JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 7, No. 2, December 2024, pp. 224~231
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: 10.36378/jtos.v7i2.3975

Application of the C4.5 Algorithm for Predicting Banana Chips Production Demand (Case Study at UD. Sinar Sejahtera Medan)

M. Teguh Wijaya[1], Rakhmat Kurniawan R[2]
[1,2]Computer Science, State Islamic University of North Sumatra

Article Info

Article history:

Received 12 05, 2024
Revised 12 17, 2024
Accepted 12 30, 2024

Keywords:

Prediction
Banana
C.45 Algorithm
Production Demand

ABSTRACT

The rapid advancement of science and technology significantly impacts various aspects of life, including business operations. Technology plays a vital role in providing information and simplifying human tasks, addressing challenges faced by growing companies, particularly in managing sales fluctuations. Factors such as market competition, product quality, and consumer interest are critical for evaluating and improving sales strategies. UD. Sinar Sejahtera Medan, a food processing industry specializing in banana chips, faces challenges such as fluctuating raw material supply, impacting production and sales. To address this, a prediction system for raw material demand was developed, leveraging the C4.5 algorithm. The C4.5 algorithm was selected for its ability to generate decision trees from historical data, providing interpretable results and high accuracy in forecasting categorical outcomes. By analyzing past trends in raw material availability and usage, the algorithm predicts future supply needs, optimizing production planning and supporting sustainable business operations. This study's findings are expected to align with previous research, offering insights for better production and sales management.

Corresponding Author:

M. Teguh Wijaya
Computer Science
State Islamic University of North Sumatra
North Sumatera, Indonesia
Email: tw03358@gmail. com

1. Introduction

The rapid development of science and technology continues to evolve in tandem with the progress of the times. Technology has become increasingly sophisticated and plays a vital role in every aspect of life. It provides information needed and simplifies human tasks, addressing various problems, especially for growing businesses. Sales performance is a crucial factor in a company's sustainability, as it reflects whether the company is advancing or regressing. However, not all businesses operate smoothly and stably, often encountering issues such as declining or increasing sales.[1]

Several factors contribute to these issues, such as competition among companies, product marketing, and product quality that affects competitiveness. Consumer interest needs to be evaluated as a reference for sales strategies[2].

UD. Sinar Sejahtera Medan is a food processing industry located at Jl. Setia Luhur, Medan, North Sumatra. Led by Mr. Iswanto, S. Pdi, banana chips are one of the flagship products of UD. Sinar Sejahtera Medan. Banana chips are a type of snack long recognized in various cultures worldwide, especially in

tropical regions where bananas are abundant. The production of banana chips significantly contributes to the economies of regions where bananas thrive. The banana chip industry often serves as a source of income for local banana farmers and craftsmen. Despite increasing demand, this industry faces challenges such as fluctuations in raw material prices (bananas), intense market competition, and production process issues that affect product quality.[3]

Entrepreneurs relying on bananas as their primary ingredient face challenges regarding supply consistency. Sometimes, the supply falls short of the required amount, while at other times, it exceeds the demand. Hence, having a prediction system to estimate the raw material supply requirements would benefit both buyers and sellers.[4]

A prediction system is a technology that utilizes historical data and algorithms to project future events or outcomes. This technology is widely applied across various sectors, including finance, healthcare, e-commerce, and weather forecasting. As computing technology advances and data availability grows, prediction systems have become increasingly sophisticated and accurate. These predictions are based on patterns in historical data, assuming these patterns will persist in the future [5]

The C4.5 algorithm is one of the algorithms used for classification, segmentation, or grouping and is predictive in nature. It is an improvement of the ID3 algorithm. Quinlan introduced the C4.5 method, which uses the Gain Ratio (GR) method for attribute splitting, replacing Information Gain.[6]

A previous study by Aghnie Kurnia et al. (2024) applied the C4.5 algorithm to predict corn production outcomes. It achieved a total accuracy of 92.82% from a dataset of 4,121 entries, divided into 80% training data and 20% testing data. The current study focuses on a different object of research, making it distinct from previous studies. It is expected that the findings of this research will align with earlier studies, providing insights for suppliers to create more profitable production plans for the next year's sales period.

