Implementation of Naïve Bayes Algorithm to Predict Food Crop Production Results

Yoga Handoko Agustin
Informatics Engineering
Garut Institute of Technology
Garut, Indonesia
yoga.handoko@itg.ac.id

Vini Oktapiani
Informatics Engineering
Garut Institute of Technology
Garut, Indonesia
2006059@itg.ac.id

Abstract- The ups and downs of food crop production each year are caused by changes in the area of land planted each year. These changes are influenced by several factors, including crop rotation, government policies, changes in agricultural practices, environmental factors such as climate, and economic pressures. In an effort to improve the efficiency and productivity of food crop production in Garut Regency, the use of technology and data analysis methods is becoming increasingly important. This research aims to predict food crop production in Garut Regency with Naïve Bayes algorithm and evaluate influential factors. This modeling is analyzed using Feature Forward selection and SMOTE techniques to determine the most influential attributes and overcome class imbalance. The method used is Cross-Industry Standard Process For Data Mining (CRISP-DM). Where the use of SMOTE successfully handles unbalanced classes, and the application of Feature selection results in the 5 most influential factors, namely crop type, added planting, realized harvest area, realized production and production. The results showed that the Naive Bayes model with Cross validation and Xgboost resulted in an Accuracy value of 82.54%, Recall value of 81.67%, Precision value of 83.34%. And the AUC value is 0.904% with the Good Classification category.

Keywords- Naïve Bayes algorithm, Feature Forward selection, Prediction, Agriculture, SMOTE technique

I. INTRODUCTION

The prediction of agricultural yields of food crops is heavily influenced by climate change. Therefore, annual weather change is an independent variable that can affect the yield of food crop production [1]. In addition to weather changes that can affect the ups and downs of food crop production, there are also changes in building land in agricultural areas resulting in reduced availability of food needed to meet community needs. Therefore, the availability of land is very important because the average production on agricultural land affects the area of planting and harvesting. To overcome this situation, it is necessary to use a Naïve Bayes method that can predict the yield of food crops [2].

Prediction is the process of forecasting the most likely future state using past data or current and past information [3], [4]. For this reason, the prediction process can be carried out using the Naïve Bayes method. The application of the Naïve Bayes method to perform prediction results in agriculture has been carried out in several studies, including the first study by [5] The research aims to forecast the production of three types of agricultural crops (corn, dryland rice, and wetland rice) in Satar Mese Sub-district, Manggarai Regency. (DES). The second research conducted by [6] The research predicts the number of palm oil harvests with the Naïve Bayes algorithm, where the results obtained are very good with 100% accuracy. The third research conducted by [7] analyzed the prediction of rice yields using the Naïve Bayes classifier algorithm, which resulted in a fairly accurate calculation with an accuracy of 89%. The fourth research conducted by [8] The research focused on the agricultural sector which experienced fluctuations in food crop production. The algorithm used in the research is K-Nearest Neighbor (K-NN). The research resulted in an accuracy rate of 92.83 percent for the corn commodity. The fifth research conducted by [9] perform data balancing on rice production prediction using Naive Bayes and CART. The model with greater accuracy is using the CART algorithm with the use of the SMOTE technique, the accuracy value before balancing the data is 47.67% and after balancing it is 55.73%.

Based on the results of previous research references that discuss the prediction of food crop production, this research uses the attributes Year, Sub-district, Type of food crop,

Realization of Planted Area (Ha), Realization of Net Harvested Area (Ha), Production (Ton) and Productivity (Kw/Ha). Based on the data obtained, there is an imbalanced class between the production results Up,Down therefore, another technique is needed, namely the SMOTE technique. This technique is used to overcome class imbalance in data [10]. This research is expected to produce an appropriate prediction model for food crop production and identify the factors that have the most influence on food crop production in Garut Regency.

II. METHODOLOGY

The methodology used in this research is the Cross Industry Standard Process for Data Mining (CRISP-DM). CRISPDM is a structured and thorough approach to the data mining process. This method consists of six phases [11]. Figure 1 shows the process stages.

