Jurnal ICT : Information and Communication Technologies, 16 (2) (2025) 320-328

Published by: Marq & Cha Institute

Jurnal ICT : Information and Communication Technologies

Journal homepage: www.ejournal.marqchainstitute.or.id/index.php/JICT

Determining Sales Based on Goods Data Classification Using the Web-Based C4.5 CRISP-DM Method

Bayu Gunawan[1], Wirhan Fahrozi[2]

[1,2] Sistem Informasi, Universitas Potensi Utama, Medan, Indonesia

Article Info

Article history

Received : Oct 20, 2025
Revised : Oct 28, 2025
Accepted : Oct 30, 2025

Keywords:

C4.5 Algorithm;
CRISP-DM;
Data Mining;
Decision Support System;
Sales Classification.

Abstract

The increasing complexity of product distribution and sales activities at PT Mitra Bersama has created challenges in accurately classifying product performance, particularly due to manual data processing that is inefficient and prone to error. To address this issue, this study aims to develop an intelligent decision-support system capable of classifying best-selling and non-best-selling products using a data-driven approach. The CRISP-DM methodology was applied to guide the overall analytical process, consisting of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The C4.5 algorithm was used to perform the classification through entropy and information gain calculations to determine the most influential attributes. The results show that Type of Food has the highest information gain (0.0113340), followed by Initial Stock, Unit, Month, and Ending Stock, indicating that product characteristics and early inventory levels play a significant role in predicting sales performance. These findings were implemented into a web-based application to facilitate real-time classification and assist decision-makers in optimizing inventory planning, distribution strategies, and sales forecasting. This research contributes to improving organizational efficiency by providing a systematic, accurate, and accessible tool that supports better strategic decision-making in product sales management.

Corresponding Author:

Bayu Gunawan
Sistem Informasi,
Universitas Potensi Utama,
Jl.K.L Yos Sudarso KM 6.5 Tj.Mulia, Medan, 20241, Indonesia
Email : gunawanbayu255@gmail.com

1. Introduction

In today's intensely competitive business landscape, the ability to forecast sales accurately is vital for sustaining long-term success. As companies vie for market share, entrepreneurs must continuously innovate and improve their strategies to maintain a competitive edge. Sales prediction plays a central role in this effort because it directly drives revenue and profitability, which are fundamental to the company's survival and growth. Without reliable forecasting, firms risk overstocking, under-stocking, or misallocating resources, jeopardizing both operational efficiency and financial stability. Therefore, a robust decision support system that leverages historic sales data to anticipate future patterns is critical. Such a system enables companies to respond proactively to market fluctuations by informing inventory management, promotional campaigns, and expansion plans. Indeed, leveraging data not merely for operational reporting but as a strategic asset is increasingly recognized as a best practice in modern business management.

The rapid evolution of information technology in the digital age has revolutionized how organizations manage and analyze their data. Traditional, manual methods of recording sales transactions are no longer sufficient to handle the volume, variety, and velocity of data generated in modern commerce. As a result, companies are shifting to automated, data-driven approaches to glean deeper insights and extract value from their records. Data mining—an advanced analytics technique—has emerged as a powerful way to extract hidden patterns and relationships within large datasets (e.g., customer buying habits, product lifecycle trends, and seasonal variations). By applying data mining, firms can move beyond descriptive reporting and adopt predictive models that anticipate future behavior. In particular, classification algorithms such as C4.5 have been successfully used in prior studies to distinguish between high-performing and low-performing products based on attributes like price, sales frequency, and time periods (Sag & Hasugian, 2024; Wijaya & Bakti, 2023).

To ensure that the data mining process is systematic, reliable, and aligned with business objectives, this study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. CRISP-DM is widely regarded as a de facto standard in the field, providing a structured, iterative framework that guides the full lifecycle of data mining projects (Shearer, 2000; Larose, 2014). The methodology comprises six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al., 2000). Its non-linear, adaptive nature allows for feedback loops; for example, insights discovered during modeling may prompt a return to data preparation or business understanding. By following this rigorous process, the research ensures that the resulting predictive model is not only technically sound but also practically relevant and actionable.

