The probability of hypothesis H is based on condition X (posterior probability).

$P(X|H)$ = The probability of X is based on the conditions in hypothesis H. (conditional probability).

$P(H)$ = The probability of hypothesis H (prior probability).

$P(X)$ = The probability of the data sought.

H = Hypothesis data which is a specific class.

X = Data with unknown class.

In the Naïve Bayes algorithm, the above equation can be explained by the equation below.

$$P(C_i|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|C_i).P(C_i)}{P(\boldsymbol{X})} \quad (2)$$

Because P(X) value is fixed for a certain point, Bayes' rules can be rewritten with the following equation (Han, Kamber & Pei, 2012).

$$P(C_i|\boldsymbol{X}) = P(\boldsymbol{X}|C_i).P(C_i) \quad (3)$$

To reduce the calculation in identifying $P(\boldsymbol{X}|C_i)$ a Naive assumption of class-conditional independence is made, which means that the value of an attribute stands alone.

$$P(\boldsymbol{X}|C_i) = \prod P(x_k|C_i)$$
$$= P(x_1|C_i).P(x_2|C_i) \ldots . P(x_n|C_i) \quad (4)$$

Where :

$P(C_i|\boldsymbol{X})$ = The probability of a specific class based on condition X (posterior probability).

$P(\boldsymbol{X}|C_i)$ = The probability of X is based on conditions in a specific class (conditional probability).

$P(C_i)$ = The probability of a specific class (prior probability).

$P(\boldsymbol{X})$ = The probability of the data looking for the class.

$\boldsymbol{X}$ = Data with unknown class.

C_i = Data that is a specific class.

$x_{1\ldots\ldots n}$ = Data of an attribute with an unknown class.

Attributes with numeric data types will usually apply a Gaussian distribution with mean, standard deviation, and variance. The Gaussian Naïve Bayes equation can be seen in equation 5.

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e - \frac{(x-\mu)^2}{2\sigma^2} \quad (5)$$

Where :

$g(x, \mu, \sigma)$ = Gaussian value is based on the value of x (Data), mean, and standard deviation σ.

π = The value of phi (3.14).

σ = The attribute standard deviation value based on a specific class.

μ = The attribute mean value based on a specific class.

x = Data value with unknown class.

To find the mean μ and standard deviation σ, the following equation can be used.

$$\mu = \frac{\sum_i^n x_i}{n} \quad (6)$$

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \mu)^2}{n-1}} \quad (7)$$

Where :

μ = The mean value of an attribute with a specific class.

σ = The standard deviation value with a specific class.

x_i = Data with a specific class.

n = The Count of data in an attribute.

In the calculation of naive Bayes data of nominal type, a conditional probability calculation will be applied for each Data with a specific class, while for data with a numeric data type, a Gaussian calculation will be applied, so that from equation (5) it can be concluded that the conditional probability value can be replaced with a value from gaussian for numeric data types.

$$P(x_1|C_i) = g(x, \mu, \sigma) \quad (8)$$

Where: $P(x_1|C_i)$ = Conditional probability value

$g(x, \mu, \sigma)$ = Gaussian value

In the application of the Naïve Bayes algorithm, a probability value with a value of 0 (zero) will be obtained when the probability of one of the attributes is 0 (zero). A probability with a value of 0 (zero) will thwart the effect of the other attribute probabilities. To overcome this, the application of Laplacian correction can be used. The equation of the Laplacian correction can be seen below.

$$\rho_i = \frac{m_i + 1}{n + k} \quad (9)$$

Where :

ρ_i = The probability of an attribute.

m_i = The number of samples in the class of an attribute.

n = Count of samples.

k = The Count of classes of an attribute.

5. Accuracy of Classification

Classification (Bhatia, 2019) is a classic method used to predict the outcome of an unknown sample. Classification is used to categorize objects into a number of already available classes. In other words, Classification (Zaki & Meira, 2014) refers to the task of predicting class labels for objects that do not yet have labels. In the classification of the model that has been made, validation must be carried out to measure the accuracy value of the model that has been made. The calculation of the accuracy value can be done using the confusion matrix method as the equation below.

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

Where: TP = The number of positive data correctly classified by the system.

TN = The number of negative data correctly classified by the system.

FP = The number of positive data but classified incorrectly by the system.

FN = The number of negative data classified incorrectly by the system.