Given these challenges, this research aims to develop a prediction system to efficiently manage banana raw material demands using the C4.5 algorithm, focusing on a case study at UD. Sinar Sejahtera.[7]

2. Research Method

A. Modeling

This is the initial stage of the research process, focusing on the general issues, particularly regarding the imbalance in the production and demand of banana chips. The main problem lies in the mismatch between the quantity of banana chips produced and the demand from consumers. This issue needs to be addressed and resolved promptly. In this stage, the focus is on identifying the existing problems to better understand the issues to be researched. In this case, the research will concentrate on predicting the production quantity of banana chips demand using the C4.5 method.[8]

B. Data collection

This technique is also an important stage in the research process, as it provides the necessary data for analysis. This research uses secondary data. Below are the two methods used to collect data:

1. Literature Review

A thorough literature review will be conducted to gather secondary data from existing studies, journals, articles, books, and other credible sources. This will help in understanding the concepts, methodologies, and findings of previous research related to the C4.5 algorithm, banana chip production, and demand forecasting. The literature review will also provide insights into the methodologies used by other researchers in similar fields and help identify gaps in existing knowledge.

2. Document Review

Secondary data will also be collected by reviewing documents and records provided by UD. Sinar Sejahtera Medan, such as production logs, sales reports, historical demand data, and raw material supply records. These documents contain valuable information on the production and demand patterns of banana chips, as well as any external factors that may influence them. The analysis of these documents will support the development of the predictive model based on the C4.5 algorithm.

These two methods will ensure that the collected data is comprehensive, reliable, and suitable for the research's goals.

C. Data Analysis

In this research, data analysis involves organizing data after the observation process to examine and directly observe the research object, as well as collecting data from UD. SINAR SEJAHTERA MEDAN. The analyzed data includes historical sales records and the production volume categorized as high or low.

The analysis method used is the C4.5 Algorithm, which relies on sample data from 2023–2024. To construct a decision tree, raw data must first be available, which can be derived from the sales and production data of UD. SINAR SEJAHTERA MEDAN for the years 2023–2024. Based on the production data,

transformation has been carried out by classifying specific attributes. After the transformation, it is explained that there are two desired targets: high and low production volumes.

The purpose of this data analysis is to create a decision tree and determine the predicted production volume of banana chip demand using several variables that will serve as considerations for this prediction, as follows:

Tabel 1. Variable Name

No.	Variable Name	Attribute
1.	Raw Material Quantity	High, Low
2.	Human Resources	High, Low
3.	Demand Quantity	High, Low
4.	Production Quantity	High, Low

The C4.5 algorithm, in its process, provides predictions by generating a decision tree through calculating the relationships between variables. The variables involved include the quantity of incoming bananas, human resources (HR), the number of chips produced, and orders. The target values are classified as *profit* and *loss*. The variable value criteria are as follows:

3. Result and Discussion

A. The application of the method

The C4.5 algorithm, in its process of making predictions, generates a Decision Tree by calculating the relationships between the variables. The variables in question include the amount of bananas received, human resources (SDM), chips produced, and orders. The target values are profit and loss. The conditions for the values of these variables are as follows

Tabel 2. Conditions and Categories

Variable	Conditions and Categories	
Raw Material Quantity (Kg)	Low	15 – 42
	Medium	43 - 70
	High	71 – 98
Human Resources	Low	8 – 10
	Medium	11 – 13
	High	14 – 15
Demand Quantity (Kg)	Low	5 – 18
	Medium	19 – 32
	High	33 - 45
Production Quantity (Kg)	Low	7 – 20
	Medium	21 – 34
	High	35 -47

B. Data Processing

One year of banana chips production and sales data was utilized, covering the period from January 2023 to December 2023. In this study, the data was normalized by assigning numeric values: "Low" was set to 1, "Medium" to 2, and "High" to 3. This normalization was done to simplify calculations when using the program. The data was then split into 80% training data and 20% testing data, resulting in a total of 240 training data points.