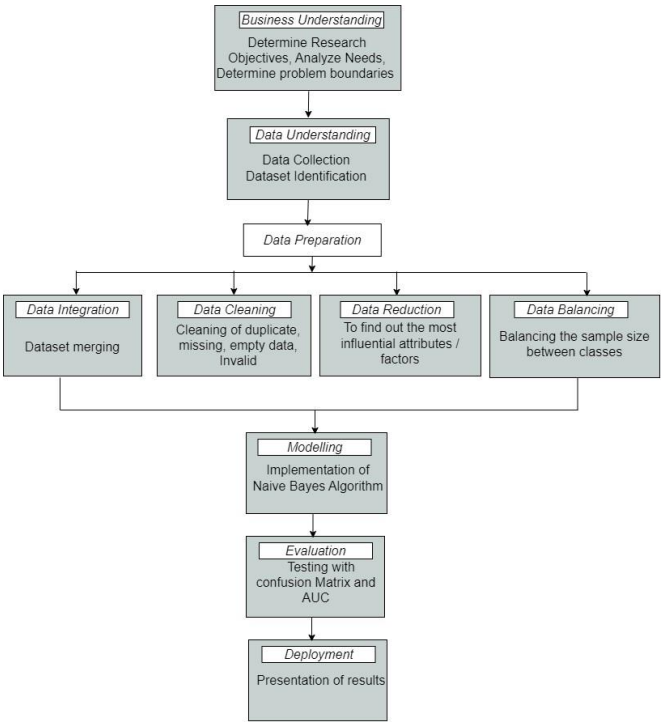

Figure 1: Research Stages

A. *Business Understanding*

This first stage is carried out to gain a more detailed understanding of the problems of food crop production . The steps to be taken include a literature study of previous research relevant to the problem. In addition, the steps taken are interviews to strengthen the research to be carried out [12].

B. *Data Understanding*

At this stage the process of collecting, understanding and preparing data will be used in the analysis [13]. Where data understanding includes preparation, checking the data used, collecting initial data, and identifying data quality. In data understanding, the data used will undergo a process that explains each of its features [14].

C. *Data Preparation*

Data Preparation is a data preparation process that involves transforming raw data into analysis-ready datasets. At this stage, preprocessing is carried out by building dataset integration, cleaning, reduction and balancing [15]. Where data is merged if the data obtained is separated, then delete missing values on attributes that have empty data, perform a selection process to identify the most influential attributes using forward selection, and perform class balancing if there are unbalanced classes using the SMOTE technique. And here are the data preprocessing stages in Figure 2.

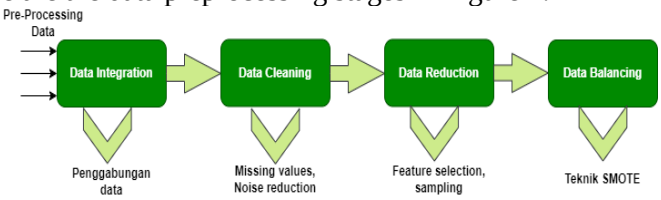

Figure 2: Data Preprocessing Stages

D. *Modelling*

This is the stage of implementing appropriate modeling techniques, displaying and identifying patterns [16]. After obtaining data that is ready to be processed, the next step is to choose the right modeling technique to get the best results. In this research, the modeling technique chosen to develop the classification model is using the Naïve Bayes algorithm.

E. *Evaluation*

The evaluation is done by measuring the results of the algorithm on training data or training using confusion matrix as the main tool for performance evaluation [9]. The evaluation carried out will show the resulting performance, namely accuracy. The following is the accuracy formula that will be used in this research [17]:

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

Description: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), *False Negatif* (FN).

F. *Deployment*

The deployment stage includes preparing reports from all the data that has been processed, developing and visualizing the data [18]. At this stage, the resulting performance model needs to be presented in a format that can be easily understood by stakeholders.

III. RESULTS AND DISCUSSION

The following are the results of research on the implementation of the Naïve Bayes algorithm to predict food crop production, the stages are in accordance with the method used.

A. *Business Understanding*

Based on the initial stages of understanding the research refers to the study of previous research literature and identifying problems by conducting interview activities with the Garut Regency Agriculture Office regarding food crop production problems. Where the problem that often occurs in food crop production is the change in the area of land planted each year, this change is influenced by several factors. Therefore, it is necessary

to implement predictions, one of which is with machine learning techniques, especially the Naive Bayes algorithm.

B. *Data Understanding*

Data Understanding is the process of understanding data. The activities carried out are data collection and dataset identification. The explanation is as follows:

1) Data Collection

At the data collection stage, the process of searching for data to be used is carried out. Where the data is obtained from the Garut Regency Agriculture Office in 2020 - 2023 with a total of 1344 records. The following is the raw data shown in Table 1.

