

Implementation Of the C4.5 Algorithm With The Backward Elimination Feature Selection For MSME Product Sales Strategy

M. Rudi Fanani^{1*}, Elvinda Bendra Agustina²

¹Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, Institut Teknologi dan Sains Nahdlatul Ulama Pekalongan

²Program Studi Fisika, Fakultas Sains dan Teknologi, Institut Teknologi dan Sains Nahdlatul Ulama Pekalongan
idurinanaf@gmail.com^{1*}

Abstract

The development of Micro, Small and Medium Enterprises (MSMEs) in Indonesia has become a major focus in advancing the economy, reducing poverty and improving community welfare. MSMEs not only play a strategic role in creating jobs and driving the local economy, but are also a source of inspiration for innovation and creativity. Kedungwuni Timur sub-district is an area in Kedungwuni sub-district, Pekalongan Regency, Central Java. The profession of the majority of Kedungwuni Timur Village residents is operates in the MSME Fashion sector. The current problem is that consumers can easily compare products. Therefore, it is necessary to carry out a sales strategy by utilizing Information Technology concepts, one of which is using Data Mining techniques. The C4.5 algorithm is a data mining algorithm that is used to predict the sales strategy for MSME products with an accuracy value of 82.78%. To increase the accuracy value, the backward elimination feature was used to produce an accuracy of 85.00%.

Keywords: C4.5 Algorithm, Forward Selection, Data Mining, Product Sales.

1. Introduction

Indonesia is entering an era of development of Micro, Small and Medium Enterprises (MSMEs) and at the same time making them important in order to reduce poverty, advance the economy, and improve community welfare. MSMEs also have a strategic role in creating jobs, stimulating the local economy, and can also be the basis for inspiring innovation and community creativity. MSMEs have become an important part of business in the Indonesian economy [1] Currently, MSMEs are in a positive trend with their numbers continuing to increase every year. This positive trend will have a good impact on the Indonesian economy. Based on data from the Ministry of Cooperatives and SMEs, the contribution of MSMEs to National GDP is 60.5%. This shows that MSMEs in Indonesia have great potential to be developed so that they can contribute even more to the economy [2]. Pekalongan Regency is an area on the north coast of the Java Sea in Central Java province, consisting of 19 sub-districts, which are further divided into 270 villages and 13 sub-districts [3]. MSMEs in the fashion sector are still the largest with 3,795 MSMEs throughout Pekalongan Regency, one of which is in Kedungwuni Timur Subdistrict where the majority of the population is engaged in fashion MSMEs [4]. Based on the results of interviews with MSME business actors, consumers currently face obstacles in easily comparing the quality, price, service and speed of a business, resulting in a decrease in demand or sales. Therefore, it is necessary to carry out an innovation or strategy when conducting business using the concept of Information Technology which aims to increase market interest and competitiveness as well as good relations with consumers, namely by using Data Mining techniques [5]. Data Mining techniques are a series of processes to find more value from a collection of data into knowledge that has not yet been known. The concept of data mining is an attempt to obtain unseen information in large amounts of data. There are many algorithms in data mining, one of which is the C4.5 algorithm [6]. The C4.5 algorithm is used to classify statistical data in determining sales strategies for MSME products [7]. The application of the backward elimination feature is also carried out to increase the accuracy of the C4.5 algorithm [8].

2. Research Methods

The method that will be used in this research uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) model. This model consists of six phases that cover the entire data mining research cycle, including: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [9].

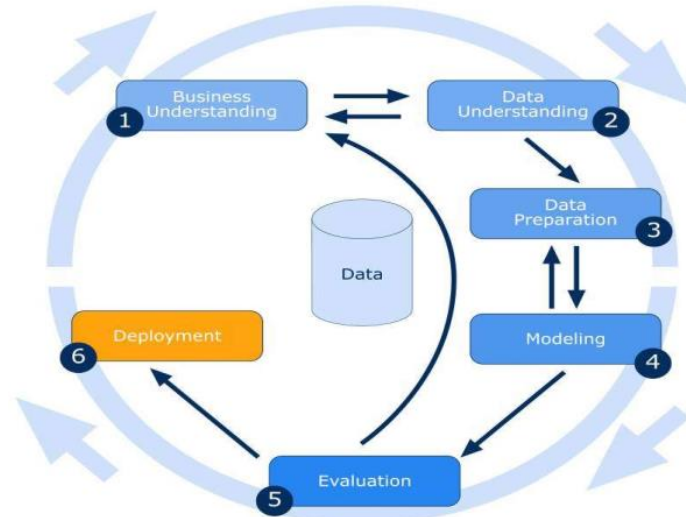


Fig. 1: CRISP-DM Methods [10]

