*Journal of Intelligent Decision Support System (IDSS) 4 (4) (2021) 186-194*

Published by: IOCSCIENCE

Journal of Intelligent Decision Support System (IDSS)

Journal homepage: www. idss.iocspublisher.org

# Implementation Of Data Mining With C4.5 Algorithm For Determining The Home Industry Product Marketing Strategy

Teresia Herniamwati Zebua, Fristi Riandari,

*Informatics Engineering, STMIK Pelita Nusantara, Sumatera Utara, Indonesia*

**Article Info**

***Article history:***

Received: 15/09/2021
Revised: 02/10/2021
Accepted: 11/11/2021
Available online 01/12/2021

***Keywords:***

Anxiety Disorder,
Naïve Bayes,
Expert System,
Web.

**ABSTRACT**

Home Industry is one of the SMEs that produce home-made products, such as pastries. Not all of these products are sold by consumers. This study uses a web-based C4.5 algorithm to determine marketing strategies for Home Industry products that are not selling well by classifying the indicators that most influence consumers in buying Home Industry products so that they can provide information about marketing strategies that will be carried out. Based on the results of the study, the highest gain value was the packaging attribute with a value of 1.86, so that the packaging attribute was used as the root in the formation of a decision tree. Then the second highest gain is the price attribute with a value of 1.26, and the third highest gain is the taste attribute with a value of 1.03, then the fourth highest gain is the service attribute with a value of 0.89.



***Corresponding Author:***

Fristi Riandari,
Informatics Engineering,
STMIK Pelita Nusantara Medan,
Jl Iskandar Muda No. 1 Medan, 20154, Indonesia.
Email: fristy.rianda@ymail.com

## 1. Introduction

The COVID-19 pandemic that entered Indonesia since January 2020 (Kompas, May 11, 2020) has made people's life patterns increasingly change. This includes the need for technology. It is inevitable that nowadays almost all activities in various fields are influenced by technology. One of them is in the field of Small and Medium Enterprises (SMEs) which today are in dire need of technology to encourage the running of the business. Where this type of SMEs is considered the most affected by the pandemic.

Products from SMEs are usually not all sold by consumers, so Home Industries need to make a marketing strategy for the results of pastries that are not selling well which are later expected to be able to classify the indicators that most influence consumers in buying products.

This research was conducted for the classification of product data that is not selling well. The approach taken for this classification is the C4.5 algorithm. The C4.5 algorithm is an algorithm used to form a decision tree. Decision trees are a very powerful and well-known method of classification and prediction. The decision tree method converts very large facts into a decision tree that represents the rule. Rules can be easily understood in natural language. And they can also be expressed in the form of database languages such as Structured Query Language to search for records in certain categories [16].

Based on the analysis conducted by Tri Bagus Tusarwenda (2018), in his research, he explains the basic problems raised in his research, namely that the transaction data in the database of sales of

goods is very large, causing the amount of data to continue to increase every day. From the accumulation of data that occurs, it can be extracted to find patterns of sales of goods that can be used to analyze the market and forecast sales in the future. The study concluded that the C4.5 algorithm has an accuracy value of training data reaching 90.59% and testing data reaching 88.00%. in classifying. This research was conducted in order to help sellers to predict the sale of their merchandise, so that they can prepare or stock goods that are predicted to increase in sales [16].

In addition, the research conducted by Choirul Anam and Harry Budi Santoso (2018) has a background problem where student scholarship applicant data that is recapitulated in a worksheet (MS Exel) still takes a lot of time and is at risk of being inconsistent due to the element of subjectivity. By using a data mining classification algorithm on existing data, it will be known the pattern of mapping the characteristics of the applicant to the decisions taken so that the selection process will be easier, more consistent and can avoid the element of subjectivity from decision makers. This study concludes that the C4.5 algorithm has a better performance than Naive Bayes in classifying [1].

Furthermore, research conducted by Erlin Elisa (2017) has a background that is the number of work accidents that occur and have an impact on the performance and work carried out by construction employees. By using the C4.5 algorithm, it is expected to collect further data which will be made a decision tree which will then generate problem solution rules. This study concludes that the factors that cause construction work accidents that often occur are the Workplace Environment, Safety Signs and Workers and Work Methods [3].