Table 1.
Raw Data

No	District	Increase Planting	Harvested Area	Production	…	Productivity
1	Cisewu	5980	5772	38444	…	66.6
2	Caringin	2851	2753	17994	…	65.4
3	Talegong	5352	5163	33910	…	65.7
…	…	…	…	…	…	…
560	Peundeuy	44	44	54	…	12,2
561	Cikajang	810	77	234	…	97,3
…	…	…	…	…	…	…
1343	Limbangan	80	80	1159	…	114,8
1344	Selaawi	31	31	452	…	145,8

2) Dataset Identification

The next step is to identify the dataset and determine the attributes used, namely Year, Sub-district, Type of food crops, Target of added planting, Realization of added planting, Added planting, Target of harvest area, Realization of harvest area, Harvest area, Production Target, Production Realization, Production, Productivity Target, Productivity Realization, Productivity and Class shown in Table 2 below.

Table 2.
Attributes of the Data

No	Attributes	Data Type	Indicator
1	Year	*Integer*	2020,2021,2022,2023
2	District	Categorical	
3	types of food crops	Categorical	Paddy, upland rice, Maize, Soybean, Groundnut, Green Bean, Cassava, Sweet Potato
4	target for additional planting	*Float*	-
5	realization of additional planting	*Float*	-
...	...	...	...
16	Class	Categorical	Up, Down

Where in Table 2. there are 16 attributes, where 12 with *float* data type, 1 with *integer* data type and 3 with categorical data type.

C. *Data Preparation*

Data Preparation is the processing of raw data that has previously been obtained in the previous stage. The following is the preprocessing stage:

1) Data *Integration*

The integration stage is the process of merging several separate data from different sources into one, into a new data storage file. The data process is combined from 2020,

2021, 2022 and 2023. As for determining the class label based on the Production value. The following is an illustration of data merging in the year shown in Figure 3.

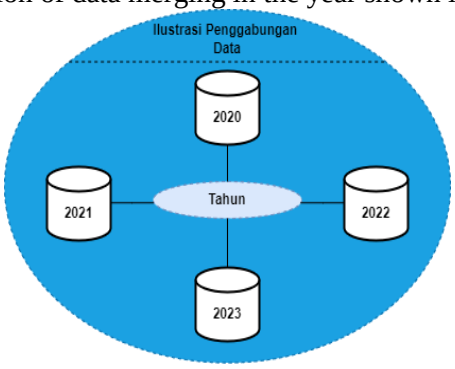

Figure 3. Illustration of Merging Year Data

2) Data *Cleaning*

The Data Cleaning stage is the stage of cleaning data or replacing values from invalid, duplicate or empty data. At this stage, the value of the data that has a missing value is changed to the average value of each attribute so that there are no empty values. The following is a sample of data that has been done data cleaning shown in Table 3.

Table 3.
Average Value on Attributes

No	Attributes	Average Value
1	Sasaran tambah tanam	582
2	Realisasi tambah tanam	634
3	Tambah tanam	644
...	...	...
7	Sasaran produksi	3.171
8	Realisasi produksi	3.928

Table 3. shows the average value of each attribute except Year, District, Type of Food Crops and Production. The average value is obtained from the summation of all values of each attribute divided by the amount of data. Then the value will replace records that have a Missing value so that there are no empty attributes.

3) Data *Reduction*

The next step is to reduce the data. This data reduction is done through a selection process, which will later be known attributes/factors that most affect the production of food crops. The reduction process applied in this research is using Forward selection which is presented in Table 4 below.

Table 4.
Forward selection result attributes

No	*Attribute*	*Weight*
1	District	0
2	Type of crop	1
3	Target t.tanam	0
4	Planting realization	0
5	Increase in planting	1
6	Targeted harvest area	0
7	Harvest area realization	1
8	Harvest area	0
9	Production target	0
10	Production realization	1
11	Production	1
12	Productivity goal	0
13	Productivity realization	0
14	Productivity	0

Therefore, the attributes that have the most influence on the production of food crops in Garut Regency are crop type, additional planting, realization of harvest area, Production Realization, Production.