- Business Understanding**
 This stage focuses on a thorough understanding of the business goals to be achieved through the data mining process. One way to gain business understanding is to create concept mapping which can be done based on data, information and business processes obtained by conducting interviews and confirmation from parties involved in the business being studied.
- Data Understanding**
 At the data understanding stage, analysts carry out initial exploration of the data available for the data mining process. The data obtained is based on data from concept mapping and interviews with parties who are the object of research to clarify the accuracy of the data and information.
- Data Preparation**
 The data will be input into digital form, then the collected data will be cleaned and processed so that it is ready to be used in the modeling process with the C4.5 algorithm. The results of this data preparation stage can be in the form of a report presented on a digital sheet. In this research, analysts use Microsoft Excel [11], [12].
- Modelling**
 Based on previously prepared data, analysts apply appropriate data mining techniques to develop descriptive and predictive models. This modeling stage is the core of the data mining process, where valuable insights and knowledge are discovered from the data, which will then help in making smarter and strategic decisions. In this research, the data mining and visualization tool used is Rapidminer. The C4.5 algorithm with Backward elimination feature selection is implemented [13], [14].
- Evaluation**
 Models and data mining results are evaluated in depth to measure their quality and relevance to predetermined business goals. The evaluation stage ensures that the resulting data mining results are useful, reliable, and in line with business needs, thereby ensuring successful implementation and better data-based decision making.
- Deployments**
 At this stage, the results of the data mining model and analysis are applied in a real business environment. Analysts collaborate with business teams to ensure that models or insights are implemented effectively and integrated with operational processes.

3. Result and Discussion

The results of this research are a decision tree and influencing factors to increase market interest and competitiveness as well as relationships with consumers. The data used is the result of interviews with MSME actors in Kedungwuni Timur District, Kec. Kedungwuni District, Pekalongan, Central Java.

3.1. Data preparation

In the initial data preparation, a number of datasets were collected from interviews with business actors in the form of Microsoft Excel reports.

Table 1: Dataset

Nama UMKM	Jenis Usaha	Harga Produk	Bahan Baku	Jenis Produk	Promosi Media Sosial	Pemasaran E-Commerce	Popularitas	Tingkat Penjualan
NR	Kemeja Polos	Bersaing	Biasa	Kurang Bervariasi	Ya	Ya	Mudah Dikenal	Kurang Memuaskan
NR	Celana Panjang Chinos	Standar	Biasa	Kurang Bervariasi	Ya	Tidak	Mudah Dikenal	Kurang Memuaskan
NR	Celana Pendek Cargo	Tinggi	Istimewa	Bervariasi	Ya	Ya	Mudah Dikenal	Kurang Memuaskan
NR	Celana Pendek Tartan	Bersaing	Biasa	Bervariasi	Tidak	Ya	Kurang Dikenal	Kurang Memuaskan
NR	Celana Pendek Chinos	Bersaing	Istimewa	Bervariasi	Ya	Ya	Mudah Dikenal	Memuaskan
NR	Celana Pendek Jeans	Tinggi	Biasa	Kurang Bervariasi	Tidak	Tidak	Kurang Dikenal	Memuaskan
HL	Celana Pendek Baby Terry	Standar	Biasa	Bervariasi	Ya	Ya	Kurang Dikenal	Memuaskan
HL	Celana Pendek Olahraga	Bersaing	Biasa	Bervariasi	Ya	Ya	Mudah Dikenal	Memuaskan
HL	Celana Formal	Bersaing	Biasa	Kurang Bervariasi	Ya	Ya	Mudah Dikenal	Memuaskan
HL	Celana Jogger	Tinggi	Istimewa	Kurang Bervariasi	Ya	Tidak	Kurang Dikenal	Kurang Memuaskan
BT	Kemeja Jeans	Standar	Biasa	Kurang Bervariasi	Tidak	Tidak	Kurang Dikenal	Kurang Memuaskan
BT	Celana Training	Bersaing	Istimewa	Bervariasi	Ya	Ya	Mudah Dikenal	Memuaskan
BT	Kemeja Shanghai	Tinggi	Istimewa	Bervariasi	Ya	Ya	Kurang Dikenal	Kurang Memuaskan