Then, in a study conducted by Luluk Elvitaria and Muhammad Havenda (2017), they raised a problem based on examining the level of interest in determining student interest in extracurricular activities using the C4.5 algorithm. In this study, the school can find out to what extent the level of interest in foreign languages in students and schools can increase extracurricular activities and students can develop their interest in foreign languages according to their wishes [6].

And finally, the research conducted by Muhammad Fauzul Arifin and Devi Fitrianah has a background problem that is not using the correct procedure in accepting sales partners so that bad credit, fictitious orders and fraud often occur. To overcome this problem, research is carried out by applying the C4.5 algorithm to determine whether the prospective sales partner is accepted or rejected. The results of this study showed that the classification results were validated by ten-fold cross validation with an accuracy rate of 96.26%, precision 100% and recall 71.43% [2].

The purpose of the research is to produce information that makes it easier for Home Industry owners to determine marketing strategies for products that are not selling well

## 2. Method

### 2.1 Research Framework

In this chapter the author will explain the research framework in Figure 2 which is made systematically. So that it can be a guide in solving problems that will be faced.

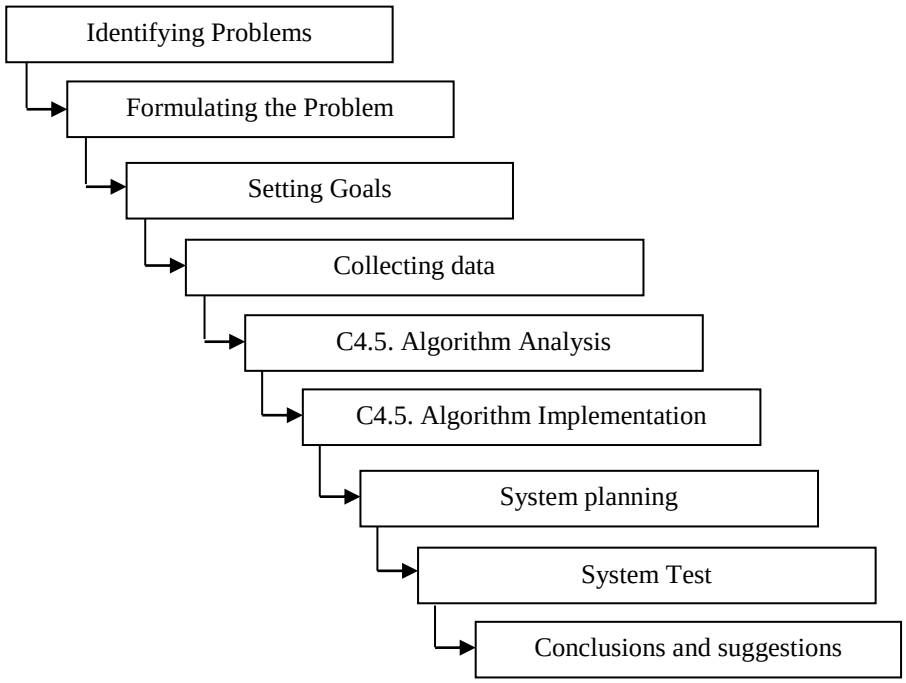


**Figure 1.** Research Framework

a. Identification of problems
This is done by researchers to make it easier to find the most appropriate solution to the problems to be discussed.
b. Formulation of the problem
This stage is carried out in order to understand the problems that have been identified.
c. Goal Setting
After the problem can be understood, then the next researcher will determine what the goals are carried out in the problem solving process.
d. Collecting Data and Information
This stage is carried out to obtain information about sales data for the last 3 months, namely November 2020, December 2020, and January 2021, as well as attribute data on products, such as packaging, price, taste and service criteria that will be assessed in determining marketing strategies for products. Home Industry. The collection of data and information is done by distributing questionnaires to consumers with a collection of questions to be answered regarding the reasons consumers decide to buy Home Industry products to get a more detailed description and explanation, so that data about Home Industry products that sell and do not sell well to determine the marketing strategy for Home Industry products that are not selling well.
e. C4.5. Algorithm Analysis
After the data is collected based on the problems that have been formulated, the next process is to analyze and process the data with the C4.5 algorithm in accordance with the KDD stages, namely:
Selection
Preprocessing/Cleaning
Transformation
Data Mining
Interpretation/Evaluation [13].
Data processing using the C4.5 algorithm is carried out by calculating the Gain value then using the formula :

$$Gain\ (\mathrm{S},\mathrm{A}) = Enthropy(\mathrm{S}) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Enthropy\ (Si) \quad (1)$$