4) Data *Balancing*

The balancing process is carried out because there are imbalanced classes in the up and down production classes. Synthetic Minority Over Sampling Technique (SMOTE) is used to augment the minority class data with synthetic data, so that the number of samples from the majority class is equal to the number of samples from the minority class. This technique is performed to balance the data in the two classes [9]. It can be seen in Figure 4 that there is data imbalance with 635 for class 0 (Up) and 709 for class 1 (Down). The following are the results before the data is balanced and after the data is balanced shown in Figure 4.

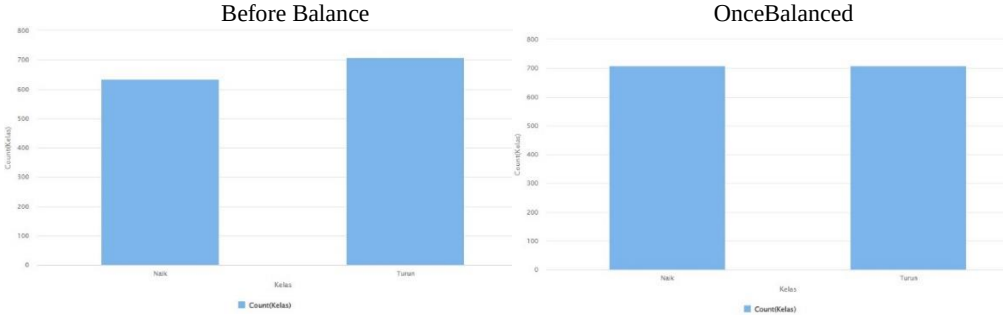

Figure 4. Before and After Oversampling Technique

Figure 4 shows the results before the data is balanced and after the data is balanced, where the left image shows the data before it is balanced, but after the smote oversampling technique in RapidMiner there is a change where the right image shows that the data is balanced, so that the total data increases from 1344 to 1418.

D. *Modelling*

The next step is to choose the right modeling technique. The modeling used is Naive Bayes.

Naive Bayes is designed to predict future probabilities with existing data from previous events, hence it is often known as Bayes' Theorem. [19]. Naïve Bayes is a technique that categorizes problems into categories based on similar or different characteristics, using statistics that can estimate the probability of each category. [20]. To find a classification model, the Naïve Bayes algorithm is applied using Feature Forward selection and SMOTE. In finding a classification model, namely by applying the Naïve bayes algorithm using Feature Forward selection and SMOTE.

To find To assess the performance of modeling results, a tool is needed to determine the accuracy value obtained. The tools used in this research are RapidMiner tools. And the following are the stages of Naive Bayes modeling in RapidMiner along with its operators shown in Figure 5.

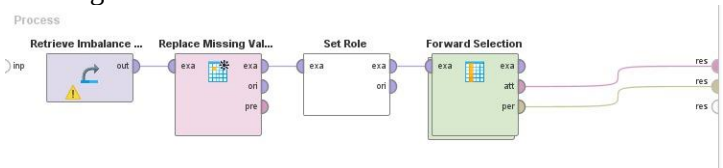

Figure 5. Data Preprocessing

In Figure 5 is the operator used in RapidMiner, where the first step is to enter the data to be modeled, then replace missing values serves to replace missing values in the dataset with other values and in this study the empty value is filled with mean or average. Set

Role serves to assign attribute labels that will be used in modeling. After that, Forward selection is applied to determine attributes that have an influence on the food crop production model. Then to continue the first stage of modeling in RapidMiner then in Forward selection there are more stages and operators, these stages are shown in Figure 6.

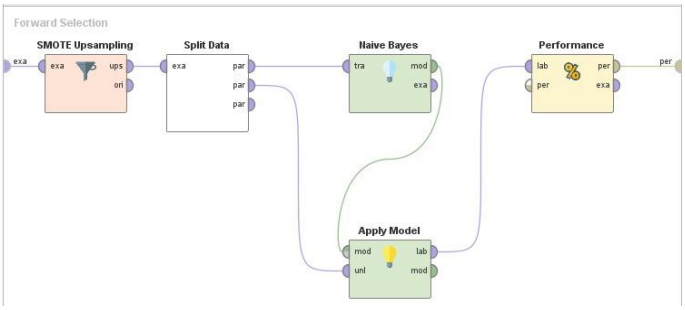

Figure 6. RapidMiner modeling

In Figure 6, the stages of using RapidMiner in the modeling process include using the SMOTE operator upsampling to balance the data, then using the data division or split operator to divide the data into two parts, namely training data and test data with a ratio of 0.8 and 0.2. Compared to the other comparison values in Table 5, the comparison value of 0.8 and 02 is considered to be the model that has the highest level of accuracy.