3.1. Testing with the Rapidminer C4.5 algorithm

This research uses rapidminer tools to test the accuracy, precision and recall values as well as the Decision tree results from the C4.5 algorithm, as well as the backward elimination features and the C4.5 algorithm.

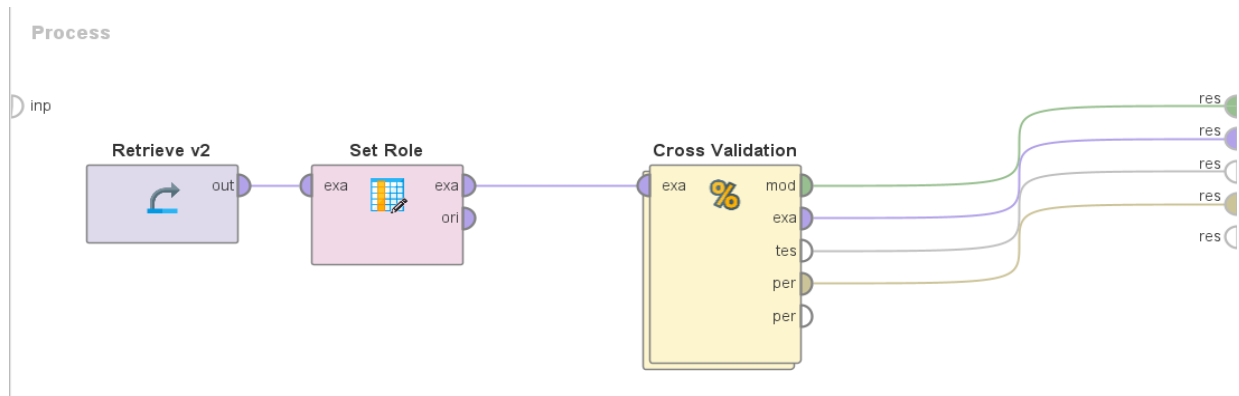


Fig. 2: C4.5 Algorithm Testing Model

This test uses Croos Validation to divide data randomly (training data and testing data) using 10 fold validation and automatic sampling type.

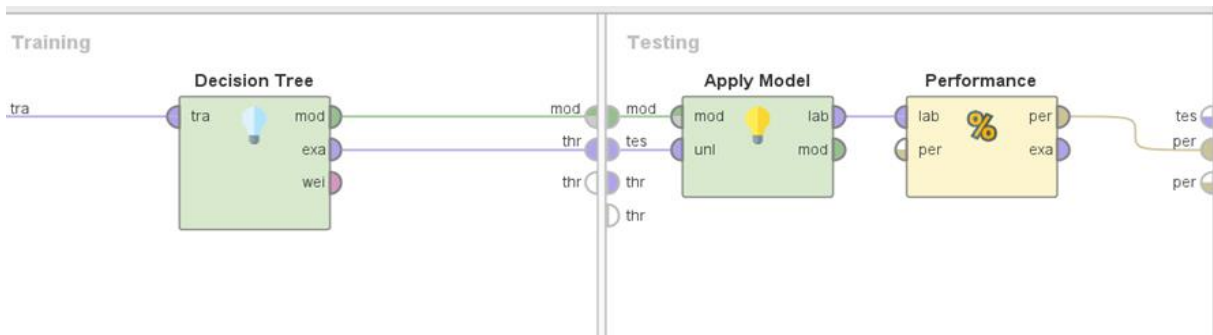


Fig. 3: C4.5 decision tree algorithm testing

In model x-validation there is a modeling process using the Decision Tree algorithm, so that it will produce accuracy and a Decision tree.



