Analysis of Consumer Shopping Interest Level through Ads on Social Media (Facebook and Instagram) Using K-Means Clustering and C4.5 Algorithm Methods

Jepri Banjarnahor[1]\*, Jessy Putrionom Hutagalung[2], Ferdinand Jery Wilkinson Sitorus[3]

Department of Sciences and Technology [1], [2], [3]

Universitas Prima Indonesia Medan - Indonesia

jepribanjarnahor@unprimdn.ac.id[1], jessyhutagalungg@gmail.com [2], ferdinandsitorus424@gmail.com [3]

***Abstract*—** **In the growing digital era, the role of social media as a marketing platform has become very important. Facebook and Instagram, two social media giants, have become the top choices for businesses to reach consumers effectively through the advertising features they provide. However, companies often face challenges in understanding consumer responses to ads posted on these platforms. With a large volume of ads published daily, it is difficult for companies to measure the effectiveness of ads and determine the level of consumer interest efficiently. To solve this problem, using K-means clustering and C4.5 algorithm methods offers an effective solution. The K-means algorithm helps companies cluster consumers based on their responses to advertisements, while the C4.5 algorithm is used to classify consumers' future shopping interest levels based on identified patterns. By applying these two methods, companies can gain deeper insights into consumer preferences and behavior, thereby optimizing their marketing strategies. This research aims to analyze the level of consumer shopping interest through advertisements on Facebook and Instagram using K-means clustering and C4.5 algorithm methods to help companies make smarter and more effective marketing decisions.**

***Keywords :*** ***Social media, K-means clustering, C4.5 Algorithm, marketing, consumer shopping interest***

# Introduction

In the growing digital era, the role of social media as a marketing platform cannot be underestimated. Facebook and Instagram, as two social media giants, have become the center of attention for many businesses that want to reach consumers effectively. Through the advertising features provided by both platforms, businesses can promote their products or services to the right segmented target audience. Advertising is a form of communication used by individuals, companies, or organizations to promote products, services, or ideas to the wider community[1]. Social media is a digital platform that allows users to create, share, and interact with content and communicate with others. Through social media, users can participate in social networks, online communities, and discussion forums that bring together individuals with similar interests[2].

In the digital era's marketing context, companies often face challenges in understanding consumer interest and response to advertisements posted on social media, especially on Facebook and Instagram platforms. With a large volume of ads published daily, it is difficult for companies to efficiently determine how consumers receive their ads and how effective they are in driving purchase actions. In addition, the success of social media advertising often depends on how well companies can tailor their advertising messages and targeting to consumer preferences and behaviors. Therefore, companies need analytical methods to help them identify patterns in consumers' responses to advertisements to optimize their marketing strategies and improve their advertising campaigns.

To overcome these challenges, using K-means clustering and C4.5 algorithm methods is an effective solution. By using K-means clustering, which is one of the most common clustering algorithms for grouping data according to similar characteristics[3]. Companies can group the type of increase in consumer interest in buying within 1 month based on the specifications of advertisements displayed on Facebook and Instagram social media, thus allowing companies to understand consumer groups who respond to advertisements by viewing the advertisements displayed. In addition, with the C4.5 algorithm, this method is used to perform a series of classification problems in machine learning and data mining[4]. Companies can classify the level of interest in future consumer spending based on the patterns identified, allowing companies to more precisely target advertisements to potential consumers who have a high level of interest. By using these two methods together, companies can gain deeper insights into consumer responses to advertisements on the Facebook and Instagram platforms, allowing them to make smarter and more effective marketing decisions.