Equation 1 shows calculated the accuracy of the Naive Bayes algorithm when the classification process has been obtained. Whereas in the classification with multiple classes, the results on the test set are often displayed in the form of a two-dimensional confusion matrix with rows and columns for each class (Witten, Frank, Hall, & Pal, 2017).

In this study, the author uses the CRISP-DM method as a development method used to make development to be more structured. In the CRISP-DM method, there are 6 (six) stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, whose flow can be seen in Figure 1.

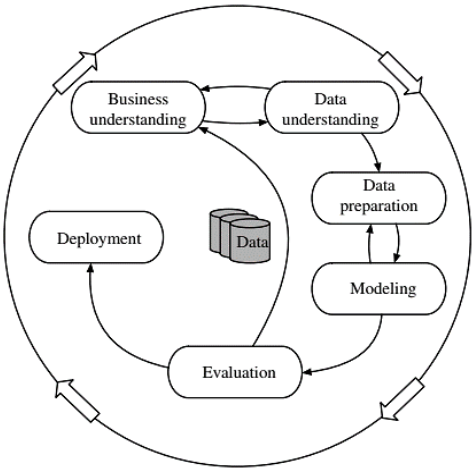

Figure 1. Stages of CRISP-DM (Witten, Frank, Hall, & Pal, 2017)

The Business Understanding stage requires knowledge of business objects, how to construct or retrieve data, and how to match modeling objectives to business objectives so that the best model can be built. At the Data Understanding stage, it is useful to examine the data, so that it can identify problems in the data. In the Modeling Stage, predictive or descriptive models are made. At the Evaluation Stage, interpretation is carried out on the results of the data mining generated in the modeling process in the previous stage. The deployment or planning phase is the last phase of CRISP-DM, namely the use of the model.

RESULTS AND DISCUSSION

1. Application of the CRISP-DM method in case studies

a. Business understanding

From the data contained in the dataset in Table 1, it is found the need for processing the data to classify the appropriate type of fertilizer based on existing data records by applying classification with the Naïve Bayes algorithm so that it can be used as a tool to help determine decisions in choosing the type of fertilizer that should be used.

b. Data understanding

From the data in Table 1, which has 99 (ninety-nine) data records, information is obtained that there are 8 (eight) attributes, namely Temperature, Humidity, Moisture, Soil Type, Crop Type, Nitrogen, Potassium, and Phosphorus, which will be used to determine the classification results of a class named Fertilizer Name.

c. Data preparation

In Table 1, there are a lot of missing values or missing data or data with a value of 0 (zero), so it is necessary to process the data before applying the algorithm at the modeling stage. We decided to eliminate the attributes of Potassium and Phosphorus due to a large number of missing values in a specific class which will result in the presence of the mean, standard deviation, and variation being 0 (zero), resulting in the probability value of a class being 0, as well as to reduce dimensions of the dataset. The Data that has been applied to data preparation can be seen in Table 4.

Table 4. The Data That Has Been Applied to Data Preparation

No	*Temperature*	*Humidity*	*Moisture*	*SoilType*	*CropType*	*Nitrogen*	*Fertilizer Name*
1	26	52	38	Sandy	Maize	37	Urea
2	29	52	45	Loamy	Sugarcane	12	DAP
3	34	65	62	Black	Cotton	7	14-35-14
4	32	62	34	Red	Tobacco	22	28-28
5	28	54	46	Clayey	Paddy	35	Urea
…	…	…	…	…	…	…	…
96	30	60	27	Red	Tobacco	4	10-26-26
97	38	72	51	Loamy	Wheat	39	Urea
98	36	60	43	Sandy	Millets	15	DAP
99	29	58	57	Black	Sugarcane	12	20-20

d. Modeling

After the data has been prepared through the data preparation stage, data mining algorithms will be applied at the modeling stage. The algorithm that the author uses in this study is the Naïve Bayes algorithm which can be used for data with nominal and numeric attributes. The flow of the stages of the Naïve Bayes algorithm can be seen in Figure 2.