Fig. 4: Decision Tree Results

The results of the Decision tree are as follows:

1. If product prices are competitive then the results are satisfactory

2. If the product price is standard then look at e-commerce marketing if yes then the results are satisfactory
3. If the product price is standard then look at e-commerce marketing, if not look at popularity, if it is easily known then the results are satisfactory, if it is less known the results are less satisfactory
4. If the product price is high then look at the type of product if it is less varied then the results are satisfactory
5. If the price of the product is high, look at the type of product, if it varies, look at its popularity, if it is easy to recognize, the results are satisfactory, if it is less known, the results are less satisfying.

3.2. Evaluation of decision tree algorithms

The process of calculating the accuracy value in the Rapidminer Tool begins with uploading the classification result data from Data Mining. This research successfully obtained an accuracy 82.78%.

accuracy: 82.78% +/- 13.41% (micro average: 82.56%)

	true Kurang Memuaskan	true Memuaskan	class precision
pred. Kurang Memuaskan	26	9	74.29%
pred. Memuaskan	6	45	88.24%
class recall	81.25%	83.33%	

Fig. 5: Accuracy Value Results

3.3. Testing with the rapidminer C4.5 algorithm with the backward elimination selection feature

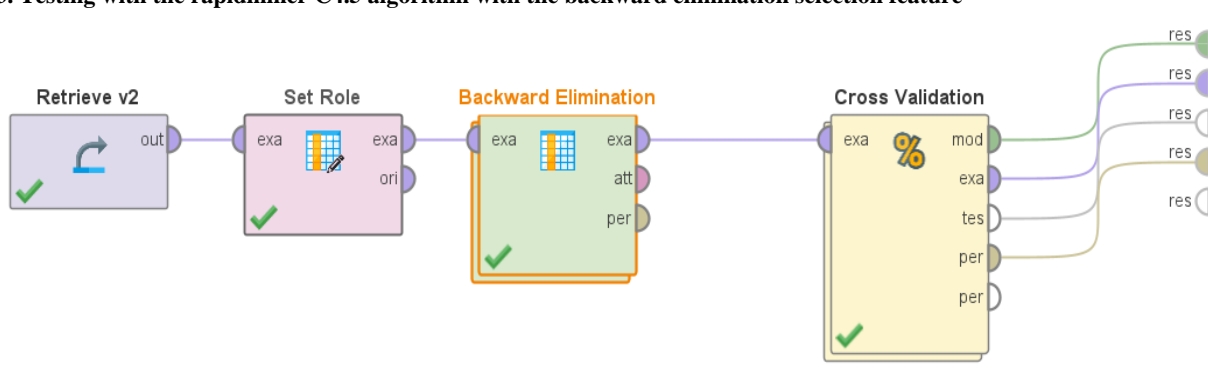


Fig. 6: Backward elimination process and Algorithm C4.5

In this process, a backward elimination feature selection is added in order to increase the accuracy of the C4.5 algorithm, namely selecting features that influence the dataset.

3.4. Evaluation of the c4.5 algorithm with backward elimination

The process of calculating accuracy values in the Rapidminer Tool begins with c4.5 algorithm data plus the backward elimination selection feature. This research succeeded in obtaining an accuracy of 85.00%.

accuracy: 85.00% +/- 16.93% (micro average: 84.88%)

	true Kurang Memuaskan	true Memuaskan	class precision
pred. Kurang Memuaskan	29	10	74.36%
pred. Memuaskan	3	44	93.62%
class recall	90.62%	81.48%	

Fig. 7: Accuracy results from adding the backward elimination feature to the C4.5 algorithm

3.5 Comparison of Accuracy Values

Based on testing using the Rapidminer tool, it can be concluded that adding the backward elimination selection feature can increase the accuracy of the C4.5 algorithm.

Table 2: Comparison of Accuracy Values

Algorithm	Accuracy Value
C4.5	82.78%
Forward Selection and C4.5	85.00%

4. Conclusion

Based on the results of the research, data was collected on MSME actors in Kedungwuni Timur Subdistrict, Pekalongan Regency, Central Java, then data was extracted and processed using data mining techniques, the algorithm used was the C4.5 algorithm which produced an accuracy value of 82.78% and the addition of the backward selection feature. elimination and the C4.5 Algorithm produce an accuracy value of 85.00%. The addition of the backward elimination selection feature has been proven to increase accuracy compared to using the C4.5 algorithm alone.