The C4.5 Algorithm consists of several steps in its prediction process. The very first step is calculating the entropy value.To calculate the entropy value, the following formula can be used:

$$Entropi(S) = -\sum_{i=1}^{k} pi * log_2(pi) \quad (1)$$

Where:

- SSS: The dataset being evaluated.
- pip_ipi: The proportion of data belonging to the iii-th category or class.

- nnn: The total number of categories or classes.

This formula measures the level of impurity or randomness in the dataset, helping to determine the best attribute for splitting in decision tree algorithms.

After calculating the entropy value of all the data, the next step is to calculate the gain value to measure how much uncertainty (entropy) is reduced after splitting the data based on a certain attribute. The formula for calculating the gain value is:

$$Gain(S, A) = Entropi(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} * Entropi\ (Sv) \quad (2)$$

Where:

- Sv = the subset of data divided based on the value of attribute A.

C. Implementation of the Method

1. Calculating the total entropy value is done as follows:

$Entropi(S) = -(0{,}02917 * log_2(0{,}02917) \pm (0{,}97083 * log_2(0{,}97083)$

$Entropi(S) = 0.1904$

The initial entropy value obtained is 0.1904. The next step is to calculate the entropy value for each subset of the variable. There are 4 variables, each with 3 subsets.

2. Calculating the entropy value for Pisang Masuk and others

"Pisang Masuk Sedikit" has a total of 136 data points, with 4 resulting in profit and 132 resulting in loss.

$Entropi(psm) = -(0{,}02941 * log_2(0{,}02941) \pm (0{,}97058 * log_2(0{,}97058)$

$Entropi(psm) = 0{,}19143$

"Pisang Masuk Sedikit" has a total of 90 data points, with 3 resulting in profit and 87 resulting in loss.

$Entropi(psm) = -(0{,}03333 * log_2(0{,}03333) \pm (0{,}96666 * log_2(0{,}96666)$

$Entropi(pms) = 0{,}21084$

"Pisang Masuk Banyak" has a total of 14 data points, but with 0 profit, the entropy is not calculated.

After obtaining all entropy values for each variable using the same calculation method as above, the gain for each variable will be calculated. The results can be seen as follows:

3. Calculating the gain value of incoming bananas and others.

$Gain(pisang\ masuk) = 0{,}190195 - \left(\frac{136}{240} * 0{,}190195 + \frac{90}{240} * 0{,}21084\right)$

$Gain(pisang\ masuk) = 0{,}002650$

$Gain(SDM) = 0{,}190195 - \left(\frac{100}{240} * 0{,}194391 + \frac{115}{240} * 0{,}21783\right)$

$Gain(SDM) = 0{,}0048187$

$Gain(keripik\ yang\ di\ hasilkan) = 0{,}190195 - \left(\frac{182}{240} * 0{,}181553 + \frac{49}{240} * 0{,}246022\right)$

$Gain(keripik\ yang\ di\ hasilkan) = 0{,}0022874$

$Gain(pesanan) = 0{,}190195 - \left(\frac{132}{240} * 0{,}266764 + \frac{92}{240} * 0{,}086504\right)$

$Gain(Pesanan) = 0{,}010314$

After obtaining the overall gain values, the next step is to create the next node in the decision tree. This node is determined by the highest gain value from the previous node. In the first node, the highest gain value was found in "orders" with a value of 0.010314, thus "orders" becomes the next node with the subsets "low" and "medium." The same process will be repeated until the final subset and variable are reached.

The following is the decision tree obtained from this research related to banana production prediction.

