Journal of Emerging Information Systems and Business Intelligence
ISSN: 2774-3993
Journal homepage: https://ejournal.unesa.ac.id/index.php/JEISBI/

Sales Performance Classification of Promotional Products Using Data Mining

Rafli Satria Iswandaru[1], Aries Dwi Indriyanti[2]
[1,2]*Universitas Negeri Surabaya, Surabaya, Indonesia*
rafli.20111@mhs.unesa.ac.id,ariesdwi@unesa.ac.id

ABSTRACT

The objective of this research is to establish a classification model to determine high sales performance promotional products using past sales records. The issue at stake is that it is very hard for business actors to forecast the appearance of high sales promotional products, taking into account different factors, such as product type, price per unit, quantity requested, and sales period. This study, based on a quantitative and experiential manner, makes use of the C4. 5 decision tree algorithms on real transaction data of HERA Promotion during 2024. The data falls into one of two types: "best-selling" and "not-selling" products. The proposed classification model achieved good generalization performance with the test accuracy of 99.48% and 5-fold cross-validation accuracy of 96.77%. Price Unit, Product Name, and Month were the most essential features in classifying products, showing that economic value and seasonal demand are major factors determining whether a product is sold. But when applied to the external data for the first few months of 2025, accuracy fell to 78%, which is a way for them to show that shifts in consumer behavior can drive changes in performance. From a theoretical point of view, this study fills the gap by incorporating the time effects as a dynamic variable into product classification models, which haven't been mentioned much in previous research. For practice, the results also encourage incorporating data-driven classification models within decision support systems to help with stock planning and promotional strategies. More work is warranted to use ensembling techniques and real-time data streams on the generalization ability and adaptability of the models.

Keywords: C4.5 Algorithm, Data Mining, Product Classification, Sales, Information Systems.

Article Info:
Article history:
Received February 09, 2026
Revised February 16, 2026
Accepted April 27, 2026

Corresponding Author
Rafli Satria Iswandaru
Universitas Negeri Surabaya, Surabaya, Indonesia
rafli.20111@mhs.unesa.ac.id

1. INTRODUCTION

The advancement of IT and business forms of digitization has significantly influenced the trade, marketing, and competitive environment. Companies need to be more agile when it comes to consumer behavior and quicker markets. Output sales are influenced by one of the problems in real-world business: determining factors influencing the level of product sales accurately and promptly. There are patterns in the increasing transaction data every day that could be used for strategic decision-making. In this scenario, data mining is a key solution to discover insights from the vast amount of sales data [1].

A classification technique that is often employed in sales research is the Decision Tree C4. 5 algorithm. This algorithm can deal with categorical and numerical data as well as generate interpretable models. It was recently reported that the C4. 5 in sales ranking provides an indication to company of which product is currently selling and not selling based on past history [2], [3]. For example, Musa [4] and Fadhila [5] found that the C4. 5 is quiteaccurate in predicting the product type, prices, and sales time.

However, the majority of the previous research used one or two sales attributes and focused on the retail and e-commerce industries, to which relatively homogeneous market assets are applied [6],[7]. There have been few studies in which the product type, sales time, unit price, and order quantity were integrated together in terms of careful consideration simultaneously, especially specializing in the seasonal product promotion field. Meanwhile, temporal time features have not been sufficiently optimized in classification modeling, and product demands are subject to seasonal wave effect [8].

In light of the above problems, this paper proposes using the C4.5 algorithm for sales data of the HERA Promotion company, which is running an E-commerce business, and selling gifts (t-shirts, tumblers, pins, mugs. The classification model relies on four dimensions: the category of product, the price range of product, the quantity ordered, and the time dimension. Using these four elements, a classification model which is accurate and business relevant [9], [10] can be generated.

The major contribution of this work is the simultaneous combination of various properties and their inclusion in one sales classifier trained on a real-world, large-scale dataset. In addition, this work evaluates the model on out-of-sample data to investigate generalization of the algorithm. The purpose of the classification results is to develop a better understanding, not just in information management for more effective marketing strategies as well. Therefore, we anticipate that the results of this study will help lay a solid theoretical foundation for data mining applications used in marketing decision support for product promotion [11]–[15].