At the heart of this study is the C4.5 algorithm, a decision-tree classifier known for its interpretability and ability to handle both categorical and continuous variables. The algorithm builds a tree based on information gain and gain ratio, producing classification rules that are easily understandable by business decision-makers. Prior research across various domains confirms the efficacy of C4.5 in predicting sales and other business outcomes. For instance, studies have applied C4.5 to forecast telecommunication and IT equipment sales (Wijaya & Bakti, 2023), classify customer traits in consumer product companies (Siswandi, Anwar, Susilo, & Hasibuan, 2024), and even predict profitability in retail settings (Calvin & Harman, 2022). These applications demonstrate that C4.5 not only achieves high accuracy in classification tasks but also yields human-readable rules that support managerial decision-making.

Building on this foundation, the present research aims to develop a web-based sales classification system for PT Mitra Bersama, a distribution company facing challenges in managing its growing variety of products and high transaction volume. Currently, reports are generated manually, which is labor-intensive, prone to errors, and slow. By implementing a CRISP-DM–guided data mining pipeline and applying the C4.5 algorithm, the system will classify items into “best-selling” and “non-best-selling” categories using features such as item type, price, sales frequency, and period of sale. The web application will allow management to access real-time classification insights, enabling rapid, data-driven decisions about inventory, promotions, and distribution. Ultimately, this research not only aims to improve operational efficiency and forecasting accuracy but also to empower PT Mitra Bersama to formulate strategic business plans grounded in data.

2. Research Methodolgy

In this study, the CRISP-DM methodology is applied as a general problem-solving approach for various business and research problems. This methodology consists of six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each stage in this methodology can be explained as follows:

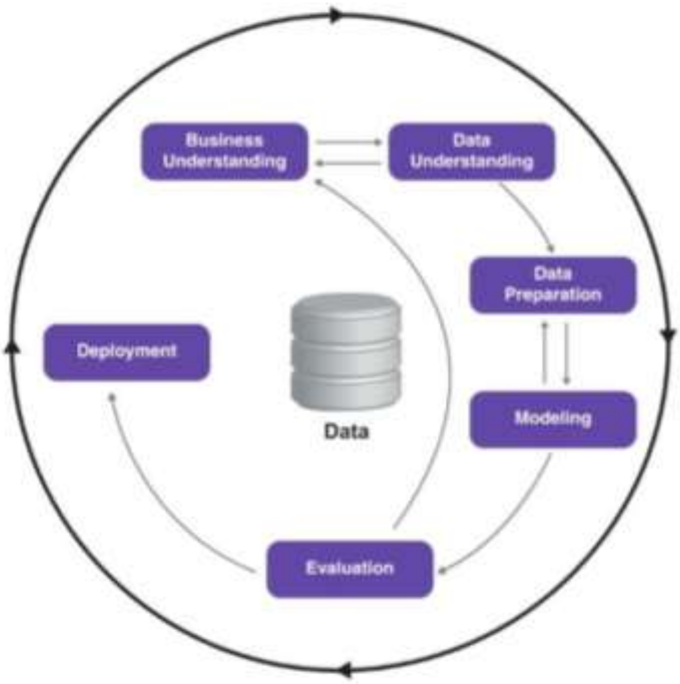

Figure 1 CRISP-DM Method

The explanation of the CRISP-DM method above is as follows:

1) Business Understanding
   This stage is the initial step, starting with understanding the needs and objectives from a business perspective, transforming knowledge into a problem definition in the context of data mining, and then planning and developing strategies to achieve the objectives in the data mining process.
   a) Observation
      Observation is a data collection method conducted by directly observing and recording events or behavior without intervention or questioning the subject. Researchers conduct observations by systematically observing the situation or object of research.
   b) Interviews
      Interviews are a data collection method that involves direct interaction between the researcher and the respondent. The researcher asks specific questions to obtain the desired information. Interviews can be conducted formally or informally. The researcher conducted direct interviews with the sales department at PT. Mitra Bersama.
   c) Sampling
      Sampling is the process of selecting a small, representative portion of the population as the object of research due to time and resource constraints. In research, it is often impossible or inefficient to collect data from the entire population. Therefore, researchers select a sample that is a subset of the population considered to represent the general characteristics of the entire population.
2) Data Understanding
   This stage begins with data collection, data description, and data quality evaluation. The process includes steps to understand the nature and characteristics of the data to be used.