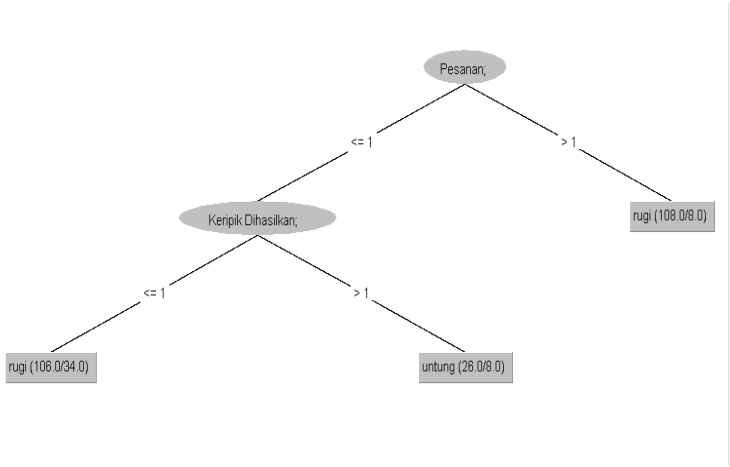


Figure 1. Decision Tree

The following is the decision tree obtained, where it can be seen that if the orders are too many, the current production will result in a loss. If the orders are few and the chips produced are plentiful, it will generate a profit. Conversely, if the chips produced are few, it will result in a loss

D. Implementation on Jupyter Notebook

```
file_path = 'uji coba.xlsx'
df = pd.read_excel(file_path)
print("Data Awal:")
print(df.head())
```

```
Data Awal:
   Pisang Masuk  SDM  Keripik Dihasilkan  Pesanan Keterangan 2
0            56   14                  25       27         rugi
1            27   11                  10       18         rugi
2            33   12                  15       10       untung
3            43   11                  17       14       untung
4            43   10                  17       25         rugi
```

Figure 2. Reading an Excel file and displaying the data.

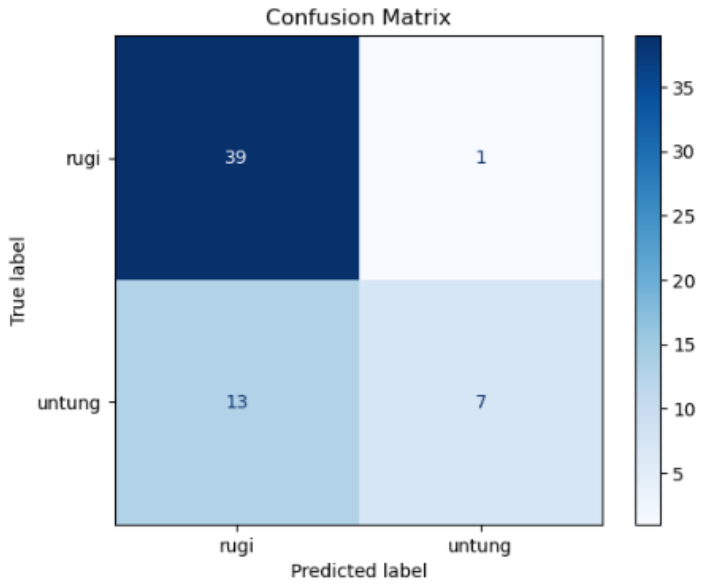
The data is read using the Pandas library and the training model utilizes Scikit-learn along with the C4.5 algorithm. Visualization is done using the Matplotlib.pyplot library.

```
print("\nPerhitungan Entropy dan Information Gain:")
for column in X.columns:
    gain = calculate_information_gain(df, column, 'Keterangan 2')
    print(f"Information Gain untuk atribut '{column}': {gain:.4f}")
```

```
Perhitungan Entropy dan Information Gain:
Information Gain untuk atribut 'Pisang Masuk': 0.2942
Information Gain untuk atribut 'SDM': 0.0886
Information Gain untuk atribut 'Keripik Dihasilkan': 0.2046
Information Gain untuk atribut 'Pesanan': 0.2093
```

Figure 3. Gain calculation results.

The following is a function to calculate the gain value in the data. Later, the data will be split into 80% for training data and 20% for testing data.




```
Confusion Matrix (angka):
[[39  1]
 [13  7]]

Classification Report:
              precision    recall  f1-score   support

       rugi       0.75      0.97      0.85        40
      untung       0.88      0.35      0.50        20

    accuracy                           0.77        60
   macro avg       0.81      0.66      0.67        60
weighted avg       0.79      0.77      0.73        60


Penjelasan:
- Total data: 60
- Prediksi benar: 46
- Prediksi salah: 14
- Akurasi: 76.67%
  - rugi:
    - Benar: 39 (diagonal confusion matrix)
    - Salah: 1 (keliru untuk kelas ini)
  - untung:
    - Benar: 7 (diagonal confusion matrix)
    - Salah: 13 (keliru untuk kelas ini)
```

Figure 4. confussion matrix.