Where:
S = Space/Sample Data used for training data.
A = Attribute
Gain (S,A) = Information Gain on attribute A
Entropy = Entropy on attribute A
Then after the gain value is obtained, the next step is to calculate the Entropy value, using the formula:

$$Enthropy(S) = -\sum_{i=1}^{i} \frac{S_i}{S} Log_2 \frac{S_i}{S} \quad (2)$$

Where:
S = Space (data) Sample used for training.
A = Attribute.
Si = Number of Samples for attribute i.

## 3. Results and Discussion

Research conducted by the author in one of the Home Industries, namely Mama Kita Cake Business, has collected data for processing as needed which aims to predict consumer reasons in deciding to buy Home Industry products in order to determine marketing strategies for products that are not selling well by applying data mining using C4.5 algorithm. The sales data obtained from Mama Kita's Cake Business is used as a guide in compiling a questionnaire that will be filled out by a number of customers. The table can be seen in table 1 below.Data Analysis

**TABLE 1.**
COOKIE SALES DATA FOR NOVEMBER 2020 TO JANUARY 2021

| Num | Types of Pastries | Weight (jar/box/pcs) | Price (Rp) | Sold | | |
|---|---|---|---|---|---|---|
| | | | | Nov | Dec | Jan |
| 1 | Nastar | 500 gr/jar | 45.000 | 5 | 45 | 18 |
| 2 | Peanut Cake | 500 gr/jar | 30.000 | 6 | 15 | 12 |
| 3 | Grilled Sagun | 500 gr/jar | 25.000 | 12 | 60 | 15 |
| 4 | Squirt Cake | 500 gr/jar | 25.000 | 15 | 60 | 25 |
| 5 | Snow Princess | 500 gr/jar | 45.000 | 15 | 25 | 28 |
| 6 | Onion Cake | Pack | 20.000 | 44 | 40 | 37 |
| 7 | Pan Flower | Pack | 15.000 | 8 | 42 | 40 |
| 8 | Root Cake | Pack | 15.000 | 5 | 45 | 42 |
| 9 | Tojen Beans | Pack | 30.000 | 8 | 22 | 20 |
| 10 | Donuts | Pcs | 1.000 | 3900 | 4800 | 4050 |

The data from the research that will be processed the assessment indicators are obtained from the results of a questionnaire that has been filled out by a number of customers at the Mama Kita Home Cake Industry. The number of questionnaire data obtained was 137 questionnaires. The output in this study is divided into two categories, namely SELLING and NOT SELLING. Based on the output to be generated, the writer uses a classification technique because there are categorical variables that will produce a number of information, such as data selection, data cleaning, data transformation, data mining and Graphical User Interface (GUI).

**TABLE 2.**
HOME INDUSTRY OVERALL PRODUCT QUESTIONNAIRE

| Code | Nastar | | | | Conclusion |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 001 | 3 | 3 | 4 | 4 | Buy |
| 002 | 3 | 3 | 3 | 3 | Buy |
| 003 | 3 | 3 | 3 | 3 | Buy |
| 004 | 3 | 3 | 2 | 3 | Buy |
| .... | .... | .... | .... | .... | .... |
| 137 | 2 | 4 | 4 | 4 | Buy |

### 3.1 Processing Questionnaire Data into the KDD Process

Based on the raw data available in the sales data of pastries for the period November 2020 to January 2021, it was found that products that sold more and those that sold less. Where the product that is in great demand is Donuts and the product that is not selling well is Peanut Cake. The data to be determined by the rule is the type of product that sells a lot and is bought by customers, where the type of product that sells a lot is Donuts. Therefore, donuts are used as a product whose rules will be determined so that it can be drawn to determine marketing strategies for products that are not selling well.