3) Data Preparation
   This stage involves creating the final dataset from the collected raw data. This process includes activities such as data cleaning, data selection (data selection), both records and attributes,

and data transformation (data transformation). These steps aim to present data ready for use as input in the modeling process.

4) Modeling
This stage applies machine learning to determine appropriate data mining techniques, tools, and algorithms. This study uses the C4.5 classification model.
5) Evaluation
This stage involves assessing the performance of the patterns generated by the algorithm. Evaluation is conducted using a Confusion Matrix to measure accuracy, precision, and recall. Used to test the model against a separate test dataset to evaluate its performance.
6) Deployment
This stage is carried out for the system design process using UML, consisting of use case diagrams, class diagrams, sequence diagrams, and activity diagrams. Then, the database design and interface design processes are carried out.

3. Results and Discussion

The steps in creating a decision tree for determining sales based on the classification of product data are as shown in the table above. The C4.5 algorithm calculation can be described as follows:
Step 1: Calculate the number of cases, the number of cases for "Not Selling" decisions, the number of cases for "Seller" decisions, and the entropy of all cases divided by the attributes: food type, unit, beginning stock, ending stock, month, and description. Then, calculate the gain for each attribute.
Step 2: Calculate the entropy for the total row using equation (1). Next, calculate the gain for each attribute using equation (2). The following are the entropy and gain calculations.
The following is the total description from the table.

Table 1 Total Table

No	Information	Amount
1	Best Seller	43
2	Not Selling Well	57

Menghitung Entropy total :

Entropy [Total] $= (-\frac{57}{100} xlog2 \frac{57}{100}) + (-\frac{43}{100} xlog2 \frac{43}{100})$
$= (-\frac{57}{100} x0{,}30103\text{x} \frac{57}{100}) + (-\frac{43}{100} x0{,}30103\text{x} \frac{43}{100})$
$= (-\frac{57}{100} x0{,}30103\text{x}0{,}57) + (-\frac{43}{100} x0{,}30103\text{x}0{,}43)$
$= (-\frac{57}{100} x\ 0{,}17158) + (-\frac{43}{100} x0{,}12944)$
$= (-0.57\ x\ 0{,}17158) + (-0{,}43\ x0{,}12944)$
= -0,0978 + (-0,0556)
=- 0,1534

Calculating entropy and gain Type of food:

Table 2.Types of food

No	Information	Best Seller	Not Selling Well	Total
1	Fresh snacks	20	42	62
2	Packaged Snacks	23	15	38

Packaged Snacks = [23,15] = 38
= (-(23/38) * 0.30103 * (23/38)) - ((15/38) * 0.30103 * (15/38))
Entropy Packaged Snacks = -0.15719
Fresh snacks = [20,42] = 62
= (-(20/62) * 0.30103 * (20/62)) - ((42/62) * 0.30103 * (42/62))

Entropy Fresh snacks = -0.16947
IG = -0.15347 - ((38/100) * -0.15719) + ((62/100) * -0.16947) = 0.0113340

Calculating entropy and gain Units:

Tabel 3. Units

No	Information	Best Seller	Not Selling Well	Total
1	Packaging/Box	1	3	4
2	Pouch	6	8	14
3	Sachet	17	27	44
4	Gram	19	19	38

Packaging/Box = [1,3] = 4
Packaging/Box = (-(1/4) * 0.30103 * (1/4)) - ((3/4) * 0.30103 * (3/4))
Entropy Packaging/Box = -0.18814
Pouch = [6,8] = 14
Pouch = (-(6/14) * 0.30103 * (6/14)) - ((8/14) * 0.30103 * (8/14))
Entropy Pouch = -0.15359
Sachet = [17,27] = 44
Sachet = (-(17/44) * 0.30103 * (17/44)) - ((27/44) * 0.30103 * (27/44))
Entropy Sachet = -0.15829
Gram = [19,19] = 38
Gram = (-(19/38) * 0.30103 * (19/38)) - ((19/38) * 0.30103 * (19/38))
Entropy Gram = -0.15052
IG = -0.15347 - ((4/100) * -0.18814) + ((14/100) * -0.15359) + ((44/100) * -0.15829) + ((38/100) * -0.15052) = 0.0024030
Calculating the entropy and gain of Initial Stock:

Tabel 4. Initial Stok

No	Information	Best Seller	Not Selling Well	Total
1	Many	23	16	39
2	Moderate	17	34	51
3	Littel	3	7	10