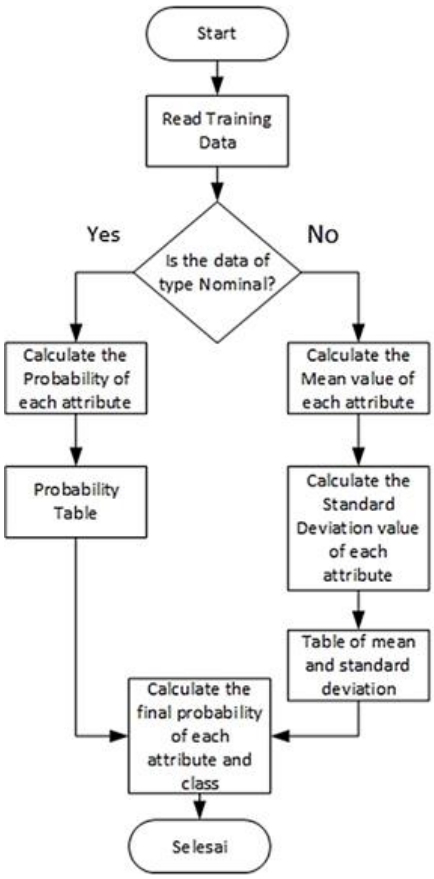

Figure 2. The Flow of The Stages of The Naïve Bayes Algorithm

From Table 4, it can be seen that the attributes of Temperature, Humidity, Moisture, and Nitrogen are of numeric data type, so the mean and standard deviation of each attribute will be calculated.

While the Soil Type and Crop Type attributes are nominal data types so the probability value of each attribute will be calculated.

From equation (3), a prior probability value is needed, which is the probability value of a specific class. Prior probability values can be found by following the calculations below.

$$P(C_{10-26-26}) = \frac{Frequency\ of\ data\ appearance}{Total\ frequency\ of\ all\ data}$$

$$= \frac{7}{99} = 0{,}071$$

The calculation above is applied to all data classes so that all probability values are obtained from all specific classes, as shown in Table 5.

Table 5. Prior Probability

Class	10-26-26	14-35-35	17-17-17	20-20	28-28	DAP	Urea
Prob.	0,071	0,141	0,071	0,141	0,172	0,182	0,222

After the prior probability is obtained, equation (4) takes the conditional probability value, which has the probability value of each attribute. Because the probability value of an attribute stands alone, the conditional probability calculation can be done by applying the calculation below.

$$P(x_1|C_i) = \frac{Data\ frequency\ with\ a\ specific\ class}{Class\ frequency\ with\ a\ specific\ attribute}$$

$$P(Soil\ Type = Black|C_{10-26-26}) = \frac{0}{7} = 0{,}000$$

The calculation above is applied to all attributes with nominal data types so that conditional probability values are obtained from the Soil Type attribute as shown in Table 6. and Crop Type in Table 7.

Table 6. Conditional Probability Soil Type

Soil Type	*10-26-26*	*14-35-14*	*17-17-17*	*20-20*	*28-28*	*DAP*	*Urea*
Black	0,000	0,214	0,143	0,286	0,118	0,111	0,318
Clayey	0,429	0,214	0,000	0,214	0,235	0,111	0,227
Sandy	0,143	0,214	0,286	0,214	0,294	0,222	0,091
Loamy	0,143	0,286	0,286	0,214	0,118	0,222	0,227
Red	0,286	0,071	0,286	0,071	0,235	0,333	0,136

Table 7. Conditional Probability Crop Type

Crop Type	10-26-26	14-35-14	17-17-17	20-20	28-28	DAP	Urea
Maize	0,000	0,143	0,143	0,000	0,118	0,000	0,045
Sugarcane	0,000	0,143	0,429	0,286	0,000	0,111	0,091
Cotton	0,000	0,214	0,143	0,071	0,059	0,222	0,091
Tobacco	0,286	0,000	0,000	0,000	0,176	0,056	0,045
Paddy	0,143	0,143	0,000	0,071	0,118	0,056	0,136
Barley	0,143	0,071	0,143	0,000	0,118	0,111	0,000
Wheat	0,143	0,071	0,000	0,071	0,118	0,056	0,136
Millets	0,000	0,000	0,000	0,214	0,118	0,111	0,182
Oil seeds	0,000	0,143	0,000	0,143	0,000	0,056	0,091
Pulses	0,286	0,071	0,000	0,143	0,118	0,056	0,091
Ground Nuts	0,000	0,000	0,143	0,000	0,059	0,167	0,091

In Table 6 and Table 7, several conditional probability values with a value of 0 (zero) must be solved by applying Laplacian correction in equation (10) as calculated.

$$\rho_{Black|10-26-26} = \frac{0+1}{7+5} = 0{,}083$$

The calculation above is applied to all attributes with conditional probabilities with a value of 0 so that the new conditional probability values are obtained in Table 8 and Table 9.