The design of the assessment on the list of questions in the questionnaire distributed to customers of Mama Kita's Home Industries Business products was compiled by the author as shown in table 3 below.

**TABLE 3.**
DONUT QUESTIONNAIRE ASSESSMENT DESIGN

| Rating Indicator | | Strongly agree | Agree | Do not agree | Strongly Disagree |
|---|---|---|---|---|---|
| 1 | Packaging | 4 | 3 | 2 | 1 |
| 2 | Price | 4 | 3 | 2 | 1 |
| 3 | Taste | 4 | 3 | 2 | 1 |
| 4 | Service | 4 | 3 | 2 | 1 |

### 3.2 Data Selection

At this stage the data from the Donut questionnaire that has been filled out by a number of customers is combined into one information for processing to the next stage. Of the 137 questionnaires distributed, 45 questionnaires were obtained that customers could fill out for Donut products and can be seen in table 4 below.

**TABLE 4.**
DONUTS QUESTIONNAIRE RESULTS DATA

| Code | Donuts | | | | Conclusion |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 001 | 3 | 3 | 4 | 2 | Buy |
| 002 | 3 | 3 | 4 | 4 | Buy |
| 003 | 3 | 3 | 4 | 1 | Buy |
| 004 | 3 | 3 | 4 | 4 | Buy |
| .... | .... | .... | .... | .... | .... |
| 045 | 2 | 4 | 4 | 4 | Buy |

### 3.3 Pre-processing

After getting the questionnaire data obtained from the data selection, it then enters the preprocessing stage, where at this stage the data from the selection will be processed to the data cleaning stage to remove inconsistent data and noise. At this stage the data obtained from the cleaning process totaled 27 questionnaires, from the previous data there were 45 questionnaires.

**TABLE 5.**
PRE-PROCESSING RESULT QUESTIONNAIRE DATA

| Code | Donuts | | | | Conclusion |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 001 | 3 | 3 | 4 | 2 | Buy |
| 002 | 3 | 3 | 4 | 4 | Buy |
| 003 | 3 | 3 | 4 | 1 | Buy |

| Code | Donuts | | | | Conclusion |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 004 | 3 | 3 | 1 | 4 | Don.t buy |
| .... | .... | .... | .... | .... | .... |
| 027 | 2 | 4 | 4 | 4 | Buy |

### 3.4 Transformation

Furthermore, after the data has been cleaned, it enters the transformation stage, where the data will be changed or combined into the assessment according to the predetermined value of the assessment indicator. Based on the results of the research conducted, there are 4 (four) assessment indicators used for the classification of Home Industry products, namely:

a. Packaging

The criteria on this package are divided into 4 (four) assessments, as can be seen in the following table.

**TABLE 6.**
PACKAGING ASSESSMENT CRITERIA

| Num | Packaging | |
|---|---|---|
| 1 | 4 | Tall |
| 2 | 3 | Enough |
| 3 | 2 | Low |
| 4 | 1 | Very low |

For the attributes of price, taste and service have the same assessment criteria.

After the questionnaire data is cleaned, then the data is then transformed from each criterion which can be seen in the following table.

TABLE 7.
TRANSFORMATION DATA

| Code | Donuts | | | | Conclusion |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 001 | Enough | Enough | Tall | Low | Buy |
| 002 | Enough | Enough | Tall | Tall | Buy |
| 003 | Enough | Enough | Tall | Very low | Buy |
| 004 | Enough | Enough | Very low | Tall | Don't buy |
| .... | .... | .... | .... | .... | .... |
| 027 | Low | Tall | Tall | Tall | Buy |

### 3.5 Data Mining

After all the data has gone through the previous 3 processes, then the classification process will be carried out, namely by forming a decision tree as the output.