Table 5.
Accuracy Comparison Value

No	Value Comparison	*Accuracy*
1	0,6 : 0,4	80,59%
2	0,7 : 0,3	82,00%
3	0,8 : 0,2	82,54%
4	0,9 : 0,1	81,38%
5	0,4 : 0,6	79,76%
6	0,3 : 0,7	75,82%
7	0,2 : 0,8	74,70%
8	0,1 : 0,9	71,71%

Then, the Naive Bayes operator is applied for the modeling process, the Apply Model operator is used to apply the trained model to new datasets. Finally, the Performance operator is used to assess the performance of the predictive model.

E. *Evaluation*

In this study, confusion matrix evaluation is used to measure the final results of the algorithm and as the main tool for performance evaluation. In this research, several models have been made in RapidMiner, and Figures 7, 8, 9, and 10 show the comparison of the application of the Naïve Bayes algorithm. Modeling using this algorithm is done several times to find the model that provides the best level of accuracy.

accuracy: 54.23%

	true Naik	true Turun	class precision
pred. Naik	38	26	59.38%
pred. Turun	104	116	52.73%
class recall	26.76%	81.69%	

Figure 7. Naïve bayes modeling 1

accuracy: 54.23%

	true Naik	true Turun	class precision
pred. Naik	38	26	59.38%
pred. Turun	104	116	52.73%
class recall	26.76%	81.69%	

Figure 8. Naïve bayes modeling with SMOTE

accuracy: 64.79%

	true Naik	true Turun	class precision
pred. Naik	71	29	71.00%
pred. Turun	71	113	61.41%
class recall	50.00%	79.58%	

Figure 9. Naïve bayes modeling with Forward selection

accuracy: 64.79%

	true Naik	true Turun	class precision
pred. Naik	71	29	71.00%
pred. Turun	71	113	61.41%
class recall	50.00%	79.58%	

Figure 10. Modeling 4 Naïve bayes, SMOTE with Forward selection

accuracy: 82.54% +/- 2.68% (micro average: 82.54%)

	true Naik	true Turun	class precision
pred. Naik	473	104	81.98%
pred. Turun	94	463	83.12%
class recall	83.42%	81.66%	

Figure 11. Modeling 5 Naive Bayes models with Cross validation and Xgboost

It can be seen in the 5 models in Figures 7 and 8 produce the same accuracy value of 54.23%, and in Figures 9 and 10 also produce the same accuracy of 64.79% and for Figure 11 produces a value of 82.54, where it can be concluded that high accuracy is in the 5th model with the accuracy obtained is quite good with the Good Classification category. Furthermore, to find out the classification category obtained is shown in the form of an ROC curve in Figure 11 below.

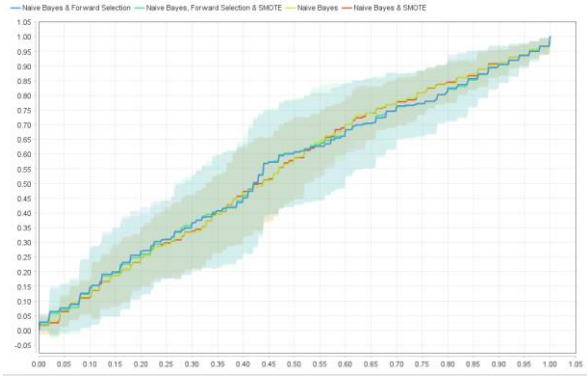

Figure 12. Naïve Bayes Algorithm Comparison Curve

Figure 12 shows the comparison curve of the application of Naïve Bayes algorithm consisting of four models. The blue curve shows the result of applying the Naïve Bayes algorithm with Forward selection, while the green curve shows the result of applying the Naïve Bayes algorithm with Forward selection and SMOTE. The yellow curve line shows the result of applying the Naïve Bayes algorithm with SMOTE, the red curve line shows the Naive Bayes algorithm with SMOTE.

The next evaluation process is displayed through AUC, to determine the type and value of this research result, as seen in Figure 12.