Table 8. Conditional Probability Soil Type Applied by Laplacian Correction

Soil Type	*10-26-26*	*14-35-14*	*17-17-17*	*20-20*	*28-28*	*DAP*	*Urea*
Black	0,000	0,214	0,143	0,286	0,118	0,111	0,318
Clayey	0,429	0,214	0,000	0,214	0,235	0,111	0,227
Sandy	0,143	0,214	0,286	0,214	0,294	0,222	0,091
Loamy	0,143	0,286	0,286	0,214	0,118	0,222	0,227
Red	0,286	0,071	0,286	0,071	0,235	0,333	0,136

Table 9. Conditional Probability Crop Type Applied by Laplacian Correction

Crop Type	*10-26-26*	*14-35-14*	*17-17-17*	*20-20*	*28-28*	*DAP*	*Urea*
Maize	0,000	0,143	0,143	0,000	0,118	0,000	0,045
Sugarcane	0,000	0,143	0,429	0,286	0,000	0,111	0,091
Cotton	0,000	0,214	0,143	0,071	0,059	0,222	0,091
Tobacco	0,286	0,000	0,000	0,000	0,176	0,056	0,045
Paddy	0,143	0,143	0,000	0,071	0,118	0,056	0,136
Barley	0,143	0,071	0,143	0,000	0,118	0,111	0,000
Wheat	0,143	0,071	0,000	0,071	0,118	0,056	0,136
Millets	0,000	0,000	0,000	0,214	0,118	0,111	0,182
Oil seeds	0,000	0,143	0,000	0,143	0,000	0,056	0,091
Pulses	0,286	0,071	0,000	0,143	0,118	0,056	0,091
Ground Nuts	0,000	0,000	0,143	0,000	0,059	0,167	0,091

The mean value is calculated with equation (7) and the standard deviation with equation (8), as shown below for data with the numeric data type.

$$\mu_{10-26-26} = \frac{28+25+26+29+36+34+30}{7} = 29{,}714$$

$$\sigma_{10-26-26} = \sqrt{\frac{(28-29{,}714)^2 + \ldots. + (30-29{,}714)^2}{7-1}} = \sqrt{\frac{97.429}{7-1}} = 4{,}030$$

The calculation above is applied to all attributes that have numeric data types so that the mean values are obtained in Table 10, and the standard deviation values are in Table 11.

Table 10. Numeric Attribute Mean Value

Fertilizer Name	*Temparature*	*Humidity*	*Moisture*	*Nitrogen*
10-26-26	29,714	58,143	39,286	7,571
14-35-14	31,357	61,143	45,214	8,214
17-17-17	29,000	57,571	47,143	12,143
20-20	29,143	57,571	45,286	11,214
28-28	29,529	58,118	41,941	22,647
DAP	31,833	61,333	42,611	12,944
Urea	30,227	58,727	41,955	38,364

Table 11. Numeric Attribute Standard Deviation Value

Fertilizer Name	*Temparature*	*Humidity*	*Moisture*	*Nitrogen*
10-26-26	29,714	58,143	39,286	7,571
14-35-14	31,357	61,143	45,214	8,214
17-17-17	29,000	57,571	47,143	12,143
20-20	29,143	57,571	45,286	11,214
28-28	29,529	58,118	41,941	22,647
DAP	31,833	61,333	42,611	12,944
Urea	30,227	58,727	41,955	38,364

e. Evaluation

At this stage, the author uses the split validation method to calculate the accurate value of the model that has been formed. Using split validation, the data in Table 1 will be divided into test and training data, and perform a classification test on the test data. The following is Figure 3 shows the accuracy values obtained from the model that has been applied.

```
score = accuracy_score(y_test, predict)
score
```

```
0.6
```

```
print(metrics.classification_report(y_test, predict))
```