Calculating the Entropy Value of each attribute:

Entropy (Total) with the following formula:

$$Enthropy(\text{Total}) = -\sum_{i=1}^{i} \frac{S_i}{S} Log_2 \frac{S_i}{S}$$

$$Enthropy(\text{Total}) = \left(-\frac{13}{27} * log_2\left(\frac{13}{27}\right)\right) + \left(-\frac{14}{27} * log_2\left(\frac{14}{27}\right)\right)$$

$$= 0{,}99$$

a. Packaging

$$Enthropy(\text{T})$$

$$= \left(-\frac{3}{3} * log_2\left(\frac{3}{3}\right)\right) + \left(-\frac{0}{3} * log_2\left(\frac{0}{3}\right)\right)$$

$$= 0$$

*Enthropy*(C)

$$= \left(-\frac{6}{11} * log_2\left(\frac{6}{11}\right)\right) + \left(-\frac{5}{11} * log_2\left(\frac{5}{11}\right)\right)$$
$$= 0{,}99$$

$$= 0{,}99$$

*Enthropy*(R)

$$= \left(-\frac{2}{5} * log_2\left(\frac{2}{5}\right)\right) + \left(-\frac{3}{5} * log_2\left(\frac{3}{5}\right)\right)$$
$$= 0{,}97$$

*Enthropy*(SR)

$$= \left(-\frac{3}{8} * log_2\left(\frac{3}{8}\right)\right) + \left(-\frac{5}{8} * log_2\left(\frac{5}{8}\right)\right)$$
$$= 0{,}95$$

For the calculation of Price entropy. Taste and Service are done using the same formula. After that calculate the gain of each attribute:

b. Gain(Total, Packaging)

$$= Enthropy(\text{S}) - \sum_{1=1}^{n} \frac{|\text{Packaging}_i|}{|\text{Total}|} * Enthropy(\text{Packaging}_i)$$
$$= 0{,}99 - \left(\left(\frac{3}{27} * 0\right)\right) + \left(\left(\frac{11}{27} * 0{,}99\right)\right) + \left(\left(\frac{5}{27} * 0{,}97\right)\right) + \left(\left(\frac{8}{27} * 0{,}95\right)\right)$$
$$= 1{,}86$$

For price gain calculation. Taste and Service are done using the same formula.

**TABLE 8.**
NODE 1 CALCULATION

| Node | Number of Cases (S) | | Buy(S1) | Don't Buy (S2) | Enthropy | Gain Info |
|---|---|---|---|---|---|---|
| 1 | Total | 27 | 14 | 13 | 0,99 | |
| | Packaging | | | | | 1,86 |
| | Tall | 3 | 3 | 0 | 0 | |
| | Enough | 11 | 6 | 5 | 0,99 | |
| | Low | 5 | 2 | 3 | 0,97 | |
| | Very low | 8 | 3 | 5 | 0,95 | |
| | Price | | | | | 1,26 |
| | Tall | 8 | 6 | 2 | 0,81 | |
| | Enough | 14 | 8 | 6 | 0,98 | |
| | Low | 4 | 0 | 4 | 0 | |
| | Very low | 1 | 0 | 1 | 0 | |
| | Taste | | | | | 1,03 |
| | Tall | 10 | 8 | 2 | 0,72 | |
| | Enough | 6 | 5 | 1 | 0,65 | |
| | Low | 8 | 1 | 7 | 0,54 | |
| | Very low | 3 | 0 | 3 | 0 | |
| | Service | | | | | 0,89 |
| | Tall | 15 | 9 | 6 | 0,97 | |
| | Enough | 8 | 3 | 5 | 0,95 | |
| | Low | 2 | 1 | 1 | 1 | |
| | Very low | 2 | 1 | 1 | 1 | |

From the calculations in table 11 above, it can be seen that the attribute with the highest gain is packaging with a gain value of 1.86, then the packaging attribute is used as the root node or node 1.
The next step is the completion to calculate node 1.1, node 1.2, and node 1.3 as roots for sufficient value, low value and very low value. The calculation is carried out in the same way as the previous method, namely by calculating the entropy value of the remaining attributes, namely price, taste and service by calculating the total entropy at sufficient packaging value, then calculating the gain for each attribute.
From the calculations that have been done, the decision tree can be as follows.