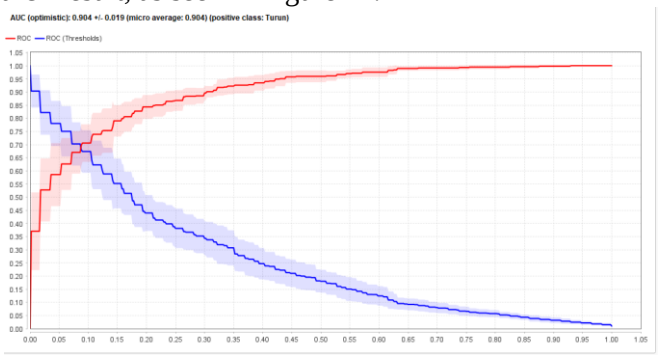

Figure 13. AUC curve

In this graph, the AUC (Area Under Curve) value of 0.904 shows the performance of the model, where the red curve close to the upper left corner indicates that the model performs well on the training data, while the blue curve on the testing data shows the generalization of the model. This result shows that although the model performed well on the training data, there is room for performance improvement on the testing data to get closer to the optimal result. Based on the generalized assessment, with an AUC value of 0.904 and an accuracy of 82.54%, the model falls into the Good Classification category.

F. *Deployment*

In this deployment stage, it becomes an illustration of how this result modeling will be used by the Agriculture Office and farmers in Garut Regency. Based on the results of food crop predictions, the Agriculture Office can provide direction and guidance to farmers to improve the efficiency and effectiveness of their production. If there is a prediction that production yields will decrease in a certain season, the government can immediately take preventive measures by providing assistance or training programs to farmers. This is expected to help farmers to increase their production and ensure food security in Garut Regency is maintained.

IV. CONCLUSION

This research was conducted to predict food crop production in Garut Regency using the Naïve Bayes algorithm and evaluate influential factors. Data was obtained from the Garut Regency Agriculture Office in 2020-2023 with a total of 1344 data. The most influential attributes are crop type, planting addition, harvest area realization, production realization, and production. Feature Forward selection and SMOTE techniques were used to improve the model.

The results showed that the Naive Bayes model with Cross validation and Xgboost resulted in an Accuracy value of 82.54%, Recall value of 81.67%, Precision value of 83.34%. And the AUC value of 0.904% with the Good Classification category.ROC and AUC curves show good model performance on training data, but still need improvement on testing data. This research emphasizes the importance of technology and data analysis in optimizing food crop production in Garut Regency.

REFERENCES

[1] A. Satria, R. M. Badri, and I. Safitri, “Prediksi Hasil Panen Tanaman Pangan Sumatera dengan Metode Machine Learning,” *Digit. Transform. Technol.*, vol. 3, no. 2, pp. 389–398, 2023, doi: 10.47709/digitech.v3i2.2852.

[2] M. K. B. Seran, F. Tedy, I. P. A. N. Samane, P. Batarius, P. A. Nani, and A. A. J. Sinlae, “Analisis Data Pertanian Tanaman Pangan untuk Memprediksi Hasil

Panen di Kabupaten Malaka Menggunakan Metode Multiple Linear Regression," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 4, no. 1, pp. 209–221, 2024, doi: 10.24002/konstelasi.v4i1.8970.

[3] S. Adiguno, Y. Syahra, and M. Yetri, "Prediksi Peningkatan Omset Penjualan Menggunakan Metode Regresi Linier Berganda," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 1, no. 4, p. 275, 2022, doi: 10.53513/jursi.v1i4.5331.

[4] M. Kafil, "Penerapan Metode K-Nearest Neighbors Untuk Prediksi Penjualan Berbasis Web Pada Boutiq Dealove Bondowoso," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 3, no. 2, pp. 59–66, 2019, doi: 10.36040/jati.v3i2.860.

[5] R. C. Wulandari, P. Batarius, and ..., "Prediksi Hasil Pertanian Tanaman Pangan Menggunakan Metode Double Exponential Smoothing," *Proc. ...*, 2023, [Online]. Available: https://conferences.ittelkom-pwt.ac.id/index.php/centive/article/view/255%0Ahttps://conferences.ittelkom-pwt.ac.id/index.php/centive/article/download/255/174

[6] W. Ananda, M. Safii, and M. Fauzan, "Prediksi Jumlah Hasil Panen Sawit Menggunakan Algoritma Naive Bayes," *TIN Terap. Inform. Nusant. Vol*, vol. 1, no. 10, pp. 513–519, 2021.