Many= [23,16] = 39
Many= (-(23/39) * 0.30103 * (23/39)) - ((16/39) * 0.30103 * (16/39))
Entropy Many= -0.15536
Moderate = [17,34] = 51
Moderate = (-(17/51) * 0.30103 * (17/51)) - ((34/51) * 0.30103 * (34/51))
Entropy Moderate = -0.16724
Littel= [3,7] = 10
Littel= (-(3/10) * 0.30103 * (3/10)) - ((7/10) * 0.30103 * (7/10))
Entropy Littel= -0.1746
IG = -0.15347 - ((39/100) * -0.15536) + ((51/100) * -0.16724) + ((10/100) * -0.1746) = 0.0098730
Calculating the entropy and gain of the ending stock:

Table 5. Ending Stock

No	Information	Best Seller	Not Selling Well	Total
1	Many	2	3	5
2	Moderate	13	16	29
3	Littel	28	38	66

Many= [2,3] = 5
Many= (-(2/5) * 0.30103 * (2/5)) - ((3/5) * 0.30103 * (3/5))
Entropy Many= -0.15654
Moderate = [13,16] = 29
Moderate = (-(13/29) * 0.30103 * (13/29)) - ((16/29) * 0.30103 * (16/29))
Entropy Moderate = -0.15213
Littel= [28,38] = 66
Littel= (-(28/66) * 0.30103 * (28/66)) - ((38/66) * 0.30103 * (38/66))
Entropy Littel= -0.15397
IG = -0.15347 - ((5/100) * -0.15654) + ((29/100) * -0.15213) + ((66/100) * -0.15397) = 0.0000950
Calculating the entropy and gain of the Moon:

Tabel 6. moon

No	Information	Best Seller	Not Selling Well	Total
1	April	8	16	24
2	Mei	23	26	49
3	Juni	12	15	27

April = [8,16] = 24
April = (-(8/24) * 0.30103 * (8/24)) - ((16/24) * 0.30103 * (16/24))
Entropy April = -0.16724
Mei = [23,26] = 49
Mei = (-(23/49) * 0.30103 * (23/49)) - ((26/49) * 0.30103 * (26/49))
Entropy Mei = -0.15108
Juni = [12,15] = 27
Juni = (-(12/27) * 0.30103 * (12/27)) - ((15/27) * 0.30103 * (15/27))
Entropy Juni = -0.15237
IG = -0.15347 - ((24/100) * -0.16724) + ((49/100) * -0.15108) + ((27/100) * -0.15237) = 0.0018370
The calculation results are shown in Table 7. Node 1 Calculation

Table 7. Entropy Calculation Results

Root node 1		Amount	Not Selling Well	Best Seller	Entropy	Gain
Total		100	57	43	-0.15347	
Types of food	Packaged Snacks	38	12	15	-0.15719	0.0113340
	Fresh snacks	62	20	42	-0.16947	
Unit	Packaging/Box	4	3	1	-0.18814	0.0024030
	Pouch	14	8	6	-0.15359	
	Sachet	44	27	7	-0.15829	
	Gram	38	19	19	-0.15052	
Initial Stok	Many	39	16	13	-0.15536	0.0098730
	Moderate	51	34	17	-0.16724	
	Littel	10	7	3	-0.1746	
Final Stock	Many	5	3	2	-0.15654	0.0000950
	Moderate	29	16	13	-0.15213	
	Littel	66	38	28	-0.15397	
Month	April	24	16	8	-0.16724	0.0018370
	Mei	49	26	23	-0.15108	
	Juni	27	15	12	-0.15237	

Thus, the conclusions from the table above can be seen as follows:

Table 8. Sequence of Entropy Calculation Results

No	Atribut	Gain
1	Jenis	0.0113340
2	Stok Awal	0.0098730
3	Satuan	0.0024030
4	Bulan	0.0018370
5	Stok Akhir	0.0000950