2. METHODS

An experimental quantitative approach has been adopted in this study in order to model a classification for bestseller and non-bestseller use cases by employing real sales information. This method is selected based on the fact that it provides statistically meaningful results and can be tested using established classification models [1], [4]. The data we used is secondary data that was derived from an internal system of HERA Promotion, and it included such attributes as product type, unit price, order quantity, and sales transaction time spanned from January to December 2024. The first step in the research was preparing data by cleaning, filling in missing values, and changing the transaction time to a variable month. The next step was data labeling: we categorized the product status into best-selling and non-best-selling based on a certain volume of sales. Descriptive analysis was carried out first to analyze patterns of sales distribution on a monthly and an annual basis, before modeling. The C4. 5 Algorithm has been selected due to its ability to work jointly with numerical data and categorical data, and generate an understandable model as a decision tree [5], [7]. The model was trained using the Python code in Decision Tree Classifier, and tuned through hyperparameter tuning with Grid Search CV. K-Fold Cross-Validation was used to prevent overfitting. Sales classification research [6], [11] generally evaluates the model performance in terms of accuracy, precision, recall, and F1-score. Moreover, the sales data of the first quarter of 2025 has also been used as external data to evaluate if such a model is able to generalize on new, not already seen samples, as stated in robust and application-oriented classification-based

machine learning papers [13], [15]. The entire flow process of the research stages is outlined in Figure 1 (Research Flowchart).