# METHODS

In conducting research, the method used is the R&D method. Research and development (R&D) is a systematic process involving investigation and experimentation to develop new knowledge and create innovation. R&D is carried out by companies or organizations to produce new and better products, services, or technologies or to improve existing ones. R&D activities include basic and applied research aimed at developing new solutions and increasing competitiveness in the market [5].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Month | Year | Information | Image | Price | Promotion |
| January | 2023 | 80 | 70 | 80 | 80 |
| February | 2023 | 75 | 80 | 89 | 76 |
| March | 2023 | 86 | 85 | 77 | 80 |
| April | 2023 | 75 | 80 | 92 | 87 |
| May | 2023 | 86 | 89 | 89 | 75 |
| June | 2023 | 87 | 75 | 77 | 79 |
| July | 2023 | 88 | 65 | 79 | 60 |
| August | 2023 | 70 | 89 | 80 | 90 |
| September | 2023 | 75 | 90 | 80 | 80 |
| October | 2023 | 92 | 78 | 84 | 80 |
| November | 2023 | 87 | 89 | 78 | 96 |
| December | 2023 | 70 | 87 | 87 | 80 |
| January | 2024 | 75 | 92 | 88 | 98 |
| February | 2024 | 78 | 80 | 79 | 67 |
| March | 2024 | 90 | 85 | 90 | 80 |
| April | 2024 | 80 | 80 | 92 | 86 |
| May | 2024 | 81 | 82 | 78 | 80 |



Figure 1: Stages of Research and Development (R&D)

The research and development phase of developing an analytical model to optimize social media marketing strategies involved several steps.First, we identified the opportunities and challenges associated with using Facebook and Instagram as effective advertising platforms and analyzed consumer shopping interest from ad interactions.Data from user interactions with ads was then collected.We then created a product design in the form of an analytical model using the K-means clustering method for segmentation and C4.5 algorithm for classification.To ensure the accuracy of the model, design validation was conducted, and improvements were made based on the validation results.Product testing was conducted to confirm functionality using a wider range of data, after which the model was revised to improve performance.After usage testing and final modifications, the model is ready for mass production.In short, it is an integrated analytics system that you can use every day to analyze consumer shopping interests and optimize your marketing strategies on Facebook and Instagram.
This method works by identifying a number of centroids in the data, and then allocating each data to its closest centroid [6]. The C4.5 algorithm creates a top-down decision tree, where the topmost attribute is the root, and the bottom is called the leaf [7]. By using a sample, researchers can make inferences or conclusions about the population as a whole [8].

# RESULT AND DISCUSSION

This research analyzes the level of consumer shopping interest through advertisements on Facebook and Instagram social media using the K-means clustering and C4.5 algorithm methods. Practically, the results of this study can be used by advertisers to optimize marketing strategies on Facebook and Instagram [9].

## Sampel Data

By using samples, researchers can make inferences or conclusions about the population as a whole (Susanto et al., 2024). In this study, using a sample of advertising specification data on social media Facebook and Instagram [10].

### Dataset

Datasets are used in various fields, including data science, statistics, machine learning, and scientific research, for analysis and decision making [11]. The datasets taken are specifications of advertisements displayed on Facebook and Instagram social media, while the datasets used in this study are as follows:

Table 1: Dataset

### Data Analysis

This process involves various techniques and methods applied to raw data to extract meaningful and relevant insights [12]. To determine the number of clusters, the sample data will be divided into two clusters (groups). Set two data points from the dataset as initial cluster centers.

Table 2: Initial Cluster

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| C1 | Up | 75 | 80 | 92 | 87 |
| C2 | Down | 86 | 85 | 77 | 80 |

After setting the k value and the initial cluster center, the next step is to calculate the distance between each student’s data and the cluster center:

$$C1=\sqrt{\left(80-75\right)^{2}+ \left(70-80\right)^{2}+ \left(80-92\right)^{2}+\left(80-87\right)^{2} }$$

$$C1=\sqrt{(5)^{2}+ (-10)^{2}+ (-12)^{2}+(-7)^{2 } }$$

$$C1=\sqrt{25+100+144+49 }$$

$$C1=\sqrt{318}$$

$C1=$17,83255

$$C2=\sqrt{\left(80-86\right)^{2}+ \left(70-85\right)^{2}+ \left(80-77\right)^{2}+\left(80-80\right)^{2} }$$

$$C2=\sqrt{(-6)^{2}+ (-15)^{2}+ (-3)^{2}+(0)^{2 } }$$

$$C2=\sqrt{36+255+9+0 }$$

$$C2=\sqrt{270}$$

$C2=$ 16,43168

The closest cluster distance for each record can be determined. Determine the cluster for each record and update the cluster center point. After the calculation of all the data is complete, the results can be seen in the table below.