Discussion

The results of the entropy and information gain calculations provide clear evidence regarding the most influential attributes in determining product sales classification using the C4.5 algorithm. Based on the findings, Type of Food emerges as the attribute with the highest information gain (0.0113340), indicating that this variable contributes the most to reducing uncertainty in differentiating between "Best Seller" and "Not Selling Well" items. This suggests that consumer purchasing behavior is highly sensitive to the categorization of goods, specifically between fresh snacks and packaged snacks. Fresh snacks exhibit a higher entropy value, implying greater variability in their sales performance, which aligns with the perishable nature and fluctuating demand typical of such products. Meanwhile, packaged snacks show more stable purchasing patterns. The second most influential attribute is Initial Stock (gain = 0.0098730), highlighting that the quantity of goods available at the beginning of the sales period significantly correlates with sales outcomes. Products with larger initial stock volumes tend to achieve better sales classifications, possibly due to higher availability and better distribution coverage. These results demonstrate that stock management and product type differentiation play critical roles in shaping sales performance at PT. Mitra Bersama.

The remaining attributes—Unit, Month, and Ending Stock—contribute lower gains, indicating a relatively weaker influence on the classification process. The Unit attribute (gain = 0.0024030) shows that packaging form (such as sachet, pouch, or box) has some relevance, though it is not a primary determinant of sales category. Similarly, temporal factors represented by the Month attribute (gain = 0.0018370) suggest minor seasonal variations in purchasing behavior but do not significantly alter classification outcomes. Ending Stock yields the lowest gain (0.0000950), indicating that remaining inventory at the end of the period offers minimal predictive value when compared to initial stock. This pattern reinforces the importance of early-period product distribution and inventory planning over end-of-period stock levels. Overall, the hierarchical ranking of attributes reflects the effectiveness of the C4.5 algorithm in identifying key drivers of sales performance, providing PT. Mitra Bersama with actionable insights for optimizing procurement, inventory management, and marketing strategies based on data-driven prioritization.

4. Conclusion

This study demonstrates that the integration of the CRISP-DM methodology and the C4.5 classification algorithm provides an effective, systematic, and data-driven approach to predicting product sales performance at PT Mitra Bersama. The analysis of entropy and information gain reveals that Type of Food is the most influential attribute in determining whether a product falls into the "Best Seller" or "Not Selling Well" category, followed by Initial Stock, Unit, Month, and Ending Stock. These findings confirm that consumer purchasing patterns are strongly associated with product characteristics and early-period inventory conditions. By implementing a web-based classification system that capitalizes on these insights, the company can significantly enhance its ability to forecast sales, manage inventory, optimize distribution, and make informed strategic decisions. Overall, the developed model not only improves operational efficiency but also contributes to more accurate and actionable predictive analytics within the organization. Given the results of this research, several strategic recommendations are proposed to further strengthen PT Mitra Bersama's data-driven decision-making capabilities. First, the company should prioritize refining product categorization and stock planning, as these attributes exhibit the highest predictive power and thus influence sales performance most significantly. Second, it is recommended that the web-based system be continuously updated with new data to allow iterative model improvements, enabling the decision-tree classifier to adapt to evolving market dynamics and consumer behavior. Additionally, future research could incorporate more advanced machine learning models such as Random Forest, Gradient Boosting, or Neural Networks to compare predictive accuracy and robustness. Integrating real-time data streams and expanding the attribute set—such as promotion intensity, competitor pricing, or customer demographics—may also enhance model performance. By adopting these recommendations, PT Mitra Bersama can build a more resilient and intelligent sales forecasting ecosystem that supports long-term business growth and competitive advantage.

References

Audrilia, M., & Budiman, A. (2020). Perancangan Sistem Informasi Manajemen Bengkel Berbasis Web (Studi Kasus : Bengkel Anugrah). *Jurnal Madani : Ilmu Pengetahuan, Teknologi, Dan Humaniora*, *3*(1), 1–12. https://doi.org/10.33753/madani.v3i1.78

Fatawa Imam Al Muftin, & Fendi Hidayat. (2024). Sistem Informasi Penjualan. *Zona Komputer: Program Studi Sistem Informasi Universitas Batam*, *13*(3), 232–237. https://doi.org/10.37776/zkomp.v13i3.1461

Gustami, E., & Astuti, K. P. (2022). Klasifikasi Barang Menggunakan Algoritma C4. 5 Dalam Penentuan Prediksi Stok Barang Pada Pt Aerofood Indonesia. *Reputasi: Jurnal Rekayasa Perangkat …*, *3*(2), 12–18.

