Scientific Journal of Informatics
Vol. 8, No. 1, May 2021
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Implementation of Data Mining Using Naïve Bayes Classifier in Food Crop Prediction

Oki Arifin[1]*, Kurniawan Saputra[2], Halim Fathoni[3]
[1,2]Department of Informatics Management, Politeknik Negeri Lampung, Indonesia
[3]Industrial Engineering and Information Enterprise, Tunghai University, Taiwan

Abstract
Purpose: This study aims to developed modeling to prediction system of food crops by data mining, with Naïve Bayes Classifier (NBC), which expected will give information and can use by the farmer and industrial food crops.
Methods: On classification, progress attributes that use there is the temperature (°C), humidity (%), rainfall (mm), photoperiodicity (hour), and production result (ton) as a class attribute. The data of research that getting there are climate data and yield of food crops by data from the Central Bureau of Statistics (BPS) and the Meteorology, Climatology and Geophysics Agency (BMKG) from 2010 to 2017 at Lampung Province. Data of food crops used in this research there are paddy, maize, and soybean.
Result: The research results about the average accuracy of modeling that development using the 10-fold cross-validation method, that had an accuracy value of 72.78% and Root Mean Square Error (RMSE) there is 0.438.
Novelty: Prediction system of food crops by data mining.

Keywords: Prediction, Food Crops, Data Mining, Naïve Bayes
Received January 2021 / **Revised** March 2021 / **Accepted** March 2021

INTRODUCTION

Indonesia is an agricultural country because most of its population works in the agricultural sector and is a place where food crops are grown [1]. Agriculture became one of the sectors that have an important role to improve the country's economy and meet the food needs of the community. This is stated in Law No. 18 of 2012 concerning Food which states that the purpose of food administration is to increase the ability to produce food independently. Lampung Province has development activities that are oriented towards the potential of natural resources in the agricultural sector, especially the food crop sub-sector.

Food crops are all kinds of crops in which there are carbohydrates, proteins, minerals, and vitamins as a source of human energy [2]. Food crops can also be regarded as a major crop consumed by humans as food to provide energy for the body's intake and survival. According to data from the Central Statistics Agency (BPS), the types of food crops and their forms of production are divided into seven groups which include paddy (milled dry grain), maize (dry shelled), soybean (dry seed), peanut (dry seed), green bean (dry seed), cassava (wet tuber), and sweet potato (wet tuber).

The phenomenon of climate anomalies is climate change and variability that is a serious concern because it has a major impact, especially on the food crops subsector [3]. Many people in land management still use manual systems, so they must pay attention to weather calculations so that crop failure does not occur, the weather that often changes greatly affects the amount of harvest, especially for food crops [4][5]. The prediction of food crop yields is one of the most critical problems encountered in the agricultural sector, this is due to the farmers' lack of knowledge about bountiful harvests, uncertain weather conditions, and fluctuating market prices that affect crop production. A method is needed to classify crop yields based on the influence of weather which is often one of the important factors affecting crop yields. The Naive Bayes Classifier (NBC) method is one of the most popular classification methods in data mining techniques

*Corresponding author.
Email addresses: okiarifin@polinela.ac.id (Arifin), kurniawan_mi@polinnela.ac.id (Saputra), d07330701@thu.edu.tw (Fathoni)
DOI: 10.15294/sji.v8i1.28354

because it is easy to interpret and easy to understand and the results of predictions can help in making decisions.

Data mining can also be interpreted as extracting new information taken from large portions of data that help in decision-making [6]. The classification process in data mining can use the classification method or supervised learning. Classification is the process of identifying objects into a category, class, or group based on predetermined procedures, definitions, and characteristics [7]. In building data mining, a training data set is required as the basis for prediction, then the classification method is determined according to the characteristics of the training data. Training data can be categorized into categories and continuous. Categories have data with discrete values or levels such as insufficient, moderate, and moderate. Meanwhile, continuous training data is in the form of continuous values or numbers [8]. To measure the accuracy of a system built using ten-fold cross-validation or often known as 10-fold cross-validation is a testing method for trained learning (supervised learning). Where in each fold it is divided into 10 subsets with the same size for each subset [6].