Table 3: Iterations 1

|  |  |  |  |
| --- | --- | --- | --- |
| Data to- | C1 | C2 | Cluster |
| 1 | 17,83255 | 16,43168 | C2 |
| 2 | 11,40175 | 17,49286 | C1 |
| 3 | 20,4939 | 0 | C2 |
| 4 | 0 | 20,4939 | C1 |
| 5 | 18,84144 | 13,60147 | C2 |
| 6 | 21,40093 | 10,0995 | C2 |
| 7 | 35,9444 | 28,42534 | C2 |
| 8 | 16,09348 | 19,51922 | C1 |
| 9 | 17,11724 | 12,4499 | C2 |
| 10 | 20,14944 | 11,57584 | C2 |
| 11 | 22,40536 | 16,55295 | C2 |
| 12 | 12,16553 | 18,97367 | C1 |
| 13 | 16,76305 | 24,79919 | C1 |
| 14 | 24,04163 | 16,18641 | C2 |
| 15 | 17,4069 | 13,60147 | C2 |
| 16 | 5,09902 | 17,94436 | C1 |
| 17 | 16,88194 | 5,91608 | C2 |

Table 4: Cluster Grouping

|  |  |  |
| --- | --- | --- |
| Data to- | C1 | C2 |
| 1 | 17,83255 | 16,43168 |
| 2 | 11,40175 | 17,49286 |
| 3 | 20,4939 | 0 |
| 4 | 0 | 20,4939 |
| 5 | 18,84144 | 13,60147 |
| 6 | 21,40093 | 10,0995 |
| 7 | 35,9444 | 28,42534 |
| 8 | 16,09348 | 19,51922 |
| 9 | 17,11724 | 12,4499 |
| 10 | 20,14944 | 11,57584 |
| 11 | 22,40536 | 16,55295 |
| 12 | 12,16553 | 18,97367 |
| 13 | 16,76305 | 24,79919 |
| 14 | 24,04163 | 16,18641 |
| 15 | 17,4069 | 13,60147 |
| 16 | 5,09902 | 17,94436 |
| 17 | 16,88194 | 5,91608 |

To update the cluster center point value, the cluster center equation can be used as follows:

|  |  |
| --- | --- |
| C1.1 = $\frac{75+75+70+70+75+80}{6}$ = 74,16667C1.2 = $\frac{80+80+89+87+92+80}{6}$ = 84,66667C1.3 = $\frac{89+92+80+87+88+92}{6}$ = 88C1.4 = $\frac{76+87+90+80+98+86}{6}$ = 86,16667 | C2.1 = $\frac{80+86+86+87+88+75+92+87+78+90+81}{11}$ = 84,54545C2.2 = $\frac{70+85+89+75+65+90+78+89+80+85+82}{11}$ = 80,72727C2.3 = $\frac{80+77+89+77+79+80+84+78+79+90+78}{11}$ = 81C2.4 = $\frac{80+80+75+79+60+80+80+96+67+80+80}{11}$ = 77,90909 |

 From the above calculations, the latest cluster point center results are obtained:

Table 5: New Cluster

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| C1 | Up | 74,16667 | 84,66667 | 88 | 86,16667 |
| C2 | Down | 84,54545 | 80,72727 | 81 | 77,90909 |

After the cluster grouping is obtained in the 3rd iteration, the next calculation process is carried out using the C4.5 algorithm to determine the cluster in the following month. Yes, the explanation can be seen as follows.