Acknowledgement

Thank you to the Government of Kedungwuni Timur Subdistrict, Pekalongan Regency, MSME actors, the research team and research assistants as well as LPPM ITSNU Pekalongan so that this research can run smoothly and be completed well.

References

- [1] M. Farras Nasrida, A. Pandahang, and D. Febrian, "Perkembangan UMKM Di Indonesia Dan Potensi Di Kota Palangka Raya," *Jurnal Manajemen Bisnis Kewirausahaan*, vol. 2, no. 1, 2023.
- [2] Kementerian Keuangan RI, "Kontribusi UMKM dalam Perekonomian Indonesia," <https://djpb.kemenkeu.go.id/>.
- [3] DPRD Kabupaten Pekalongan, "Selayang Pandang Kabupaten Pekalongan," <https://dprd-pekalongankab.go.id/halaman/selayang-pandang>.
- [4] Dinas Perindustrian dan Perdagangan (Dinperindag), "Jumlah UMKM Menurut Bidang di Kabupaten Pekalongan," https://data.pekalongankab.go.id/zh_TW/dataset/jumlah-umkm-menurut-bidang-di-kabupaten-pekalongan#.
- [5] A. M. Nazmi and A. Pambudi, "Journal of Artificial Intelligence and Engineering Applications Classification Analysis Of The Eligibility Of Recipients Of Non-Cash Food Assistance And Family Hope Programs In The City Of Sukabumi Using The Naïve Bayes Classifier Algorithm," 2024. [Online]. Available: <https://ioinformatic.org/>
- [6] M. R. Fanani and K. Fari, "Seleksi Fitur PSO untuk Klasifikasi Jenis Kekerasan dengan Algoritma C4.5," *Smart Comp*, 2022.
- [7] T. Widodo, D. Setiawan, A. Syahputri, S. Informasi, and S. Triguna Dharma, "Data Mining Menentukan Minat Konsumen Memilih Sepeda Motor Idaman Dengan Algoritma C4.5," *Jurnal Sistem Informasi TGD*, 2022, [Online]. Available: <https://ojs.trigunadharma.ac.id/index.php/jsi>
- [8] R. A. Anggraini, "Algoritma Naïve Bayes Dengan Backward Elimination Pada Dataset Breast Cancer," 2023. [Online]. Available: <http://ejurnal.ubharajaya.ac.id/index.php/JKI>
- [9] N. Lundén, E. T. Bekar, A. Skoogh, and J. Bokrantz, "Domain Knowledge in CRISP-DM: An Application Case in Manufacturing," in *IFAC-PapersOnLine*, Elsevier B.V., Jul. 2023, pp. 7603–7608. doi: 10.1016/j.ifacol.2023.10.1156.
- [10] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model," in *Procedia CIRP*, Elsevier B.V., 2019, pp. 403–408. doi: 10.1016/j.procir.2019.02.106.
- [11] Rika Nur Adiha, "APPLICATION OF DATA MINING IN DETERMINING SOCIAL ASSISTANCE RECIPIENTS WITH C4.5 ALGORITHM", *j. of artif. intell. and eng. appl.*, vol. 1, no. 2, pp. 118–127, Feb. 2022.
- [12] Adeita A. Ndraha, J. T. Hardinata, and Y. P. Purba, "Application Of The C4.5 Algorithm To Determining Student's Level Of Understanding", *j. of artif. intell. and eng. appl.*, vol. 1, no. 2, pp. 162–167, Feb. 2022.
- [13] R. Arfika and M. Safii, "Analysis of Decreased Public Awareness in the Application of Health Protocols with the C4.5 . Algorithm", *j. of artif. intell. and eng. appl.*, vol. 1, no. 1, pp. 66–70, Oct. 2021.
- [14] F. W. Sari and M. Safii, "Application of the C4.5 Algorithm in Teaching Teachers' Skills on Learning Effectiveness", *j. of artif. intell. and eng. appl.*, vol. 1, no. 1, pp. 54–59, Oct. 2021.