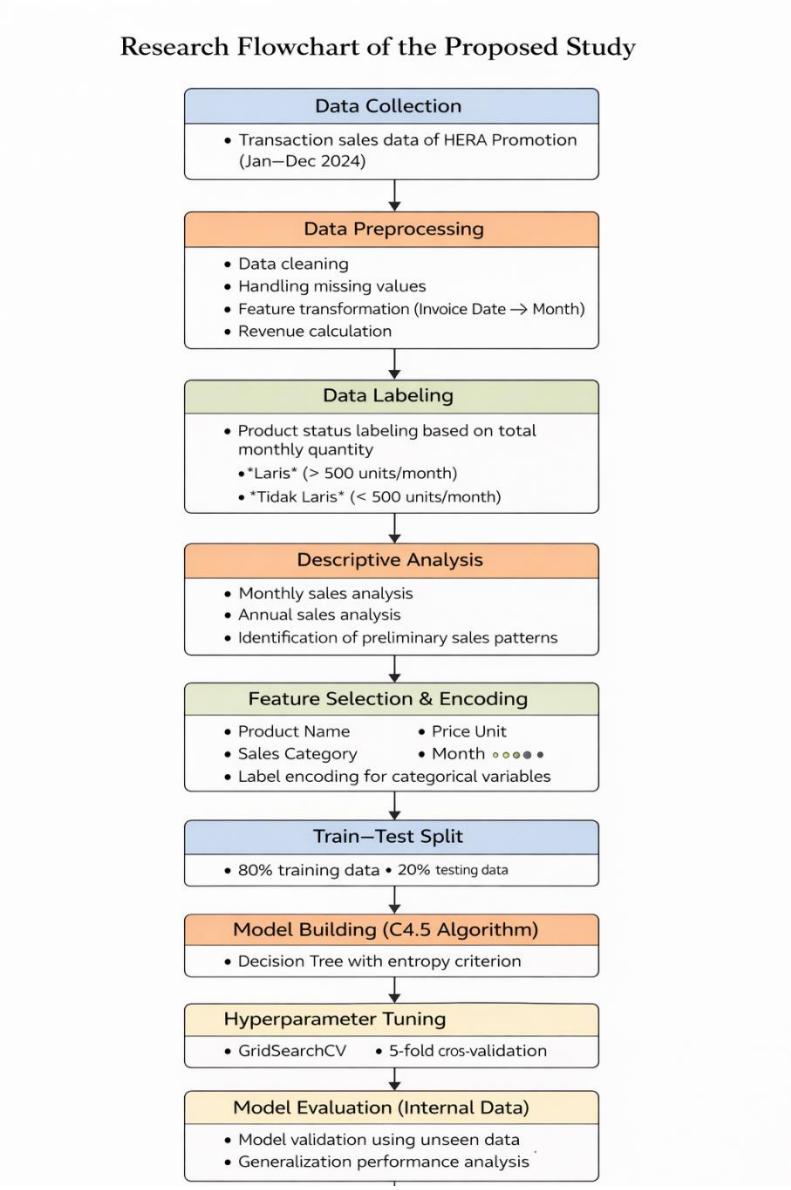

Figure 1. Research Flowchart

3. RESULTS AND DISCUSSION

3.1 Results

This paper generates a model of classification of sales of promotional products with the C4.5 approach using 2024 sales data from HERA Promotion. According to the data processing and training results, the C4. 5 can generate a decision tree structure to summarize the relationship among the primary attributes Product Name, Price Unit, Order Quantity, and Month. (3) The model built can classify sales data into ‘selling’ and ‘not selling’ successfully.

The visualization of the decision tree model is shown in Figure 2, where the principal partition leads to the Price Unit category, which means that product price plays as the most crucial factor to distinguish products' sales. The next split is via Product Name and Month, hinting to us that the demand for the product has a seasonal effect. This model also demonstrates that low-priced and a few types of products fall in the best-selling class, for example, keychains and Brochure Printing, etc.

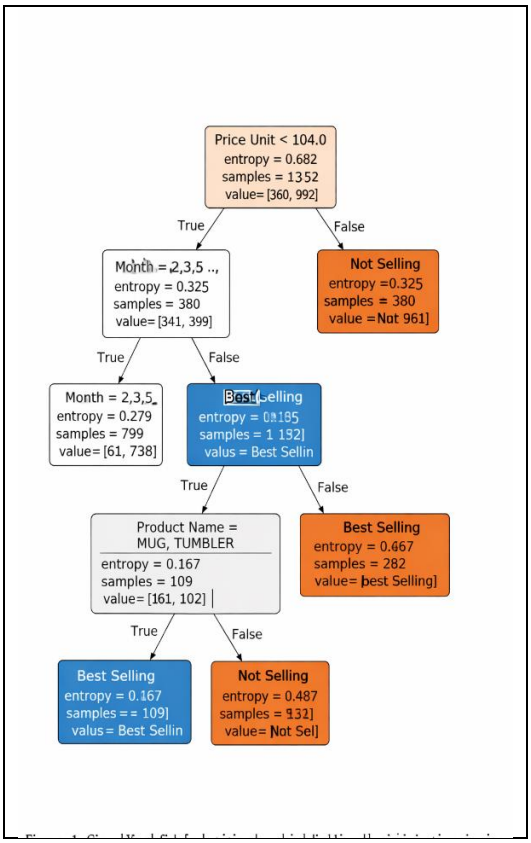

Figure 2. Visualization of the C4.5 decision tree for sales classification

On test data, the model had an accuracy of 99.48%, along with precision, recall, and f1-score greater than 0.98. These findings suggest that the model does an excellent job of discriminating between both classes, and with very few classification errors.

Table 1. Classification model performance on the test data assessment outcomes

Metric	Value (%)
Accuracy	99.48%
Precision	99.22%
Recall	99.42%
F1-Score	99.32%

Moreover, we conducted the cross-validation evaluation with k-fold cross-validation with 5 folds and found that the average accuracy based on four trials was 96.77%, which enhanced the stability of performance of models on different subsets of data. The predictive performance reduced to 78% when the model was applied on a new dataset (first quarter of 2025), suggesting that PP sales patterns change over time, and the model should be updated frequently so as to be able to capture the relevant information.