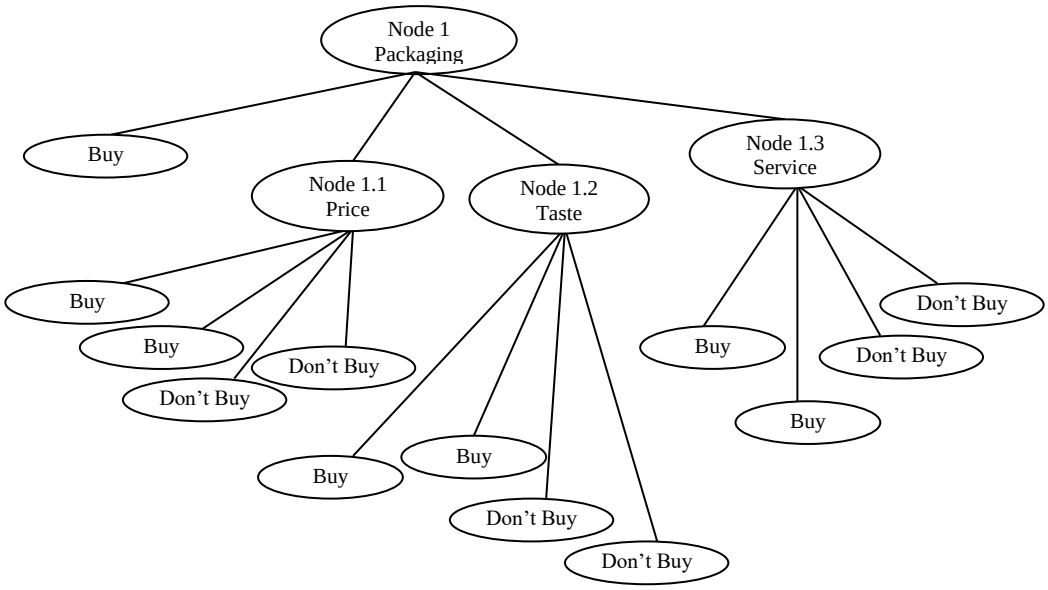


**Figure 2.** Calculation Result Decision Tree

The rules or rules formed based on the last decision tree as shown in Figure 4.4 above (Riandari and Simangunsong, 2019) are as follows:
1. IF Packaging = High, THEN Decision = Buy
2. IF Packaging = Enough AND Price = High, THEN Decision = Buy
3. IF Packaging = Enough AND Price = Enough, THEN Decision = Buy
4. IF Packaging = Enough AND Price = Low, THEN Decision = Don't Buy
5. IF Packaging = Enough AND Price = Very Low, THEN Decision = Don't Buy
6. IF Packaging = Low AND Service = High, THEN Decision = Buy
7. IF Packaging = Low AND Service = Enough, THEN Decision = Buy
8. IF Packaging = Low AND Service = Low, THEN Decision = Don't Buy
9. IF Packaging = Low AND Service = Very Low, THEN Decision = Don't Buy
10. IF Packaging = Very Low AND Taste = High, THEN Decision = Buy
11. IF Packaging = Very Low AND Taste = Enough, THEN Decision = Buy
12. IF Packaging = Very Low AND Taste = Low, THEN Decision = Don't Buy
13. IF Packaging = Very Low AND Taste = Very Low, THEN Decision = Don't Buy

Based on the decision tree and the rules formed, it can be concluded that the marketing strategy that can be done for products that are not selling well is to improve the packaging, because packaging is a root attribute that can be the main factor the product will sell or not. Then fix the price by adjusting the taste of the product and service from the employees.

## 4. Conclusion

Processing the questionnaire data to classify the indicators that most influence consumers in buying Home Industry products using the C.45 algorithm, a decision tree is obtained that can be used as a guide to determine marketing strategies for products that are not selling well. The decisions that are formed based on the decision tree and rules can be concluded that the marketing strategy that can be done for products that are not selling well is to improve the packaging, because packaging is a root attribute that can be the main factor the product will sell or not. Then fix the price by adjusting the taste of the product and service from the employees.