After obtaining the results, error evaluation is performed using a confusion matrix on the test data, which consists of 60 samples, achieving an accuracy of 76.6%.

4. Conclusion

Based on the final results of the research process for predicting the production quantity of banana chips demand, several conclusions can be drawn, as follows:

- The prediction of banana chip demand using one year of sales data with the C4.5 algorithm produced a decision tree with the following details:
- If orders are low and the resulting chips are few, there will be a loss.
- If orders are low and the resulting chips are more than few, there will be a profit.
- If orders are more than few, there will be a loss.
- The accuracy achieved is not very high, with a value of 76% on 40 test data points out of a total of 300 data points.

Acknowledgement

I would like to extend my heartfelt gratitude to all parties who have provided invaluable support, guidance, and contributions throughout the research process. The successful completion of this research would not have been possible without the significant roles of those who sincerely assisted, guided, and motivated me at every stage of its development.

References

[1] R. T. Aldisa and P. Maulana, "Analisis sentimen opini masyarakat terhadap vaksinasi booster Covid-19 dengan perbandingan metode Naive Bayes, Decision Tree dan SVM," Building of Informatics, Technology and Science (BITS), vol. 4, no. 1, pp. 106–109, 2022. doi: 10.47065/bits.v4i1.1581.
[2] R. Amaliyyah, "Title," February, p. 6, 2021.
[3] A. Armansyah and R. K. Ramli, "Model prediksi kelulusan mahasiswa tepat waktu dengan metode Naïve Bayes," Edumatic: Jurnal Pendidikan Informatika, vol. 6, no. 1, pp. 1–10, 2022. doi: 10.29408/edumatic.v6i1.4789.
[4] N. M. A. J. Astari, D. G. H. Divayana, and G. Indrawan, "Analisis sentimen dokumen Twitter mengenai dampak virus Corona menggunakan metode Naive Bayes Classifier," Jurnal Sistem dan Informatika (JSI), vol. 15, no. 1, pp. 27–29, 2020. doi: 10.30864/jsi.v15i1.332.
[5] M. R. Fahlevvi, "Analisis sentimen terhadap ulasan aplikasi pejabat pengelola informasi dan dokumentasi kementerian dalam negeri republik Indonesia di Google Playstore menggunakan metode Support Vector Machine," Jurnal Teknologi dan Komunikasi Pemerintahan, vol. 4, no. 1, pp. 1–13, 2022. doi: 10.33701/jtkp.v4i1.2701.
[6] F. Wulandari, E. Haerani, M. Fikry, and E. Budianita, "Analisis sentimen larangan penggunaan obat sirup menggunakan algoritma Naive Bayes Classifier," Jurnal Coscitech (Computer Science and Information Technology), vol. 4, no. 1, pp. 88–96, 2023. doi: 10.37859/coscitech.v4i1.4781.
[7] M. Furqan, S. Sriani, and S. M. Sari, "Analisis sentimen menggunakan K-Nearest Neighbor terhadap new normal masa Covid-19 di Indonesia," Techno.Com, vol. 21, no. 1, pp. 51–60, 2022. doi: 10.33633/tc.v21i1.5446.
[8] D. G. Nugroho, Y. H. Chrisnanto, A. Wahana, and F. Matematika, "Analisis sentimen pada jasa ojek online menggunakan metode Naïve Bayes," unpublished.
[9] H. Irsyad, A. Farisi, and M. R. Pribadi, "Klasifikasi opini masyarakat terhadap jasa ISP MyRepublic dengan Naïve Bayes," JNTETI, vol. 8, no. 1, 2019. Available: https://t.co/q3btia6mof.
[10] M. Luqman, "Keamanan perangkat lunak pada bahasa pemrograman Node.js untuk aplikasi berbasis web," Institut Teknologi Bandung, 2016.
[11] S. Mandasari, B. H. Hayadi, and R. Gunawan, "Nomor 2," Volume, vol. 5, pp. 118–126, 2022. Available: https://ojs.trigunadharma.ac.id/index.php/jsk/index.
[12] I. Nabillah and I. Ranggadara, "Mean absolute percentage error untuk evaluasi hasil prediksi komoditas laut," JOINS (Journal of Information System), vol. 5, no. 2, pp. 250–255, 2020. doi: 10.33633/joins.v5i2.3900.
[13] D. Nugraha and D. Gustian, "Analisis sentimen penggunaan aplikasi transportasi online pada ulasan Google Play Store dengan metode Naive Bayes Classifier," unpublished.
[14] S. Nurul, J. Fitriyyah, N. Safriadi, E. Esyudha, and P. #3, "Analisis sentimen calon presiden Indonesia 2019 dari media sosial Twitter menggunakan metode Naive Bayes," JEPIN (Jurnal Edukasi dan Penelitian Informatika), vol. 5, no. 3, pp. 279–285, 2019.
[15] L. Oktasari, Y. H. Chrisnanto, R. Program, S. Informatika, F. Matematika, P. Pengetahuan, and J. A. Yani, "Text mining dalam analisis sentimen asuransi menggunakan metode Naïve Bayes Classifier," unpublished.
[16] M. I. Petiwi, A. Triayudi, and I. D. Sholihati, "Analisis sentimen GoFood berdasarkan Twitter menggunakan metode Naïve Bayes dan Support Vector Machine," Jurnal Media Informatika Budidarma, vol. 6, no. 1, p. 542, 2022. doi: 10.30865/mib.v6i1.3530.
[17] G. Pringgo Digdo, Modul Lengkap JavaScript, 2015.