Table 2. Comparison of model accuracy on training data, cross-validation, and external data

Dataset	Accuracy (%)
Test Data (2024)	99.48
Cross Validation (5-Fold)	96.77
External Data (2025)	78.00

3.2 Discussion

Findings The key findings of this experiment are that the features Price Unit, Product Name, and Month play a significant role in determining the sales class of various products. An interpretation of the decision tree reveals that if the price for a product is less than or equal to some cut-off value, and has already classified as a 'light' promotional type each time,e.g., Keychains / Brochure Printing, it gets classified as a "selling" product in every case. That is to say, price still has a significant influence on consumer purchasing behavior in the product promotion industry. This not only extends Musa [1] and Putri's [7] findings about price acting as a prominent determinant in sales classification, but also adds Month, a continuously used but seldom explored temporal characteristic.

Furthermore, these results show evidence of seasonality in some products where demand appears higher in those months like t-1 before big promotions or for a national holiday. This result is also similar to that in van Steenbergen [5] on the seasonal demand for new products, although in our case the application focus lies with promotional seasonal products, which become highly at risk of remaining unused if their marketing strategy does not respond to market feedback. Therefore, in this research, we are proving that the price and product type attributes' importance as described by Leonardi [3] and Fadhila [4]; moreover, we give an insight into the temporal aspect (month of sale) as one of the predictor variables, which should not be omitted from the sales classification system.

Compared with the literature, which only takes advantage of static attributes such as price or order quantity [6], [10], up to now, the interrelations of these attributes have not been taken into account in determining the sales pattern. This research presents a more synthetic way that is richer in data as it encompasses four features at the same time, and is able to construct a model that can better capture some of the complexities in the real market. Moreover, the externally evaluated approach (2025) is crucial in terms of added value as it demonstrates the generalization capability of the model that is very limited with respect to newcomers. This is further proof that the market has dynamic and static models that should be checked time after time [14].

Despite being able to achieve a very high accuracy on the test set (99.48%), such a model generalizes not so well on external data (78%). This points to an overfit on the training data, or a failure of the model in adapting itself to real-time market trend changes. These

inaccuracies are consistent with those observed in the study of Ruliansyah [2], specifically that C4. 5-based classification models would lose accuracy without regular data updates or a mechanism by the ensemble method, such as Random Forest techniques.

In terms of system science, it can be structured that these results contribute to a data mining decision support system. The proposed classification model could be built into a management dashboard in order to visualize products according to demand level and forecast the stocks required, suggesting appropriate prices and launch times. Reasoning built based on a decision tree is tangible and not very difficult to transfer into the brains of a non-technical person due to visualization and logic. It is an instance of a knowledge-based system that integrates analytic techniques and data-driven heuristic methods, as part of modern information systems that value decision making based on information [15].

CONCLUSION

This study has substantially solved all the posed formulations for marketing of promotional product sales levels in HERA Promotion. According to the modeling parasite and the discussion, the C4.5 algorithm can categorize products as Best-Selling or Not-Selling correctly by ProductName, Price Unit, Order Quantity, and Month. High test and cross-validation performance of the model reveals that the data mining-based classification can effectively result in sales patterns that were not shown at previous times with traditional analysis.

From an empirical point of view, the results of this study verify that product price has the greatest influence on sales volume, followed by component to product type and sales period. This interpretation indicates that the market of a promotional product is not only affected by price and time as a combination instead of strictly historical sales. Therefore, two positives of the research findings are that: (i) we can see how particular products may be differentiated by their sales volume, and (ii) it is problematic why some may sell better than others within a certain period.

Theoretically, our research is a valuable addition to information systems and data mining literature, mainly in using the C4. 5 algorithm in the promotion product area that can have a seasonal or dynamic atypicality. This enrichment of the analysis approach is also evident from at least the inclusion of temporal variables (Month) in the classification model - a field that so far is still primarily shaped through static factors. Second, it generates explicit and easy-to-interpret knowledge in the form of a decision tree, which is even directly applicable/mappable on managerial support systems.

But the decrease in accuracy in out-of-sample data testing confirms that this model remains imperfect in terms of adaptive success to market evolution. This shows that having regular and actualized data would be appropriate as well as a more dynamic model. Hence, additional studies are suggested to integrate the C4. 5 algorithm with ensemble methods or add other variables (e.g., demand trend and digital promotion context) to enhance the generalizability of the model.