## References

[1] Anam, C., & Santoso, H. B. (2018). Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa. Jurnal Ilmiah Ilmu-Ilmu Teknik, 8(1), 13–19. https://ejournal.upm.ac.id/index.php/energy/article/view/111/449.

[2] Arifin, M. F., & Fitrianah, D. (2018). Penerapan Algoritma Klasifikasi C4.5 Dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus : PT Atria Artha Persada. InComTech, 8(2), 87–102. https://doi.org/10.22441/incomtech.v8i1.2198.

[3] Elisa, E. (2017). Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti. Jurnal Online Informatika, 2(1), 36. https://doi.org/10.15575/join.v2i1.71

[4] Firman, A., Wowor, H. F., Najoan, X., Teknik, J., Fakultas, E., & Unsrat, T. (2016). Sistem Informasi Perpustakaan Online Berbasis Web. E-Journal Teknik Elektro Dan Komputer, 5(2), 29–36.

[5] Hendini, A. (2016). Pemodelan Uml Sistem Informasi Monitoring Penjualan Dan Stok Barang. Jurnal Khatulistiwa Informatika, 2(9), 107–116. https://doi.org/10.1017/CBO9781107415324.004

[6] Luluk Elvitaria, M. H. (2017). Smk Analisis Kesehatan Abdurrab Menggunakan Algoritma. (Jurnal Teknologi Dan Sistem Informasi Univrab, 2(2), 220–233.

[7] Lumbantoruan, R. (2015). Analisis data mining dan warehousing. Ilmiah, 19(Vol 19 No 1 (2015): Vol 19 No 1 (2015): Buletin Ekonomi ISSN: 1410-3842), 66–76. http://ejournal.uki.ac.id/index.php/beuki/article/view/308.

[8] Munawar. 2019. Analisis Perancangan Sistem Berorientasi Objek dengan UML. Informatika Bandung.

[9] Nursanti, Z. A. (2019). PERAN HOME INDUSTRY DALAM RANGKA PEMBERDAYAAN EKONOMI MASYARAKAT (Studi Pada Produksi Roti Jahe SARI Desa Lebeng Kecamatan Sumpiuh Kabupaten Banyumas) Zahra Aulia Nursanti NIM. 1522104032.

[10] Pratiwi, Y. A., Ginting, R. U., Situmoran, H., & Sitanggang, R. (2020). Perancangan Sistem Informasi Akademik Berbasis Web Di Smp Rahmat Islamiyah. Jurnal Teknologi, Kesehatan Dan Ilmu Sosial, 2(1), 27–32.

[11] Riandari, F., & Simangunsong, A. 2019. Penerapan Algoritma C4.5 Untuk Mengukur Tingkat Kepuasan Mahasiswa. CV. Rudang Mayang.

[12] Riandari, F., & Simangunsong, A. 2019. Buku Ajar Data Mining: Dengan Penerapan Algoritma C4.5. CV. Rudang Mayang.

[13] Safii, M. (2019). Implementasi Data Mining Dengan Metode Pohon Keputusan Algoritma Id3 Untuk Menentukan Status Mahasiswa. Jurnal Mantik Penusa, 2(1), 82–86.

[14] Samsudin, M., Abdurahman, M., & Abdullah, M. H. (2019). Sistem Informasi Pengkreditan Nasabah Pada Koperasi Simpan Pinjam Sejahtera Baru Kota Ternate Berbasis Web. Jurnal Ilmiah ILKOMINFO - Ilmu Komputer & Informatika, 2(1), 11–23. https://doi.org/10.47324/ilkominfo.v2i1.16.

[15] Suendri. (2018). Implementasi Diagram UML (Unified Modelling Language) Pada Perancangan Sistem Informasi Remunerasi Dosen Dengan Database Oracle (Studi Kasus: UIN Sumatera Utara Medan). Jurnal Ilmu Komputer Dan Informatika, 3(1), 1–9. http://jurnal.uinsu.ac.id/index.php/algoritma/article/download/3148/1871.

[16] Tusarwenda, T. R. I. B. (2018). Penerapan data mining dengan algoritma c4.5 dalam prediksi penjualan botol pada cv. seribukilo.