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       1.00      0.40      0.57         5
           2       0.00      0.00      0.00         3
           3       0.00      0.00      0.00         3
           4       1.00      1.00      1.00         3
           5       0.38      0.75      0.50         4
           6       1.00      1.00      1.00        10

    accuracy                           0.60        30
   macro avg       0.48      0.45      0.44        30
weighted avg       0.65      0.60      0.60        30
```

Figure 3. Accuracy Value of The Model

f. Deployment

At this stage, the author will use the Python programming language to build a classification tool so that it can be used to make decisions more quickly.

2. Application of Naive Bayes classification to test data

After the model has been formed, the model can be used to classify the test data to get the classification results from the model that has been formed. The test data used in this study can be seen in Table 12.

Table 12. Data Test

Temparature	*Humidity*	*Moisture*	*Soil Type*	*Crop Type*	*Nitrogen*	*Fertilizer Name*
12	60	53	Clayey	Paddy	18	-

From the data in Table 12, the nominal data type data will look for the probability value in Table 8 and Table 9. Data with a numeric data type will be calculated using equation (6) as calculated below with the example of class 10-26-26.

$$G_{(Temperature=12,\mu,\sigma)} = \frac{1}{\sqrt{2*\pi*4{,}030}} * e - \frac{(12-29{,}714)^2}{2*4{,}030^2} = 6{,}300*10^{-6}$$

$$G_{(Humidity=60,\mu,\sigma)} = \frac{1}{\sqrt{2*\pi*6{,}694}} * e - \frac{(60-58{,}143)^2}{2*6{,}694^2} = 0{,}057$$

$$G_{(Moisture=53,\mu,\sigma)} = \frac{1}{\sqrt{2*\pi*8{,}712}} * e - \frac{(53-39{,}286)^2}{2*8{,}712^2} = 0{,}013$$

$$P_{(Soil\ Type=Clayey|10-26-26)} = 0{,}333$$

$$P_{(Crop\ Type=Paddy|10-26-26)} = 0{,}111$$

$$G_{(Nitrogen=18,\mu,\sigma)} = \frac{1}{\sqrt{2*\pi*3{,}505}} * e - \frac{(18-7{,}571)^2}{2*3{,}505^2} = 1{,}361*10^{-3}$$

$$P(\boldsymbol{X}|C_{10-26-26}) = \left(6{,}300 * 10^{-6}\right) * 0{,}057 * 0{,}013 * 0{,}333 * 0{,}111 * \left(1{,}361 * 10^{-3}\right)$$
$$= 2{,}417 * 10^{-13}$$

The calculation above is applied to all test data for a specific class until all posterior probability values are obtained, as shown in Table 13.

Table 13. Posterior Probability

Class	$P(\boldsymbol{X}\|C_i).P(C_i)$	*Posterior Probability*
10-26-26	$(2{,}417 * 10^{-13}) * 0{,}071$	$1{,}709 * 10^{-14}$
14-35-35	$(2{,}291 * 10^{-26}) * 0{,}141$	$3{,}240 * 10^{-27}$
17-17-17	$(1{,}172 * 10^{-14}) * 0{,}071$	$8{,}289 * 10^{-16}$
20-20	$(2{,}036 * 10^{-15}) * 0{,}141$	$2{,}879 * 10^{-16}$
28-28	$(1{,}918 * 10^{-16}) * 0{,}172$	$3{,}293 * 10^{-17}$
DAP	$(3{,}683 * 10^{-16}) * 0{,}182$	$6{,}696 * 10^{-17}$
Urea	$(2{,}724 * 10^{-30}) * 0{,}141$	$6{,}054 * 10^{-31}$

From Table 13, it is found that the highest value is in class 10-26-26, so the data class sought in Table 12 can be filled with the class with the highest value obtained from the calculation process that has been carried out shown in Table 14.

Table 14. Data Test with Classified Classes

Temparature	*Humidity*	*Moisture*	*Soil Type*	*Crop Type*	*Nitrogen*	*Fertilizer Name*
12	60	53	Clayey	Paddy	18	10-26-26

3. Model accuracy value measurement

After the model is successfully obtained, the model needs to be measured to determine the accuracy of the model formed. Using Split Validation, Table 4 will be divided into 30% test data and 90% training data. The distribution of the classification results of 30% of the tested data against 70% of the trained data can be seen in Table 15.

Table 15. Distribution of The Classification Results

Temparature	*Humidity*	*Moisture*	*Soil Type*	*Crop Type*	*Nitrogen*	*Fertilizer Name*	*Classified*	*Explanation*