Table 6: K-Means Clustering Result

|  |  |  |  |
| --- | --- | --- | --- |
| No | Month | Year | Cluster |
| 1 | January | 2023 | C2 |
| 2 | February | 2023 | C1 |
| 3 | March | 2023 | C2 |
| 4 | April | 2023 | C1 |
| 5 | May | 2023 | C2 |
| 6 | June | 2023 | C2 |
| 7 | July | 2023 | C2 |
| 8 | August | 2023 | C1 |
| 9 | September | 2023 | C1 |
| 10 | October | 2023 | C2 |
| 11 | November | 2023 | C1 |
| 12 | December | 2023 | C1 |
| 13 | January | 2024 | C1 |
| 14 | February | 2024 | C2 |
| 15 | March | 2024 | C2 |
| 16 | April | 2024 | C1 |
| 17 | May | 2024 | C2 |
| 18 | June | 2024 | ? |

From the table above, what is taken to become training data for data testing, that is, only the attribute data, as for the attribute data taken, A1, A2, A3, A4. Where the training data has 17 records, it will be classified with the testing data, which has 1 record. For example, the calculation can be explained as follows:

EntropyA1 (>=80) = (-($\frac{2}{10}$) x$ Log\_{2}(\frac{2}{10})$ + ($\frac{8}{10}$) x$ Log\_{2}(\frac{8}{10})) $= 0,7219

EntropyA1 (<80) = (-($\frac{6}{7}$) x$ Log\_{2}(\frac{6}{7})$ + ($\frac{1}{7}$) x$ Log\_{2}(\frac{1}{7})) $= 0,5917

GainA1 = 0,997 – $((\frac{10}{17})$ x 0,7219) - $((\frac{7}{17})$ x 0,5917) = 0,3292

And so on.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Node | Attributes | Value | Case | C1 | C2 | Entropy | Gain |
| 1 | Total |  | 17 | 8 | 9 | 0,997502546 |  |
|  | A1 | >=80 | 10 | 2 | 8 | 0,721928095 | 0,329209 |
|  |  | <80 | 7 | 6 | 1 | 0,591672779 |  |
|  | A2 | >=80 | 13 | 8 | 5 | 0,961236605 | 0,262439 |
|  |  | <80 | 4 | 0 | 4 | 0 |  |
|  | A3 | >=80 | 11 | 7 | 4 | 0,945660305 | 0,156185 |
|  |  | <80 | 6 | 1 | 5 | 0,650022422 |  |
|  | A4 | >=80 | 12 | 7 | 5 | 0,979868757 | 0,093499 |
|  |  | <80 | 5 | 1 | 4 | 0,721928095 |  |

Table 7 Node

From the node table above, which has gone through several stages in the formation of nodes, a decision tree is formed, which can be seen below:



Figure 2 Decision Tree

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No | Month | Year | Information | Image | Price | Promotion | Interest  |
| 18 | June | 2024 | 85 | 87 | 70 | 89 | C2 |

Table 8 C4.5 Classification

So it can be explained in the 18th data the results of the classification of the C4.5 algorithm, then for data 18 enter the cluster "C2".

### System Implementation

System implementation refers to applying or implementing a new system or changes to an existing system in a production or operational environment. This involves hardware setup, software installation, network configuration, and user testing and training [13]. The explanation can be seen as follows:



Figure 3: Weka K-Means Clustering Classification Results

The Weka analysis above uses the K-means algorithm to group the data into two clusters. The model went through four iterations with a sum of squared errors value of 3.7644334066637334. The starting point of clustering was initialized randomly. The final cluster shows that:

1. Cluster 0 has centroids A1: 78.7143, A2: 71.4286, A3: 65.5714, and A4: 78.4286.

2. Cluster 1 has centroids A1: 83.1429, A2: 79.2857, A3: 74.7143, and A4: 82.4286.

A total of 41% of the data is included in Cluster 0 and 59% in Cluster 1. This model takes 0 seconds.



Figure 3 Final Result of Weka Clustering

The figure above explains the results of the comparison between manual calculations and calculations using the Weka application, there are 'Y' errors totaling '4', while valid 'T' totals '13', more valid results than error results, so the results of this calculation can be used to help companies adjust their ad targeting more precisely.



Figure 4: Classify the C4.5 algorithm Weka Result

The Weka analysis results above show the performance of the classification model using the stratified cross-validation method. Out of 18 instances, the model classified 83.333% (15 instances) correctly and 16.6667% (3 instances) incorrectly. The kappa statistic obtained was 0.6582. The resulting error values are as follows: mean absolute error (0.1667), root mean squared error (0.383), relative absolute error (33.5294%), and root relative squared error (76.61%). The accuracy details show that for class C2, the precision value is 0.818, recall 0.900, and F-measure 0.857. For class C1, the precision value is 0.857, recall is 0.750, and F-measure is 0.800. The area under the ROC curve is 0.913 for both classes. The confusion matrix shows that 9 instances of class C2 and 6 instances of class C1 are correctly classified, while 2 instances of class C1 are classified as C2.