In general, the C4. 5-based sales classification serves not only as a technological tool but also as an imagination and research foundation to build the information system in terms of data-driven decision-making for alignment of marketing strategy and product management more effectively.

REFERENCES

[1] D. M. Musa, "Penerapan data mining untuk klasifikasi data penjualan pakan ternak menggunakan algoritma C4.5," *J. Tek. Inform. Dan Komput.*, vol. 2024. [Online]. Available: https://journal.thamrin.ac.id/index.php/jtik/article/view/1985

[2] M. J. Ruliansyah, "Penerapan metode C4.5 dalam prediksi penjualan Tim Bev 1," *J. Ilm. Manaj. Inform. Dan Komput.*, vol. 3, no. 2, 2024. [Online]. Available: https://journal.stmiki.ac.id/index.php/jimik/article/view/664

[3] M. Leonardi, "Prediksi penjualan produk rokok pada PT Indomarco Adi Prima Tbk menggunakan algoritma C4.5," *J. Paradigma*, vol. 25, no. 2, 2023. [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/paradigma/article/view/11151

[4] F. Fadhila, "Application of C4.5 algorithm to predict sales at PT. Sumber Sayur Segar," *J. Intell. Decis. Sci. Syst.*, vol. 5, 2022. [Online]. Available: https://idss.iocspublisher.org/index.php/jidss/article/view/45

[5] R. M. van Steenbergen, "Forecasting demand profiles of new products with C4.5 decision tree," *Decis. Support Syst.*, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923620301561

[6] A. K. Lalo, "Implementasi algoritma C4.5 untuk klasifikasi penjualan barang pada Swalayan Dutalia," *J. Tek. Inform. Univ. Suryakancana*, vol. 5, no. 2, 2021. [Online]. Available: https://ejournal.ust.ac.id/index.php/JTIUST/article/view/1089

[7] S. R. Putri, "Implementation of the decision tree method and C4.5 algorithm for sales classification," *InfoSains: J. Ilmu Komput. Dan Inform.*, 2024. [Online]. Available https://ejournal.seaninstitute.or.id/index.php/InfoSains/article/view/4103

[8] P. W. Sari, "Implementasi algoritma C4.5 untuk klasifikasi data insentif karyawan," *J. Komput. Dan Inform.*, 2024. [Online]. Available: https://ejournal.mediaantartika.id/index.php/jka/article/view/318

[9] S. Haafizh *et al*., "Classification of product predicates based on sales rate using C4.5 decision tree," *Int. J. Multimedia and Inform.*, 2024. [Online]. Available: https://journal.antispublisher.com/index.php/IJMI/article/download/201/155/894

[10] R. Sulastri, "Identifikasi tingkat penjualan produk herbal HWI menggunakan algoritma C4.5," *Infeb: J. Inform. Dan FEB*, vol. 3, no. 2, 2022. [Online]. Available: https://infeb.org/index.php/infeb/article/view/141

[11] A. H. Nasrullah, "Klasifikasi produk laris menggunakan decision tree C4.5," *J. Ilmu Komput. Dan Inform.*, vol. 14, no. 2, 2021. [Online]. Available: https://ejournal.lppm- unasman.ac.id/index.php/jikom/article/view/203

[12] I. Ifongki, “Data mining dengan algoritma C4.5 untuk penjualan kopi,” *J. Sains dan Inform. Ind.*, 2021. [Online]. Available: https://ejournal.lppm-unbaja.ac.id/index.php/jsii/article/view/836

[13] N. Surojudin, “Penerapan data mining dengan algoritma C4.5 untuk prediksi penjualan bahan bangunan,” *Infeb: J. Inform. Dan FEB*, 2025. [Online]. Available: https://www.infeb.org/index.php/infeb/article/view/1241

[14] A. K. Ramadhan, “Analisis dampak digitalisasi penjualan dengan algoritma C4.5,” *SMSJ: Sustain. Manag. and Sci. J.*, 2026. [Online]. Available: https://journal.independentresearchcenter.com/smsj/article/view/208

[15] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.