36	68	50	Loamy	Wheat	12	10-26-26	DAP	False
36	68	41	Red	Ground Nuts	41	Urea	Urea	True
29	58	30	Red	Tobacco	13	10-26-26	DAP	False
26	52	44	Sandy	Maize	23	28-28	28-28	True
33	64	50	Loamy	Wheat	41	Urea	Urea	True
34	65	38	Clayey	Paddy	39	Urea	Urea	True
29	58	65	Black	Cotton	14	DAP	DAP	True
30	60	58	Loamy	Sugarcane	10	14-35-14	20-20	False
28	54	38	Clayey	Pulses	40	Urea	Urea	True
31	62	48	Sandy	Maize	14	17-17-17	DAP	False
27	53	34	Black	Oil seeds	42	Urea	Urea	True
31	62	49	Black	Sugarcane	10	17-17-17	20-20	False
31	62	32	Red	Tobacco	39	Urea	Urea	True
30	60	61	Loamy	Cotton	8	14-35-14	14-35-14	True
33	64	39	Clayey	Paddy	13	20-20	DAP	False
28	54	41	Clayey	Paddy	36	Urea	Urea	True
33	64	34	Clayey	Pulses	38	Urea	Urea	True
27	53	35	Black	Oil seeds	37	Urea	Urea	True
32	62	41	Clayey	Paddy	24	28-28	28-28	True
34	65	62	Black	Cotton	7	14-35-14	14-35-14	True
31	62	44	Sandy	Barley	21	28-28	28-28	True
36	60	43	Sandy	Millets	15	DAP	DAP	True
32	62	30	Loamy	Sugarcane	38	Urea	Urea	True
25	50	64	Red	Cotton	9	20-20	17-17-17	False
26	52	31	Red	Ground Nuts	14	DAP	17-17-17	False
33	64	51	Sandy	Maize	5	14-35-14	10-26-26	False
25	50	56	Loamy	Sugarcane	11	17-17-17	20-20	False
28	54	43	Clayey	Paddy	10	14-35-14	20-20	False
33	64	31	Red	Ground Nuts	13	DAP	DAP	True
36	68	33	Black	Oil seeds	13	20-20	DAP	False

From Table 15, information is obtained from the 30 data tested, 18 correct results were obtained with the classes already contained in the tested Data so we have obtained data whose distribution is shown in the confusion matrix in Table 16.

Table 16. Confusion Matrix

Kelas	*10-26-26*	*14-35-14*	*17-17-17*	*20-20*	*28-28*	*DAP*	*Urea*
10-26-26	**0**	0	0	0	0	2	0
14-35-14	1	**2**	0	2	0	0	0
17-17-17	0	0	**0**	2	0	1	0
20-20	0	0	1	**0**	0	2	0
28-28	0	0	0	0	**3**	0	0
DAP	0	0	1	0	0	**3**	0
Urea	0	0	0	0	0	0	**10**

From Table 16, the classification results with accurate values will be divided by the total Count of data tested, as shown in the calculation below.

$$Accuracy = \frac{0+2+0+0+3+3+10}{30} = \frac{18}{30} = 0.6$$

From the above calculation, according to equation 1, the number of correct values is 18 divided by the sum of all valid values and incorrect values, which is 30. The accuracy value obtained is 0.6 or 60% which means that the accuracy value of the classification is good.

In his research, Lopez uses a confusion matrix to calculate the accuracy of the classification process for orthogonal defect automation (Lopez, 2020). Akshatha uses the confusion matrix to determine the accuracy of the fertilizer classification in the system for soil fertility (Akshatha, 2022). The results of their research state that the calculation of the accuracy of the classification method with the confusion matrix is considered good if it has an accuracy of more than 50%.

CONCLUSIONS AND RECOMMENDATIONS

From the discussion that has been carried out in this study, it can be concluded that the application of classification with the Naïve Bayes algorithm to the Fertilizer Prediction dataset gives a classification result in the form of the name of the fertilizer which is one of the 7 (seven) classes sought. The accuracy value obtained from modeling with the Naïve Bayes algorithm obtained a value of 60%, which is the result of calculations from test data obtained by the split validation method by calculating the amount of data classified accurately based on the confusion matrix. With an accuracy value of 60%, it can be concluded that the model formed is accurate enough to be used. Further researchers can add optimization methods such as GAFS or PSO into the Naïve Bayes model in this study to increase the accuracy value of the model that has been used in this study.

REFERENCES