[7] I. B. K. D. S. Negara, I. P. K. Negara, and N. Y. Arso, "Prediksi Hasil Panen Padi Di Kabupaten Jembrana Menggunakan Metode Naive Bayes Classifier," *J. Teknol. Inf. dan Komput.*, vol. 9, no. 3, pp. 260–265, 2023.

[8] K. C. Pelangi, "Prediksi Produksi Tanaman Pangan Di Provinsi Gorontalo Menggunakan Metode K-NN (K- Nearest Neighbor)," vol. 6, no. 2, pp. 2–6, 2021.

[9] K. Akbar and M. Hayaty, "Data Balancing untuk Mengatasi Imbalance Dataset pada Prediksi Produksi Padi Balancing Data to Overcome Imbalance Dataset on Rice Production Prediction," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 2, no. 02, pp. 1–14, 2020.

[10] A. M. A. Rahim, Inggrid Yanuar Risca Pratiwi, and Muhammad Ainul Fikri, "Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier," *Indones. J. Comput. Sci.*, vol. 12, no. 5, pp. 2995–3011, 2023, doi: 10.33022/ijcs.v12i5.3413.

[11] D. Kurniawan and M. Yasir, "Optimization Sentimen Analysis using CRISP-DM and Naive Bayes Methods Implemented on Social Media," *Cybersp. J. Pendidik. Teknol. Inf.*, vol. 6, no. 2, p. 74, 2022, doi: 10.22373/cj.v6i2.12793.

[12] Y. Suhanda, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," *J. Teknol. Inform. dan Komput.*, vol. 6, no. 2, pp. 12–20, 2020, doi: 10.37012/jtik.v6i2.299.

[13] Y. Christian and K. O. Y. R. Qi, "Penerapan K-Means pada Segmentasi Pasar untuk Riset Pemasaran pada Startup Early Stage dengan Menggunakan CRISP-DM," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 4, p. 966, 2022, doi: 10.30865/jurikom.v9i4.4486.

[14] A. Hardirega and I. Jaelani, "IMPLEMENTASI CONVOLUTIONAL NEURAL NETWORK (CNN) KLASIFIKASI MOTIF BATIK MENGGUNAKAN EFFICIENTNET-B1," vol. 8, no. 5, pp. 10023–10028, 2024.

[15] M. Rafi Muttaqin, T. Iman Hermanto, M. Agus Sunandar, P. Studi Teknik Informatika, and S. Tinggi Teknologi Wastukancana, "Penerapan K-Means Clustering dan Cross-Industry Standard Process For Data Mining (CRISP-DM) untuk Mengelompokan Penjualan Kue," *Journal.Unpak.Ac.Id*, vol. 191. Rafi, no. 1, pp. 38–53, 2022, [Online]. Available: http://journal.unpak.ac.id/index.php/komputasi/article/view/3976

[16] N. C. Sastya and I. Nugraha, "Penerapan Metode CRISP-DM dalam Menganalisis Data untuk Menentukan Customer Behavior di MeatSolution," *Unistek*, vol. 10, no. 2, pp. 103–115, 2023, doi: 10.33592/unistek.v10i2.3079.

[17] T. T. Widowati and M. Sadikin, "Analisis Sentimen Twitter terhadap Tokoh Publik dengan Algoritma Naive Bayes dan Support Vector Machine," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 11, no. 2, pp. 626–636, 2021, doi: 10.24176/simet.v11i2.4568.
[18] S. Shedriko and M. Firdaus, "Penentuan Klasifikasi Dengan Crisp-Dm Dalam Memprediksi Kelulusan Mahasiswa Pada Suatu Mata Kuliah," *Semnas Ristek (Seminar Nas. Ris. dan Inov. Teknol.*, vol. 6, no. 1, pp. 826–831, 2022, doi: 10.30998/semnasristek.v6i1.5814.
[19] Juanda, "Jurnal Mantik Penerapan Naive Bayes Dalam Memprediksi Penjualan Tuan Kentang Palembang," vol. 6, no. 36, pp. 2502–2507, 2022.
[20] S. Lestari, A. Akmaludin, and M. Badrul, "Implementasi Klasifikasi Naive Bayes Untuk Prediksi Kelayakan Pemberian Pinjaman Pada Koperasi Anugerah Bintang Cemerlang," *PROSISKO J. Pengemb. Ris. dan Obs. Sist. Komput.*, vol. 7, no. 1, pp. 8–16, 2020, doi: 10.30656/prosisko.v7i1.2129.