Hidayati, N., Suntoro, J., & Setiaji, G. G. (2021). Perbandingan Algoritma Klasifikasi untuk Prediksi Cacat Software dengan Pendekatan CRISP-DM. *Jurnal Sains Dan Informatika*, *7*(2), 117–126. https://doi.org/10.34128/jsi.v7i2.313

Indah Sari, A., & Sewaka. (2022). Prediksi Penjualan Sepatu Dengan Algoritma C4.5 Di T&T Collection Di Tangerang. *OKTAL : Jurnal Ilmu Komputer Dan Science*, *1*(9), 1435–1442.

Janis, J. W., Mamahit, D. J., Sugiarso, B. A., Rumagit, A. M., Elektro, T., Sam, U., & Manado, R. (2020). Rancang Bangun Aplikasi Online Sistem Pemesanan Jasa Tukang Bangunan Berbasis Lokasi. *Jurnal Teknik Informatika*, *15*(1), 1–12. https://doi.org/10.35793/jti.15.1.2020.29023

Leasiwal, A. A., Andrian, A. M., & Masala, I. V. (2021). Implementasi Algoritma C4. 5 Untuk Klasifikasi Harga Emas. *Jurnal Teknik Informatika Unika St. Thomas (JTIUST)*, *06*, 1–12.

Narayana Sakti Aji, Fauzan Natsir, & Siti Istianah. (2023). Penentuan Penjualan Barang Berdasarkan Pengelompokan Produk dengan K-Means Clustering Metode CRISP-DM Pada CV Sembako Dina. *Journal Zetroem*, *5*(2), 119–126. https://doi.org/10.36526/ztr.v5i2.3041

Navisa, S., Luqman Hakim, & Aulia Nabilah. (2021). Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM. *Jurnal Sistem Cerdas*, *4*(2), 114–125. https://doi.org/10.37396/jsc.v4i2.162

Pala'langan, E. I., Saerang, D. P. E., & Gamaliel, H. (2020). Analysis of Accounting Information System Sales At Pt. Wahana Wirawan Manado-Nissan Datsun Martadinata. *Jurnal EMBA*, *8*(4), 824–834.

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.

Priyatna, A., & Sanwani. (2024). Evaluasi Pemahaman Siswa Dalam Proses Belajar Secara Online dengan Menggunakan Algoritma C5.0. *Resolusi : Rekayasa Teknik Informatika Dan Informasi*, *4*(3), 300–309.

Ramayu, I. M. S., Susanto, F., & Mahendra, G. S. (2022). Penerapan Data Mining Dengan Algoritma C4.5 Dalam Pemesanan Obat Guna Meningkatkan Keuntungan Apotek. *Prosiding Seminar Nasional Manajemen, Desain & Aplikasi Bisnis Teknologi (SENADA)*, *5*, 237–245.

Salsabilla, T., & Sulastri, S. (2022). Implementasi Algoritma C4.5 Untuk Klasifikasi Produk Laris Sepeda Motor Honda Pada Cv Cendana Motor Cepiring. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *7*(2), 164–171. https://doi.org/10.36341/rabit.v7i2.2489

Saputra, A. D., & Qoiriah, A. (2022). Penerapan Algoritma C4.5 Untuk Mengatur Persediaan Stok Barang Berbasis Website. *Journal of Informatics and Computer Science (JINACS)*, *3*(04), 481–493. https://doi.org/10.26740/jinacs.v3n04.p481-493

Shedriko, S., & Firdaus, M. (2022). Penentuan Klasifikasi Dengan Crisp-Dm Dalam Memprediksi Kelulusan Mahasiswa Pada Suatu Mata Kuliah. *Semnas Ristek (Seminar Nasional Riset Dan Inovasi Teknologi)*, *6*(1), 826–831. https://doi.org/10.30998/semnasristek.v6i1.5814

Suhartini, Sadali, M., & Putra, Y. K. (2020). Sistem Informasi Berbasis Web Sma Al- Mukhtariyah Mamben Lauk Berbasis Php Dan Mysql . *Jurnal Informatika Dan Teknologi*, *3*(1), 79–83.

Wahyudi, M. D. (2023). Penerapan Data Mining Dengan Algoritma C4. 5 Dalam Prediksi Penjualan Buku. *Jurnal Teknorama (Informatika Dan …*, *1*(1), 1–6.