Based on research [2] conducted research using the J48 decision tree algorithm with data mining techniques. This study aims to make a prediction system model for food crops in North Sulawesi province to determine the effect of weather on crop yields. From the results of experiments conducted with existing data, the average accuracy value is 69.48%. Then the research [1] using NBC aims to recommend planting food crops in the Special Region of Yogyakarta Province based on weather, yields and selling prices with an accuracy of 85.71%.

Based on previous research, it will be reviewed studies on the implementation of data mining methods for the prediction of crop NBC in Lampung Province. Different from the data taken and the methods used. To measure the performance of the classification model using confusion matrix and Root Mean Square Error (RMSE). So that a food crop prediction system modeling will be made using data mining techniques with the NBC method to find out the effect of weather on crop production so that it can be used as a basis for decision making.

METHODS

Based on this research, we developed a food crop prediction system modeling using data mining techniques using the NBC method. The prediction system can be used to assist farmers in determining the types of crops to be planted by considering the weather conditions and in the hope of increasing food crop production. Also, this system can also assist decision-makers in industries that manage food crop products, especially in Lampung Province. The methodological stages used are data collection, data pre-processing, model development, and the final model as shown in Figure 1.

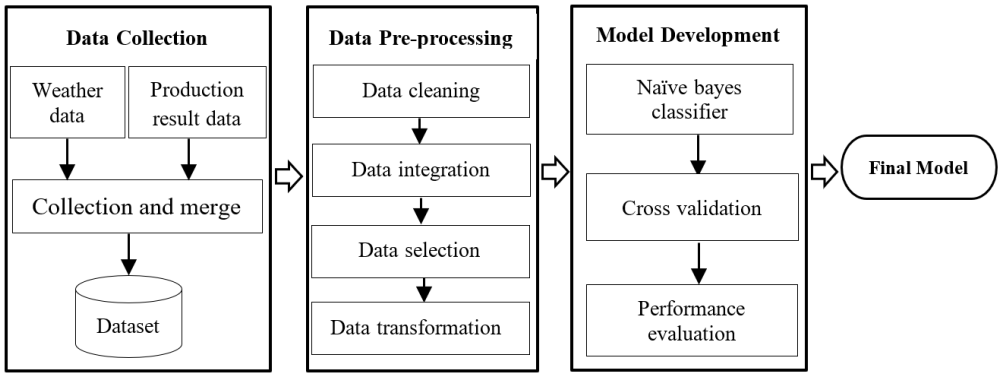

Figure 1. Research stage

Data Collection

At this stage, data is obtained from weather data and food crop production results from online data from the Central Statistics Agency (BPS) and the Meteorology, Climatology and Geophysics Agency (BMKG) [9][10]. The food crops data used in this study were 3 food crops with the highest production yields, namely paddy, maize, and soybeans from 2010 to 2017 in Lampung province. The amount of data obtained is 120

data for each type of food crop. The following is a sample of online paddy data from BPS and BMKG which is shown in Table 1 and Table 2.

Table 1. Production of paddy from BPS

Region	Production of Paddy by Regency/City (Ton)							
	2010	2011	2012	2013	2014	2015	2016	2017
Lampung Barat	160080	165342	177810	177810	121668	112063	141374	147606
Tanggamus	208553	201067	212317	226628	222360	284643	283379	354549
Lampung Selatan	370060	395437	399900	441113	434969	488079	494629	579534
Lampung Timur	431981	443552	492315	509949	494722	564315	638817	662291
Lampung Tengah	570968	654545	660443	673564	765007	782604	805261	733033
Lampung Utara	117088	131155	139319	150339	153627	168942	196136	214329
Way Kanan	120487	145472	137161	151674	158051	149178	209076	219282
Tulang Bawang	187412	186728	185674	186781	228049	242728	291031	329220
Pesawaran	139159	146317	150526	153472	146428	170073	205442	214455
Pringsewu	111239	113284	113342	120275	134274	137193	156541	136796
Mesuji	113822	87195	144304	129791	132000	186216	186230	266847
Tulang Bawang Barat	60245	49155	66182	73473	79606	88443	95839	111288
Pesisir Barat	-	-	-	72506	72213	77605	84751	85335
Bandar Lampung	9336	8631	6752	9220	8966	9997	10201	10190
Metro	23443	24988	22555	27027	18251	34410	33216	25899
Provinsi Lampung	2623873	2752869	2908600	3042419	3170191	3496489	3831923	4090654

Table 2. Average temperature (Celsius) from BMKG

Month	Average Temperature (Celcius)							
	2010	2011	2012	2013	2014	2015	2016	2017
January	26.4	26.3	26.5	26.2	25.8	26.2	27.5	26.6
February	27	26.6	26.4	26.7	26.6	26.6	27.1	26.4
March	26.6	26.3	26.8	27	27.1	26.9	27.6	26.8
April	27.8	27	26.9	27.1	27.4	27.1	27.4	27.2
May	27.5	27.2	27.4	27.1	27.5	27.4	27.7	27.1
June	26.2	26.7	26.5	27.2	27.2	27.2	27.1	26.7
July	25.9	26.1	26.2	25.8	27	27	26.7	26.5
August	26.3	26.6	26.4	26.2	26.8	27.3	27	26.5
September	26.2	27.6	27.2	26.8	27.7	27.7	27.2	27.4
October	27.1	27.1	27.7	27.4	28.2	28.7	27	27.7
November	26.7	27.2	27.3	26.8	27.2	28.2	26.9	27.3
December	26.6	26.9	26.6	25.8	26.6	27.2	27	26.7

Data Pre-processing

At this stage, the data has been collected and combined to become a dataset, then the preprocessing process is carried out, which is divided into several stages including:

a. Data cleaning is the process of removing inconsistent data and removing noise. The data obtained are then cleaned where the missing value in all data, both weather data and production data, is replaced with a value of 0 (zero).
b. Data integration is the process of combining data if you have a data source in the data mining system.

c. Data selection is the collection of relevant data that will be used in the data mining process.
d. Data transformation is a process where data is transformed into forms suitable for data mining processes. The data that has been cleaned is then carried out by a transformation process, in this case changing the numerical data on the production price is transformed into nominal data. This process is carried out using the binning method, namely Equal Width Binning. This method divides the data into k intervals of equal size. Equation (1) in Equal Width Binning [2].

$$w = (max - \min)/k \quad (1)$$

w is the value that determines the boundary value for each k interval. The obtained interval limits are $min + w$, $min + 2w$,…, $min + (k\text{-}1)\ w$. Based on this method, the production data are categorized into 2 class label classifications, namely low and high. The low class states that the production results are below the average value, while the high class states that the production results are above the average value. This production yield parameter will then be used as a classification parameter for each type of food crop. The final result of this stage is a training dataset that will be used for system modeling.

Development Model

At this stage, make a system model using the RapidMiner Studio 9.6 software. The process of classifying weather data and crop production uses the NBC algorithm. The Bayes theorem algorithm is an algorithm that refers to the concept of probability for all independent or interdependent attributes given by values in class variables [11]. Naïve Bayes Classifier (NBC) or commonly called Bayesian Classification is a classification method with probability and statistical methods based on the Bayes theorem, which can be used to predict the probability of class membership [12]. NBC is the most effective and efficient algorithm for machine learning and data mining [13]. NBC is proven to have high accuracy and speed when applied to large databases. The training dataset is a subset used to train classification algorithms by using known values to predict the future and unknown values [14]. Equation (2) in the Naïve Bayes Classifier [15].

$$P(H|X) = \frac{P(X|H)\ P(H)}{P(X)} \quad (2)$$

Where is the formula for equation (2):

P(H|X) = H hypothesis probability based on condition X (posterior probability)
H = Data hypothesis X is a specific class
X = Data with an unknown class
P(X|H) = Probability of X based on the conditions in hypothesis H
P(H) = The initial (prior) hypothesis of the H hypothesis occurs regardless of any evidence.
P(X) = The initial probability (prior) of evidence X occurs regardless of the hypothesis / other evidence.

To determine the performance of the model using 10-fold cross-validation so that it is known how accurate the model is made to predict these food crops. Ten-fold cross-validation or often known as 10-fold cross-validation is a technique for estimating the performance of the training model that has been built [6]. Where in each fold divides the data into 10 subsets with the same size for each subset. After that, doing training ten times using 9 folds is used to create a training model and 1 other fold is used as test data. At each iteration, the error value is calculated and accumulated then divided by ten to get the overall error value [16].

To measure the resulting performance using RMSE. The results of the calculation of RMSE ≤ 1. The results of the smallest RMSE calculations, the better the accuracy [17]. Equation (3) on RMSE [18].

$$RMSE = \frac{\sum_{i=1}^{n} \sqrt{(x_i - f_i)^2}}{n} \quad (3)$$

Where the variable n is the amount of data and the variable x is the predictive value generated by the model, while the variable f is the actual desired value.

Final Model
This stage is the final result of the process of creating and training the algorithms used. This model is expected to be a recommendation model for predicting food crops to determine production yields in Lampung Province.

RESULT AND DISCUSSION
Parameters used in the classification process use the NBC method with annual data, namely temperature (°C), humidity (%), rainfall (mm), photoperiodicity (hours), and production result (ton) as attributes class. The data collection collected was 120 data from each type of food crop, namely paddy, maize, and soybeans. The distribution and testing of the dataset use 10-fold cross-validation. The purpose of this test is to analyze the NBC model algorithm based on the level of accuracy, precision, recall, and RMSE values.

Paddy Dataset Testing
Dataset on paddy collected was 120 obtained from BMKG and BPS. The following is a sample dataset for paddy food crops, as shown in Table 3.

Table 3. Paddy sample dataset

No	Temperature	Humidity	Photoperiodisity	Rainfall	Production Result
1	26.7	82.3	50.6	223.7	high
2	26.8	77.8	56.4	130.7	high
3	26.8	79.2	64.6	133.8	high
4	26.7	81.3	64.2	204.7	high
5	27.1	80	61.7	140.2	low
6	27.3	78.3	67.9	135.7	low
7	27.2	83	58.1	193.1	low
8	26.9	80.7	56.1	152.1	high
9	26.7	82.3	50.6	223.7	low
10	26.8	77.8	56.4	130.7	low
…	…	…	…	…	…
120	26.9	80.7	56.1	152.1	low

The results of the classification performance analysis for the paddy dataset using the confusion matrix are shown in Table 4.

Table 4. Result of confusion matrix paddy dataset

Class	Accuracy	Precision	Recall
high	67.50%	64.44%	55.77%
low		69.33%	76.47%

In Table 4, the accuracy of the model using the NBC algorithm with the paddy dataset produces an accuracy of 67.50%. Meanwhile, the RMSE value is 0.436, which is shown in Figure 1.

Figure 1. RMSE value for paddy dataset

Maize Dataset Testing
Dataset on maize collected was 120 obtained from BMKG and BPS. The following is a sample dataset for maize food crops, as shown in Table 5.

Table 5. Maize sample dataset

No	Temperature	Humidity	Photoperiodisity	Rainfall	Production Result
1	26.7	82.3	50.6	223.7	high
2	26.8	77.8	56.4	130.7	high
3	26.8	79.2	64.6	133.8	high
4	26.7	81.3	64.2	204.7	low
5	27.1	80	61.7	140.2	low
6	27.3	78.3	67.9	135.7	low
7	27.2	83	58.1	193.1	low
8	26.9	80.7	56.1	152.1	low
9	26.7	82.3	50.6	223.7	high
10	26.8	77.8	56.4	130.7	high
…	…	…	…	…	…
120	26.9	80.7	56.1	152.1	high

The results of the classification performance analysis for the maize dataset using the confusion matrix are shown in Table 6.

Table 6. Result of confusion matrix maize dataset

Class	Accuracy	Precision	Recall
high	79.17%	70.00%	56.76%
low		82.22%	89.16%

In Table 6, the accuracy of the model using the NBC algorithm with a maize dataset yields an accuracy of 79.17%. Meanwhile, the RMSE value is 0.431, which is shown in Figure 2.

Figure 2. RMSE value for maize dataset

Soybean Dataset Testing

Dataset on soybean collected was 120 obtained from BMKG and BPS. The following is a sample dataset for soybean food crops, as shown in Table 7.

Table 7. Soybean sample dataset

No	Temperature	Humidity	Photoperiodisity	Rainfall	Production Result
1	26.7	82.3	50.6	223.7	low
2	26.8	77.8	56.4	130.7	low
3	26.8	79.2	64.6	133.8	low
4	26.7	81.3	64.2	204.7	low
5	27.1	80	61.7	140.2	low
6	27.3	78.3	67.9	135.7	high
7	27.2	83	58.1	193.1	low
8	26.9	80.7	56.1	152.1	low
9	26.7	82.3	50.6	223.7	low
10	26.8	77.8	56.4	130.7	high
…	…	…	…	…	…
120	26.9	80.7	56.1	152.1	low

The results of the classification performance analysis for the soybean dataset using the confusion matrix are shown in Table 8.

Table 8. Result of confusion matrix soybean dataset

Class	Accuracy	Precision	Recall
high	71.67%	25.00%	6.67%
low		75.00%	93.33%

In Table 8, the accuracy of the model using the NBC algorithm with the soybean dataset produces an accuracy of 71.67%. Meanwhile, the RMSE value is 0.446, which is shown in Figure 3.

Figure 3. RMSE value for soybean dataset

Soybean Dataset Testing

Following are the results of accuracy for each type of food crop, as shown in Table 9.

Table 9. Performance results of the naïve bayes classifier

No	Type of Food Crops	Accuracy	RMSE
1	Paddy	67.50%	0.436
2	Maize	79.17%	0.431
3	Soybean	71.67%	0.446

Based on Table 9, it can be seen that the accuracy value of the model built using the NBC method on different types of food crops. This happens because of differences in production per year. The results of the overall accuracy and average RMSE of the model are shown in Figure 4.

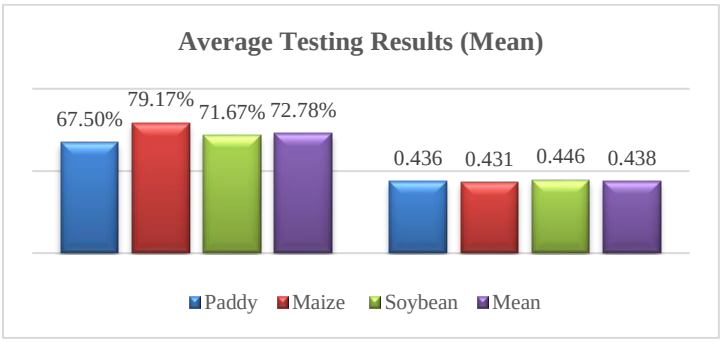

Figure 4. Average testing results (mean)

Figure 5 shows that the average accuracy of the model is 72.78% and the RMSE performance is 0.438. Based on these results, the model we have developed can be taken into consideration for decision-makers in industries that manage food crop products in Lampung Province.

CONCLUSION

This study succeeded in building a modeling system capable of classifying food crops using the Naïve Bayes method. Based on the results of the performance level with 10-fold cross-validation and RMSE with 120 data for each type of food crop with an accuracy value of 67.50% paddy, 79.17% maize, and 71.67% soybeans. The RMSE value of paddy was 0.436, maize was 0.431, and soybean was 0.446. For the overall average accuracy value of the types of food, crops are 72.78% and RMSE 0.438. These results indicate that the model developed can be used to assist farmers in determining the types of food crops to be planted by considering whether factors in the hope that production yields will increase so that they can be used as a basis for decision making. The results of this study need to be further developed by comparing other classification algorithms for better accuracy and adding other parameters such as soil conditions, topography, and water potential.

REFERENCES

[1] T. Setiadi, F. Noviyanto, H. Hardianto, A. Tarmuji, A. Fadlil, & M. Wibowo, "Implementation of Naïve Bayes Method in Food Crops Planting Recommendation," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 4750–4755, 2020.

[2] F. J. Kaunang, R. Rotikan, & G. S. Tulung, "Pemodelan Sistem Prediksi Tanaman Pangan Menggunakan Algoritma Decision Tree Crop Prediction System Using Decision Tree Algorithm," *CogITo Smart J.*, vol. 4, no. 1, pp. 213–218, 2018.

[3] Y. Apriyana, E. Susanti, & F. Ramadhani, "Analysis of Climate Change Impacts on Food Crops Production in Dry Land and Design of Information System," *J. Inform. Pertan.*, vol. 25, no. 1, pp. 69–80, 2016.

[4] Zhao et al., "Temperature Increase Reduces Global Yields of Major Crops in Four Independent Estimates," *Proc. Natl. Acad. Sci. U. S. A.*, Indonesia: PNAS, 2017.

[5] M. Springmann, D. Mason-D'Croz, S. Robinson, T. Garnett, H. C. J. Godfray, D. Gollin, M. Rayner, P. Ballon, & P. Scarborough, "Global and Regional Health Effects of Future Food Production Under Climate Change: A Modelling Study," *The Lancet*, vol. 387, no. 10031, pp. 1937–1946, 2015.

[6] O. Arifin, & T. B. Sasongko, Analisa Perbandingan Tingkat Performansi Metode Support Vector Machine dan Naïve Bayes Classifier," *Semin. Nas. Teknol. Inf. dan Multimed*, Indonesia: Researchgate, 2018.

[7] O. Arifin, "Sistem Klasifikasi Berita Daring Faktor Kejahatan Penyalahgunaan Narkotika Berbasis Algoritma Naive Bayes," *Telematika*, vol. 11, no. 2, pp. 27, 2018.

[8] S. Maesaroh, "Sistem Prediksi Produktifitas Pertanian Padi Menggunakan Data Mining," *Energy*, vol. 7, no. 2, pp. 25–30, 2017.

[9] BPS, "Tanaman Pangan. Badan Pusat Statisktik Indonesia," 2020. [Online]. Available: https://www.bps.go.id/subject/53/tanaman-pangan.html.

[10] BMKG, "Data Online Pusat Database – BMKG," 2020. [Online]. Available: http://dataonline.bmkg.go.id/home.

[11] F. Y. Azalia, M. A. Bijaksana, & A. F. Huda, "Name Indexing in Indonesian Translation of Hadith using Named Entity Recognition with Naïve Bayes Classifier," *Procedia Comput. Sci.*, vol. 157, pp. 142–149, 2019.

[12] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, & I. Cholissodin, "Optimasi Naïve Bayes Classifier dengan Menggunakan Particle Swarm Optimization pada Data Iris," *J. Teknol. Inform. dan Ilmu Komput.*, vol. 4, no. 3, pp. 180, 2017.

[13] A. R. Susanti, T. Djatna, & W. A. Kusuma, "Twitter's Sentiment Analysis on GSM Services Using Multinomial Naïve Bayes," *Telkomnika (Telecommun. Comput. Electron. Control)*, vol. 15, no. 3, pp. 1354–1361, 2017.

[14] A. Wibowo, D. Manongga, & H. D. Purnomo, "The Utilization of Naive Bayes and C.45 in Predicting The Timeliness of Students' Graduation," *Sci. J. Inform.*, vol. 7, no.1, pp. 99–112, 2020.

[15] P. L. Kumalasari, & R. Arifudin, "Decision Making System to Determine Childbirth Process with Naïve Bayes and Forward Chaining Methods," *Sci. J. Inform.*, vol. 7, no. 2, pp. 180–188, 2020.

[16] D. Sofi, "Analisis dan Prediksi Kinerja Mahasiswa Menggunakan Teknik Data Mining," *Syntak*, vol. 2, pp. 1–10, 2013.

[17] W. T. Parmadi, & B. M. Sukojo, "Analisa Ketelitian Geometric Citra Pleiades Sebagai Penunjang Peta Dasar RDTR," *J. Tek. ITS*, vol. 5, no. 2, pp. A411–A415, 2013.

[18] T. Chai, R. R. Draxler, & C. Prediction, "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against avoiding RMSE in The Literature," *Geosci. Model Dev.*, vol. 7, pp. 1247–1250, 2014.