[18] M. Radhi, D. R. H. Sitompul, S. H. Sinurat, and E. Indra, "Analisis big data dengan metode exploratory data analysis (EDA) dan metode visualisasi menggunakan Jupyter Notebook," Jurnal Sistem Informasi dan Ilmu Komputer Prima, vol. 4, no. 2, 2021.
[19] G. T. Santoso, "Analisis sentimen pada tweet dengan tagar #BPJSrasarentenir menggunakan metode Support Vector Machine (SVM)," pp. 12–13, 2021.
[20] F. V. Sari and A. Wibowo, "Analisis sentimen pelanggan toko online JD.ID menggunakan metode Naïve Bayes Classifier berbasis konversi ikon emosi," Jurnal Simetris, vol. 10, no. 2, 2019.
[21] G. I. E. Soen, Marlina, and Renny, "Implementasi cloud computing dengan Google Colaboratory pada aplikasi pengolah data Zoom Participants," Journal Informatic Technology and Communication, vol. 6, no. 1, pp. 24–30, 2022.
[22] A. Syakuro, "Analisis sentimen masyarakat terhadap e-commerce pada media sosial menggunakan metode Naive Bayes Classifier (NBC) dengan seleksi fitur Information Gain (IG)," unpublished.
[23] M. Syarifuddin, "Analisis sentimen opini publik mengenai Covid-19 pada Twitter menggunakan metode Naïve Bayes dan KNN," INTI Nusa Mandiri, vol. 15, no. 1, pp. 23–28, 2020. doi: 10.33480/inti.v15i1.1347.
[24] W. Yulita, E. D. Nugroho, M. H. Algifari, P. Studi Teknik Informatika, I. Teknologi Sumatera, J. Terusan Ryacudu, W. Huwi, J. Agung, and L. Selatan, "Analisis sentimen terhadap opini masyarakat tentang vaksin Covid-19 menggunakan algoritma Naïve Bayes Classifier," JDMSI, vol. 2, no. 2, pp. 1–9, 2021.

[25] P. Yuniar and Kismiantini, "Analisis sentimen ulasan pada Gojek menggunakan metode Naive Bayes," Statistika, vol. 23, no. 2, pp. 164–175, 2023. doi: 10.29313/statistika.v23i2.2353.
[26] I. Zufria and N. Fadhillah, "Prediksi penjualan ikan dengan metode fuzzy time series," Journal of Science and Social Research, no. 3, 2024. Available: