Informatics and Software Engineering
p-ISSN/e-ISSN: 2988-2818/2988-2222
Homepage: https://sanscientific.com/journal/index.php/ise
2(1) 31-41 (2024)
https://doi.org/10.58777/ise.v2i1.239

Research Article

Analysis of Land Plot Sales Using the C4.5 Algorithm in Property and Housing Sector

Rizky Alfi Syahrin
Sistem Informasi, Universitas Bina Sarana Informatika

Received: June 6, 2024; Accepted: June 13, 2024

Abstract
This research examines the application of the C4.5 algorithm in analyzing land plot sales data at PT Piliruma Rosa Land. Using data from two main projects, namely Harmoni Farm House and Nuansa Alam Agroewisata and Luxury, this study aims to build a decision tree model that is effective in predicting consumer interest. The analysis was conducted by evaluating the contribution of various sales attributes, specifically Sold Area and Number of Sold Lots, which proved to have a significant impact on the classification of buyer interest. The results of the model showed an accuracy rate of 91.18%, signifying the effectiveness of the C4.5 algorithm in assisting strategic decision-making in the property sector. The findings provide important insights for real estate developers in optimizing marketing and sales strategies while offering a reliable analytical method for similar sectors.

Keywords: Data Mining, Plot Land Sales, C4.5 Algorithm, Decision Tree, Consumer Interest Prediction

Abstrak
Penelitian ini mengkaji penerapan algoritma C4.5 dalam menganalisis data penjualan tanah kavling pada PT Piliruma Rosa Land. Melalui data dari dua proyek utama, yaitu Harmoni Farm House dan Nuansa Alam Agroewisata dan Luxury, penelitian ini bertujuan untuk membangun model pohon keputusan yang efektif dalam memprediksi minat konsumen. Analisis dilakukan dengan mengevaluasi kontribusi dari berbagai atribut penjualan, khususnya luas terjual dan jumlah kavling terjual, yang terbukti memberikan dampak signifikan terhadap klasifikasi minat pembeli. Hasil dari model ini menunjukkan tingkat akurasi sebesar 91.18%, menandakan efektivitas algoritma C4.5 dalam membantu pengambilan keputusan strategis di sektor properti. Temuan ini memberikan wawasan penting bagi pengembang real estat dalam mengoptimalkan strategi pemasaran dan penjualan, sekaligus menawarkan metode analitis yang bisa diandalkan untuk sektor serupa.

Kata Kunci: Data Mining, Penjualan Tanah Kavling, Algoritma C4.5, Pohon Keputusan, Prediksi Minat Konsumen

How to cite: Syahrin, R. A. Analysis of Land Plot Sales Using the C4.5 Algorithm in Property and Housing Sector. *Informatics and Software Engineering*, *2*(2). Retrieved from https://sanscientific.com/journal/index.php/ise/article/view/239

Corresponding author: Rizky Alfi Syahrin (Rizky.alfi811@gmail.com)

1. Introduction

In the property industry, understanding the factors that influence consumer interest is crucial to determining effective marketing strategies. PT Piliruma Rosa Land faces the challenge of managing and analyzing land plot sales data from its two main projects, Harmoni Farm House and Nuansa Alam Agroewisata and Luxury. Data mining algorithms, specifically C4.5, can provide deep insights into sales patterns and assist in strategic decision-making.

Previous research has used the C4.5 algorithm for various classification applications, including in determining marketing eligibility and predicting workforce placement. For example, research by (Fitriani et al., 2020) used the C4.5 algorithm for marketing placement classification and produced an accuracy rate of 91.10%. In addition, research by Susanti and Jefi (2018) using data mining to support decisions on determining contract employees to become permanent employees shows that data mining techniques are effective in making employee decisions. Yuni and Putri (2023) also found that the use of the C4.5 algorithm in predicting the amount of palm oil production can provide accurate results with a high level of accuracy.

Furthermore, Dewi et al. (2020) showed that the C4.5 algorithm can effectively classify sales data by producing accurate decision trees. Research by Azwanti (2018) also stated that this algorithm is very effective in classifying product sales data, which is relevant to the context of this research. In addition, Rohman and Rufiyanto (2019) emphasized the importance of evaluating classification models using cross-validation to ensure optimal model performance in various data mining applications.

The use of data mining to analyze land plot sales data is not only limited to predicting consumer interest, but this research also includes price analysis and land plot sales trends (Anggraini et al., 2018). Research by Han et al. (2011) has shown that the sales prediction model built using the C4.5 algorithm method can help companies identify the main factors that influence consumer purchasing decisions. This research also emphasizes that attributes such as the area sold and the number of lots sold are significant indicators in determining consumer interest. The prediction model generated from this research has high accuracy, providing reliability in using historical data to make strategic decisions in marketing and sales (Han et al., 2011; Jin et al., 2020).

The uniqueness of the previous research is in the application of the C4.5 algorithm method to analyze land plot sales, which needs to be explored more in the literature. This research aims to identify significant attributes that can influence consumer interest and build a decision tree model to predict consumer interest based on historical sales data. Thus, this research is expected to contribute to the development of a more effective marketing strategy at PT Piliruma Rosa Land. The purpose of this research is to build a decision tree model using the C4.5 algorithm that can predict consumer interest in land plot sales at PT Piliruma Rosa Land. This model is expected to help companies make better strategic decisions regarding marketing and sales of land plots (Puspita et al., 2022).

Data mining is the process of extracting meaningful information from large data sets. This process involves using statistical, mathematical, artificial intelligence, and machine learning techniques to find significant patterns in data (Han et al., 2011). Data mining has been used in various applications such as sales analysis, consumer behavior prediction, and inventory management (Anggraini et al., 2018). In the context of the property industry, data mining can be used to identify factors that influence consumer interest and trends in land plot sales. This research shows that data mining techniques, particularly the C4.5 algorithm, enable companies to make better strategic decisions based on in-depth data analysis (Yuni & Putri, 2023). Kirkpatrick (2019) mentioned that the data mining process consists of a series of

steps, including data cleaning, data integration, data selection, data transformation, pattern evaluation, and visualization. These steps are interrelated and often require interaction with users to generate usable knowledge. The stages of the data mining process are illustrated in Figure 1.

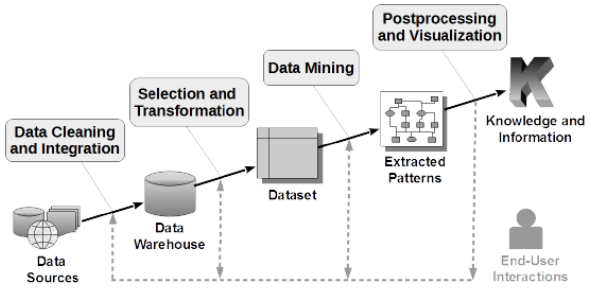

Figure 1. Illustration of Data Mining Stages

The C4.5 algorithm is one of the most well-known and widely used algorithms in data mining to form decision trees. J. Ross Quinlan introduced it as an extension of the ID3 algorithm. C4.5 is used for classification and prediction by transforming data into a decision tree consisting of nodes and leaves. Each node represents a data attribute, while the leaves represent decision classes or categories (Dewi et al., 2020). The process of forming a decision tree using the C4.5 algorithm involves several main steps, namely selecting the most informative attribute as the root of the tree, creating a branch for each value of the attribute, dividing the dataset based on the attribute value, and repeating this process for each branch until all cases in the branch have the same class. C4.5 works by calculating entropy and gain values for each attribute. Entropy measures the level of uncertainty in the data, while gain measures the decrease in uncertainty by dividing the dataset based on a particular attribute. The attribute with the highest gain value is selected as the root of the decision tree (Rohman & Rufiyanto, 2019). In recent research, the use of the C4.5 algorithm has proven effective in a variety of applications, including in sales analysis and consumer behavior prediction (Fadhila & Hasugian, 2022; Muttaqien et al., 2021).

Classification is one of the main techniques in data mining used to separate data into predefined classes. Decision trees are a very popular classification model due to their ability to generate rules that are easy to understand and interpret. In this context, decision trees are used to predict consumer interest based on certain attributes such as area sold, number of lots sold, price per m², and sales projects (Fitriani et al., 2020). Research by Witten and Frank (2002) shows that a decision tree model built with the C4.5 algorithm can help companies identify important attributes that affect consumer interest. This model allows companies to focus their marketing strategies on the most significant attributes. The use of the C4.5 algorithm in this study proved effective in the classification and prediction of consumer interest, making it easier for companies to understand consumer behavior and improve marketing strategies (Quinlan, 2014; Witten & Frank, 2002).

Evaluation of classification models is done using several evaluation metrics such as accuracy, precision, recall, and confusion matrix. Accuracy measures the percentage of correct predictions, while precision and recall measure the quality of predictions for each class. The confusion matrix is a table that shows the number of correct and incorrect predictions for each class. Cross-validation techniques are often used to measure overall model performance by dividing the data into subsets and performing validation repeatedly (Puspita et al., 2022). Good evaluation is essential to ensure that the resulting model is not only accurate but also reliable under various conditions. Research by Dewi et al. (2020) emphasized the importance of using cross-validation to avoid overfitting and ensure that the model has good generalizability. The Random Forest technique introduced by Jin et al. (2020) also showed that the

combination of multiple decision trees can significantly improve model accuracy. By integrating these techniques, this research aims to build a reliable decision tree model to predict consumer interest in land plot sales at PT Piliruma Rosa Land. The resulting model is expected to provide valuable insights for companies in developing more effective and efficient marketing strategies.

2. Methods

This research uses a quantitative approach with survey and experimental method to collect primary and secondary data. The stages of the research method are detailed as follows:

2.1. Data Collection

Primary data was collected through direct interviews with PT Piliruma Rosa Land staff to gather detailed information about the sales process, factors that influence consumer interest, and specific attributes of the land plots. The interviews included questions on sales strategies, feedback from consumers, and challenges faced in the market. In addition, the researcher observed the sales process first-hand at the company office and project site to gain first-hand insights into consumer behavior and sales tactics. Observations focused on consumer interactions, sales presentations, and the decision-making process. Surveys were also distributed to consumers who had shown interest or made a purchase to collect data regarding their preferences, satisfaction levels, and factors influencing their purchasing decisions. The survey included both multiple-choice and open-ended questions to capture a range of consumer insights.

Secondary data was collected from PT Piliruma Rosa Land's sales records for 2023/2024. These records include information on the number of lots sold, area sold, price per m², and consumer interest.

2.2. Data Processing and Analysis

The collected primary and secondary data were integrated into a unified dataset for comprehensive analysis. The datasets were cleaned to remove inconsistencies, missing values or irrelevant information that could affect the accuracy of the analysis. The cleaned data is then converted into a format suitable for analysis using the C4.5 algorithm. This involved categorizing variables such as Area Sold (large, normal, small), Number of Lots Sold (many, few), and Price per m² (expensive, cheap).

The C4.5 algorithm is applied to calculate the entropy and gain for each variable. Entropy measures the level of uncertainty in the data, while gain measures the reduction of entropy by splitting the dataset based on certain attributes. The attribute with the highest gain value is selected as the node to form the decision tree. This process is repeated in an iterative manner, splitting the data at each node until all branches produce a uniform class (high interest or not). The decision tree model is validated using cross-validation techniques to ensure its accuracy and reliability in predicting consumer interest.

The variables used to predict consumer interest in the sale of land plots at PT Piliruma Rosa Land include Project (project name, either Harmoni Farm House or Nuansa Alam Agrotourism and Luxury), Year (month and year of sale, namely 2023 and 2024), Number of Plots sold (categorized into many or few), Area sold (divided into large, normal, or small area categories), Price per m² (classified as expensive or cheap), and Interest (divided into two categories, namely high or not high). This data is processed and analyzed to determine sales patterns and factors that influence consumer interest.
Data analysis is a process carried out to collect and process data that has been collected. The data used in this research consists of primary data and secondary data. Primary data is data that has a current nature and to obtain it, researchers must go directly to the field for interviews, observations, and discussions with PT Piliruma Rosa Land staff (Dewi et al., 2020). Primary data in this study includes interviews with staff, direct observation, and consumer surveys. Secondary data was collected from PT Piliruma Rosa Land's sales records for the year 2023/2024. These records include information on the number of lots sold, area sold, price per m², and consumer interest.

The variables used to predict consumer interest in land plot sales at PT Piliruma Rosa Land include several important aspects. First, Project, which is the name of the project which includes Harmoni Farm House and Nuansa Alam Agroewisata and Luxury. Second, Year, which refers to the month and year of sales, specifically for 2023 and 2024. Third, Number of Sold Lots, which is categorized into many or few.

Fourth, Sold Area, which is divided into large, normal, or small categories. Fifth, Price per m², which is classified as expensive or cheap. Finally, Interest, which is divided into two categories, namely high or not high. Using these variables, analysis can be conducted to understand the patterns and factors that influence consumer interest in land plot sales. The data used has passed the processing stage and is in accordance with the type of data class in this study can be seen in Table 1.

Table 1. Plot Sales Data 2023/2024

No	Project	Year 2023/2024	Number of Lots Sold	Sold Area	Price per m²	Interests
1	Harmoni Farm House	January	Many	Large	Expensive	High
2	Harmoni Farm House	February	Many	Large	Expensive	High
3	Harmoni Farm House	March	Many	Normal	Expensive	Not High
4	Harmoni Farm House	April	Little bit	Small	Expensive	Not High
5	Harmoni Farm House	Mei	Little bit	Small	Expensive	Not High
6	Harmoni Farm House	June	Many	Normal	Expensive	High
7	Harmoni Farm House	July	Many	Large	Expensive	High
8	Harmoni Farm House	August	Many	Large	Expensive	High
9	Harmoni Farm House	September	Many	Large	Expensive	High
10	Harmoni Farm House	October	Many	Large	Expensive	High
11	Harmoni Farm House	November	Many	Large	Expensive	High
12	Harmoni Farm House	December	Little bit	Small	Expensive	Not High
13	Harmoni Farm House	January	Many	Large	Expensive	High
14	Harmoni Farm House	February	Many	Large	Expensive	High
15	Harmoni Farm House	March	Many	Normal	Expensive	Not High
16	Harmoni Farm House	April	Little bit	Normal	Expensive	Not High
17	Harmoni Farm House	Mei	Little bit	Small	Expensive	Not High
18	Nuansa alam Agroewisata dan Luxury	January	Little bit	Normal	Cheap	Not High
19	Nuansa alam Agroewisata dan Luxury	February	Little bit	Small	Cheap	Not High
20	Nuansa alam Agroewisata dan Luxury	March	Little bit	Small	Cheap	Not High
21	Nuansa alam Agroewisata dan Luxury	April	Little bit	Normal	Cheap	Not High
22	Nuansa alam Agroewisata dan Luxury	Mei	Many	Normal	Cheap	High
23	Nuansa alam Agroewisata dan Luxury	June	Many	Normal	Cheap	Not High
24	Nuansa alam Agroewisata dan Luxury	July	Little bit	Small	Cheap	Not High
25	Nuansa alam Agroewisata dan Luxury	August	Little bit	Small	Cheap	Not High
26	Nuansa alam Agroewisata dan Luxury	September	Little bit	Small	Cheap	Not High
27	Nuansa alam Agroewisata dan Luxury	October	Little bit	Normal	Cheap	Not High
28	Nuansa alam Agroewisata dan Luxury	November	Many	Normal	Cheap	Not High
29	Nuansa alam Agroewisata dan Luxury	December	Many	Normal	Cheap	High
30	Nuansa alam Agroewisata dan Luxury	January	Little bit	Small	Cheap	Not High
31	Nuansa alam Agroewisata dan Luxury	February	Little bit	Normal	Cheap	Not High
32	Nuansa alam Agroewisata dan Luxury	March	Little bit	Normal	Cheap	Not High
33	Nuansa alam Agroewisata dan Luxury	April	Little bit	Normal	Cheap	Not High
34	Nuansa alam Agroewisata dan Luxury	Mei	Many	Large	Cheap	High

Data analysis in Table 1 includes entropy and gain calculations for each variable to determine the most informative attributes in forming a decision tree using the C4.5 algorithm. C4.5 is one of the algorithms used to build decision trees. This algorithm is a development of the ID3 (Iterative Dichotomizer 3)

algorithm developed by J. Ross Quinlan, with the ability to handle missing values, continuous data, and pruning to reduce overfitting. The C4.5 algorithm can handle missing values by filling in the missing values using certain methods, such as mean/mode imputation (Shurrab & Duwairi, 2021), and is effective in managing continuous data by dividing it into smaller intervals for use in decision trees (Cherfi et al., 2020). In addition, this algorithm uses a pruning method to remove branches of the decision tree that do not contribute significantly to the accuracy of the model, thereby reducing the risk of overfitting (Mijwil & Abttan, 2021). In general, the C4.5 algorithm for building a decision tree is as follows: Select an attribute as the root, create branches for each value, divide the cases into branches, and repeat the process for each branch until all cases on the branch have the same class. The C4.5 algorithm continues to be developed to improve its efficiency in handling complex and large data (Liu et al., 2019).

The study selects an attribute as the root in the decision tree; the C4.5 algorithm uses the highest gain value of the existing attributes. The formula for calculating gain is as follows:

$$Gain(S,A) = Entropy(S) - i = 1\sum n(|\ S\ ||\ Si\ |\ * Entropy(Si)) \quad (1)$$

In the context of this research, some notations are used to facilitate understanding of the concepts and calculations performed. For example, *S* refers to the set of cases, while *A* is the attribute being analyzed. The number of partitions of attribute *A* is expressed by *n*, and */Si/* indicates the number of cases in the i-th partition. Meanwhile, */S/* represents the number of cases in the set *S*. The calculation of the entropy value, which is the basis of this analysis, can be seen in the following equation.

$$Entropy(S) = -i = 1\sum npilog2pi \quad (2)$$

Furthermore, to clarify the calculation, some additional notations are used. *S* denotes the set of cases, while the proportion of *Si* to *S* is expressed as *pi*, which indicates the relative share of *Si* in the set *S*. In addition, the number of partitions of *Si* is represented by *n*, which denotes the number of categories or groups in the set. In general, the C4.5 algorithm for building a decision tree starts by selecting an attribute as the root. Then, branches are created for each value of the attribute. After that, the cases are divided within each branch. This process is repeated for each branch until all cases on the branch have the same class.

4. Results

Analysis using the C4.5 algorithm shows that the Sold Area attribute is the most significant attribute in predicting consumer interest in land plot sales. The high gain value for this attribute indicates this. The following is the calculation of the gain for the Sold Area attribute:

$$Entropy(S) = -(3413log23413) - (3421log23421) = 0.9597 \quad (3)$$

For the Sold Area attribute:
Large:

$$Entropy(Large) = -(1010log21010) - (100log2100) = 0 \quad (4)$$

Normal:

$$Entropy(Normal) = -(143log2143) - (1411log21411) = 0.7496 \quad (5)$$

Small:

$$Entropy(Small) = -(100log2100) - (1010log21010) = 0 \quad (6)$$

Gain for the Sold Area attribute:

$$Gain(S, Sold\ Area\ ) = 0.9597 - (3410 \times 0 + 3414 \times 0.7496 + 3410 \times 0) = 0.9597 - 0.3087 = 0.651 \quad (7)$$

The Number of Lots Sold attribute also contributes significantly to this decision tree model. The gain calculation for the Number of Lots Sold attribute is as follows:

Many:

$$Entropy(Many) = -(1713log21713) - (174log2174) = 0.7871 \quad (8)$$

Little bit:

$$Entropy(Little\ bit) = -(170log2170) - (1717log21717) = 0 \quad (9)$$

Gain for the Number of Lots Sold attribute:

$$Gain(S, Number\ of\ Lots\ Sold) = 0.9597 - (3417 \times 0.7871 + 3417 \times 0) = 0.9597 - 0.3936 = 0.566 \quad (10)$$

The following is a calculation table for nodes in the decision tree, with the Sold Area attribute as the root node and the Number of Sold Lots as the next node.

Table 2. Root Node Calculation

		Total (S)	High (Si)	Not High (Si)	Entropy	Gain
Total		34	13	21	0.959686894	
Project						0.134830576
	Harmoni	17	10	7	0.977417818	
	Nuansa & Luxury	17	3	14	0.672294817	
Number of Lots Sold						0.566123601
	Many	17	13	4	0.787126586	
	Little	17	0	17	0	
Sold Area						**0.651030023**
	Large	10	10	0	0	
	Normal	14	3	11	0.749595257	
	Small	10	0	10	0	
Price per m²						0.134830576
	Expensive	17	10	7	0.977417818	
	Cheap	17	3	14	0.672294817	

The root node is the starting point of the decision tree. In this study, the Sold Area attribute has the highest gain value of 0.651030023, so it was chosen as the root node. The table below shows the entropy and gain calculations for the Sold Area attribute. In table 2 above, it can be seen that the Area Sold with the Large category has an entropy of 0 because all data in this category has high interest. The same goes for the Small category, which has a non-high interest with an entropy of 0. While the Normal category has an entropy of 0.749595, showing the variation between high and non-high interest. The highest gain value on this attribute makes it ideal as the root node.

Table 3. Calculation of Node 1.1

		Total (S)	High (Si)	Not High (Si)	Entropy	Gain
Total		14	3	11	0.749595257	
Project						0.015358808
	Harmoni	3	1	2	0.918295834	
	Nuansa & Luxury	11	2	9	0.684038436	
Number of Sold Lots						**0.321023829**
	Many	6	3	3	1	
	Few	8	0	8	0	
Price per m²						0.098924879
	Expensive	3	1	3	0.528320834	
	Cheap	11	2	9	0.684038436	

After determining the root node, the calculation continues on the next node with the Number of Sold Lots attribute. The following table shows the entropy and gain calculations for node 1.1. In Table 3 above, the Number of Sold Lots attribute with the Many category has entropy 1 because the data in this category is equally divided between high and not high interest. At the same time, the Few category has an entropy of 0 because all data in this category has a non-high interest. The gain value for this attribute is 0.321024, making it significant to use in the next node.

Table 4. Calculation of Node 1.2

		Total (S)	High (Si)	Not High (Si)	Entropy	Gain
Total		6	3	3	1	0
Project						
	Harmoni	2	1	1	1	
	Nuansa & Luxury	4	2	2	1	
Price per m²						0
	Expensive	2	1	1	1	
	Cheap	4	2	2	1	

The next node analyzed is the project. Table 4 shows the calculation of entropy and gain for node 1.2. The Project attribute with the categories Harmoni Farm House and Nuansa Alam Agroewisata and Luxury both have an entropy of 1 because the data in these two categories are equally divided between high and not high interest. The gain value for this attribute is 0, indicating that this attribute does not provide significant additional information for further separation on the decision tree. Data processing in this study was carried out using RapidMiner software, which is one of the powerful and easy-to-use data analysis tools for building predictive models. The resulting decision tree structure can be seen in Figure 2 below:

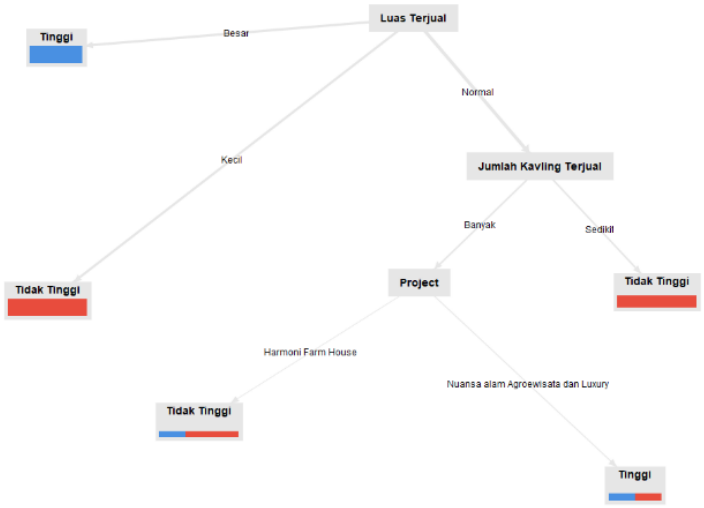

Figure 2. Decision tree result

Figure 2 shows the final decision tree pattern after calculating and testing the data on each attribute with the C4.5 algorithm. This figure provides a more detailed visualization of how each attribute and its values are used to predict consumer interest.

Tree

```
Luas Terjual = Besar: Tinggi {Tinggi=10, Tidak Tinggi=0}
Luas Terjual = Kecil: Tidak Tinggi {Tinggi=0, Tidak Tinggi=10}
Luas Terjual = Normal
|   Jumlah Kavling Terjual = Banyak
|   |   Project = Harmoni Farm House: Tidak Tinggi {Tinggi=1, Tidak Tinggi=2}
|   |   Project = Nuansa alam Agroewisata dan Luxury: Tinggi {Tinggi=2, Tidak Tinggi=2}
|   Jumlah Kavling Terjual = Sedikit: Tidak Tinggi {Tinggi=0, Tidak Tinggi=7}
```

Figure 3. Decision tree description

The text visualization in the second Figure shows how each attribute and its values are used to predict consumer interest. The Sold Area attribute is the main factor that determines interest, followed by the Number of Sold Lots and Projects. The evaluation results show that the resulting model has an accuracy rate of 91.18%. The confusion matrix shows that the predictions for the High and Not High classes are quite accurate, with high precision and recall values. The confusion matrix can be seen in Figure 4:

accuracy: 91.18%

	true Tinggi	true Tidak Tinggi	class precision
pred. Tinggi	12	2	85.71%
pred. Tidak Tinggi	1	19	95.00%
class recall	92.31%	90.48%	

Figure 4. Accuracy Value of C4.5 Algorithm

This confusion matrix shows that out of a total of 34 data, 12 High-class data were correctly predicted, while 2 Non-high-class data were incorrectly predicted as High. Conversely, 1 High-class data was incorrectly predicted as Not High, and 19 Non-high-class data were correctly predicted. The precision value for the High-class prediction is 85.71%, which indicates that out of all High predictions, 85.71% are actually data with high interest. The precision value for the Not High class is 95.00%, which indicates that out of all Not High predictions, 95.00% are really data with not high interest.

The recall value for the High class is 92.31%, which means that of all the data that actually has high interest, 92.31% was correctly predicted as High. The recall value for the Not High class is 90.48%, which means that out of all the data that actually had a not high interest, 90.48% were correctly predicted as Not High.

PerformanceVector

```
PerformanceVector:
accuracy: 91.18%
ConfusionMatrix:
True:   Tinggi  Tidak Tinggi
Tinggi: 12      2
Tidak Tinggi:   1       19
```

Figure 5. Performance Vector Value

The model formed has been tested for accuracy using data tested from training data with split validation in the RapidMiner 10.4.0 data analysis application.

5. Discussion

This research successfully shows that the application of the C4.5 algorithm significantly predicts consumer interest in the sale of land plots at PT Piliruma Rosa Land. From the analysis, the Sold Area attribute emerged as the main factor affecting consumer interest, where large lots are more likely to be in demand than small lots. In addition, the Number of Sold Lots and Project attributes also make significant contributions to predicting consumer interest. The results of this study indicate that the Nuansa Alam Agroewisata and Luxury projects are more attractive than the Harmoni Farm House project. This can be a reference for PT Piliruma Rosa Land in designing a more effective marketing strategy, focusing on projects that appeal more to consumers.

Evaluation of the model using several evaluation metrics, such as accuracy, precision, recall, and confusion matrix, shows that the model built has an accuracy rate of 91.18%. The high precision and recall values for the High and Not High classes indicate that the predictions made by this model are accurate and reliable enough to be used in strategic decision-making. These results are in line with previous research showing that the C4.5 algorithm is effective in data classification and prediction.

In addition, the finding that high price per square meter is not always the determining factor of consumer interest, but rather the area sold and the number of lots sold are more dominant, provides new insights for PT Piliruma Rosa Land. The company can focus more on managing lot size and the number of lots sold in its sales strategy.

This research significantly contributes to our understanding of the factors that influence consumer interest in land plot sales and shows that the use of the C4.5 algorithm can improve the effectiveness of marketing and sales strategies. The findings are expected to be applied by other companies in similar industries to optimize sales performance and better understand their consumers' preferences.

For future research, explore the use of other algorithms that might further improve prediction accuracy. In addition, the addition of other attributes or external factors, such as economic indicators and demographic data, may provide a more comprehensive and reliable prediction model for the property sector. This research provides a strong foundation for the development of better prediction models and more effective marketing strategies in the future.

6. Conclusion

This research successfully applies the C4.5 algorithm to analyze land plot sales data at PT Piliruma Rosa Land. From the analysis, it was found that the Sold Area attribute is the main factor affecting consumer interest, with large lots tending to attract high interest and small lots tending to be less desirable. In addition, the attributes, Number of Sold Lots, and Project were also influential, especially in the normal sold lots category. The project Nuansa Alam Agroewisata dan Luxury is more desirable than Harmoni Farm House. The resulting decision tree model shows a high accuracy of 91.18%, indicating that this model is reliable enough to predict consumer interest. The results of this study can assist PT Piliruma Rosa Land in designing more effective marketing strategies based on a better understanding of the factors that influence consumer interest.

References

Anggraini, S., Defit, S., & Nurcahyo, G. W. (2018). Analisis Data Mining Penjualan Ban Menggunakan Algoritma C4.5. *Jurnal Ilmu Teknik Elektro Komputer Dan Informatika*, *4*(2), 136–143.

https://core.ac.uk/download/pdf/295348196.pdf
Azwanti, N. (2018). Analisa Algoritma C4.5 Untuk Memprediksi Penjualan Motor Pada Pt. Capella Dinamik Nusantara Cabang Muka Kuning. *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, *13*(1), 33. https://doi.org/10.30872/jim.v13i1.629
Cherfi, A., Nouira, K., & Ferchichi, A. (2020). MC4.5 Decision tree algorithm: an improved use of continuous attributes. *Int. J. Comput. Intell. Stud.*, *9*, 4–17. https://doi.org/10.1504/ijcistudies.2020.10028137
Dewi, K. R., Mauladi, K. F., & Masruroh. (2020). Analisa Algoritma C4.5 untuk Prediksi Penjualan Obat Pertanian di Toko Dewi Sri. *Seminar Nasional Inovasi Teknologi*, *25*, 109–114.
Fadhila, F., & Hasugian, P. S. (2022). Application of C4.5 Algorithm to Prediction Sales at PT. Sumber Sayur Segar. *Journal of Intelligent Decision Support System (IDSS)*, *5*(1), 10–19. https://doi.org/10.35335/idss.v5i1.45
Fitriani, E., Aryanti, R., Saepudin, A., & Ardiansyah, D. (2020). Penerapan Algoritma C4.5 Untuk Klasifikasi Penempatan Tenaga Marketing. *Paradigma - Jurnal Komputer Dan Informatika*, *22*(1), 72–78. https://doi.org/10.31294/p.v22i1.6898
Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier Science. https://books.google.co.id/books?id=pQws07tdpjoC
Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020). RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12343 LNCS*, 503–515. https://doi.org/10.1007/978-3-030-62008-0_35
Kirkpatrick, K. (2019). 1. What Is Data Mining? *Data Mining for the Social Sciences*. https://doi.org/10.4135/9781526493095
Liu, J., Ning, B., & Shi, D. (2019). Application of Improved Decision Tree C4.5 Algorithms in the Judgment of Diabetes Diagnostic Effectiveness. *Journal of Physics: Conference Series*, *1237*(2). https://doi.org/10.1088/1742-6596/1237/2/022116
Mijwil, M. M., & Abttan, R. A. (2021). Utilizing the Genetic Algorithm to Pruning the C4.5 Decision Tree Algorithm. *Asian Journal of Applied Sciences*. https://doi.org/10.24203/AJAS.V9I1.6503
Muttaqien, R., Pradana, M. G., & Pramuntadi, A. (2021). Implementation of Data Mining Using C4.5 Algorithm for Predicting Customer Loyalty of PT. Pegadaian (Persero) Pati Area Office. *International Journal of Computer and Information System (IJCIS)*, *2*(3), 64–68. https://doi.org/10.29040/ijcis.v2i3.36
Puspita, D., Aminah, S., & Arif, A. (2022). Prediction System for Credit Eligibility Using C4.5 Algorithm. *Journal of Informatics and Telecommunication Engineering*, *6*(1), 148–156. https://doi.org/10.31289/jite.v6i1.7311
Quinlan, J. R. (2014). *C4.5: Programs for Machine Learning*. Elsevier Science. https://books.google.co.id/books?id=b3ujBQAAQBAJ
Rohman, A., & Rufiyanto, A. (2019). Implementasi Data Mining Dengan Algoritma Decision Tree C4 . 5 Untuk Prediksi Kelulusan Mahasiswa Di Universitas Pandaran. *Proceeding SINTAK 2019*, 134–139.
Shurrab, S., & Duwairi, R. (2021). Effect of Missing Data Treatment on the Predictive Accuracy of C4.5 Classifier. *International Journal on Communications Antenna and Propagation (IRECAP)*. https://doi.org/10.15866/irecap.v11i3.19721
Susanti, M., Kom, M., & Kom, M. (2018). *Prediksi Pengangkatan Karyawan Kontrak Menjadi Karyawan Tetap Menggunakan Decision Tree Pada PT . Baskara Cipta Pratama*. *6*(1), 1–7.
Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *SIGMOD Rec.*, *31*(1), 76–77. https://doi.org/10.1145/507338.507355
Yuni, R., & Putri, R. A. (2023). Penerapan Algoritma C4 . 5 Untuk Prediksi Jumlah Produksi Kelapa Sawit. *Jurnal Media Informatika Budidarma*, *7*(4), 1749–1757. https://doi.org/10.30865/mib.v7i4.6861