Bhatia, P. (2019). *Data Mining and Data Warehousing: Principles and Practical Techniques.* Cambridge, United Kingdom: Cambridge University Press. Retrieved from www.cambridge.org/9781108727747

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques (3rd ed.).* Waltham, USA: Morgan Kaufmann Pub. doi: 10.1016/C2009-0-61819-5

Jha, G.K., Ranjan, P. and Gaur, M. (2020). A Machine Learning Approach to Recommend Suitable Crops and Fertilizers for Agriculture. In Recommender System with Machine Learning and Artificial

Intelligence (eds S.N. Mohanty, J.M. Chatterjee, S. Jain, A.A. Elngar and P. Gupta). doi: 10.1002/9781119711582.ch5

R. Priya, D. Ramesh, and E. Khosla, (2018) "Crop Prediction on the Region Belts of India: A Naïve Bayes MapReduce Precision Agricultural Model," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 99-104, doi: 10.1109/ICACCI.2018.8554948

Roy, A. H. (2017). Fertilizers and Food Production. In J. A. Kent, T. V. Bommaraju, & S. D. Barnicki, *Handbook of Industrial Chemistry and Biotechnology (13th ed.)* (pp. 757-804). Cham, Switzerland: Springer International Publishing.

Suntoro, J. (2019). *Data Mining : Algoritma dan Implementasi dengan Pemrograman PHP*. Jakarta, Indonesia: PT Elex Media Komputindo.

Walida, H., Harahap, F. S., Dalimunthe, B. A., Hasibuan, R., Nasution, A. P., & Sidabukke, S. H. (2020). Pengaruh Pemberian Pupuk Urea Dan Pupuk Kandang Kambing Terhadap Beberapa Sifat Kimia Tanah Dan Hasil Tanaman Sawi Hijau. Jurnal Tanah Dan Sumberdaya Lahan, 7(2), 283–289. doi: https://doi.org/10.21776/ub.jtsl.2020.007.2.12

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques (4th ed.).* United States: Morgan Kaufmann.

Zaki, M. J., & Meira, W. (2014). *Data Mining and Analysis: Fundamentals Concepts and Algorithms.* New York, USA: Cambridge University Press.

Zelle, J. M. (2017). *Python Programming: an Introduction to Computer Science (3rd ed.).* USA: Franklin, Beedle & Associates.

Wang, J. L., Liu, K. L., Zhao, X. Q., Zhang, H. Q., Li, D., Li, J. J., & Shen, R. F. (2021). *Balanced fertilization over four decades has sustained soil microbial communities and improved soil fertility and rice productivity in red paddy soil.* Science of The Total Environment, 793, 148664. doi: 10.1016/j.scitotenv.2021.148664

Pahalvi, H. N., Rafiya, L., Rashid, S., Nisar, B., & Kamili, A. N. (2021). *Chemical fertilizers and their impact on soil health*. *In Microbiota and Biofertilizers, Vol 2 (pp. 1-20). Springer, Cham.* doi: 10.1007/978-3-030-61010-4_1

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations, and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277-2293. doi: 10.1007/s00500-020-05297-6

Bhatt, M. K., Labanya, R., & Joshi, H. C. (2019). Influence of long-term chemical fertilizers and organic manures on soil fertility-A review. *Universal Journal of Agricultural Research*, *7*(5), 177-188. doi: 10.13189/ujar.2019.070502

Jha, G. K., Ranjan, P., & Gaur, M. (2020). A Machine Learning Approach to Recommend Suitable Crops and Fertilizers for Agriculture. *Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries*, 89-99. doi: 10.1002/9781119711582.ch5

De Souza, G. F. M., Melani, A. H. D. A., Michalski, M. A. D. C., & Da Silva, R. F. (Eds.). (2021). *Reliability Analysis and Asset Management of Engineering Systems*. Elsevier.

Lopes, F., Agnelo, J., Teixeira, C. A., Laranjeiro, N., & Bernardino, J. (2020). Automating orthogonal defect classification using machine learning algorithms. *Future Generation Computer Systems*, *102*, 932-947. doi: 10.1016/j.future.2019.09.009

Akshatha, G. C., & Shastry, K. A. (2022). Crop and Fertilizer Recommendation System Based on Soil Classification. In *Recent Advances in Artificial Intelligence and Data Engineering* (pp. 29-40). Springer, Singapore. doi: 10.1007/978-981-16-3342-3_3