Table 5: Final Result Classify Weka

The figure above explains the results of the comparison between manual calculation and calculation using the Weka application. In the 18th data, the results of manual classification and Weka have similarities for the results, namely in the 18th data into the "C2" cluster, so that the results of this calculation can be used in helping companies adjust their ad targeting more precisely.

##### Acknowledgment

Thank you to the Prima Indonesia University campus for providing the opportunity to complete this research and create this paper.

##### References

1. Armanto, R., & Gunarto, M. (2022). Analysis of the impact of social media on housing sales: An empirical study of Facebook and Instagram ad usage. Journal of Business, Management and Economics, 3(1), e-ISSN 2745-7281.
2. Hartawan, E., Liu, D., Handoko, M. R., Evan, G., & Widjojo, H. (2021). The effect of advertising on Instagram social media on public buying interest in e-commerce. Sam Ratulangi University Scientific Journal of Business Management and Innovation, 8(1), 217-228.
3. Sembiring, C. S. D., Hanum, L., & Tamba, S. P. (2022). Application of data mining using the K-Means algorithm to determine thesis titles and research journals (Case study of FTIK UNPRI). JUSIKOM PRIMA (Journal of Information Systems and Computer Science Prima), 5(2).
4. Anestiviya, V., & Pasaribu, A. F. O. (2021). Pattern Analysis Using the C4.5 algorithm Method for Student Major Specialization Based on Curriculum (CASE STUDY: SMAN 1 NATAR). Journal of Information Technology and Systems (JTSI) Vol. 2, No. 1, March 2021, 80 - 85.
5. Fransiscaa, S., & Putri, R. N. (2019). Utilization of RFID technology for school inventory management with (R&D) method (Case study: SMK Global Pekanbaru). Journal of Computer and Information Technology Application Students, 1(1), 72-75.
6. Mawarni, Q. I., & Budi, E. S. (2022). Implementation of K-Means clustering algorithm in student discipline assessment. Journal of Computer and Information Systems (JSON), 3(4), 522-528.
7. Anestiviya, V., & Pasaribu, A. F. O. (2021). Pattern Analysis Using the C4.5 algorithm Method for Student Major Specialization Based on Curriculum (CASE STUDY: SMAN 1 NATAR). Journal of Information Technology and Systems (JTSI) Vol. 2, No. 1, March 2021, 80 - 85.
8. Susanto, P. C., Arini, D. U., Yuntina, L., Soehaditama, J. P., & Nuraeni. (2024). Quantitative research concepts: Population, sample, and data analysis (A literature review). Journal of Computer Science, 3(1).
9. Utomo, D. P., & Mesran. (2020). Comparative analysis of data mining classification methods and attribute reduction on heart disease datasets. JOURNAL OF BUDIDARMA INFORMATICS MEDIA, 4(2), 437-444.
10. Wiratama, M. A., & Pradnya, W. M. (2022). Optimization of data mining algorithms using backward elimination for diabetes disease classification. National Journal of Informatics Engineering Education: JANAPATI, 11(1).
11. Pramudito, D. K. (2022). Data mining implementation on Java North Coast weather forecast dataset using C4.5 algorithm. Journal of Technology Pelita Bangsa, 13(3).
12. Pambudi, A., Abidin, Z., & Permata. (2023). Application of CRISP-DM using MLR K-Fold on PT Telkom Indonesia (Persero) Tbk (TLKM) stock data (Case study: Indonesia Stock Exchange 2015-2022). JDMSI, 4(1), 1-14.
13. Rahmadi, M., Kaurie, F., & Susanti, T. (2020). Test the accuracy of postoperative patient dataset using Naïve Bayes algorithm using Weka tools. JURIKOM (Journal of Computer Research), 7(1).