

CLASSIFICATION OF PRODUCT PREDICATES BASED ON SALES RATE USING THE C4.5 DECISION TREE ALGORITHM IN RETAIL COMPANIES

Salman Haafizh¹, Agustiena Merdekawati², Yuri Yuliani³

Information Systems, Bina Sarana Informatika University

E-mail: ¹salmanxfxa@gmail.com, ²agustiena.atd@bsi.ac.id, ³yuri.yyi@bsi.ac.id

Abstract

General Background: Retail companies serve as crucial intermediaries between producers and end consumers, providing a wide range of products to meet daily needs. **Specific Background:** However, many retail companies encounter challenges in inventory management and stock provision, often stemming from insufficient analysis of sales and inventory data. **Knowledge Gap:** Existing research on inventory management in retail lacks a focus on predictive analytics techniques that leverage sales data to optimize ordering strategies. **Aims:** This study aims to identify sales patterns using the Decision Tree C4.5 algorithm, with the goal of predicting sales for various products to enhance ordering strategies. **Results:** Employing primary data collected via the company's API and direct interviews with Order Management staff and the regional director of Jabodetabek, sales data spanning six months (November 2023 to April 2024) was analyzed using data mining techniques on the RapidMiner platform. The findings reveal that the Decision Tree algorithm effectively identifies product sales predicates, achieving a model accuracy of 96.20%. **Novelty:** This research introduces a data-driven approach to inventory management in retail, utilizing advanced decision tree algorithms for enhanced sales prediction. **Implications:** The implementation of the proposed model is expected to significantly improve the efficiency and effectiveness of the company's ordering processes, ultimately leading to better inventory control and customer satisfaction.

Keywords : Sales Analysis, Determining Sales Predicate

Corresponding Author; Salman Haafizh

E-mail: salmanxfxa@gmail.com

DOI: <https://doi.org/10.61796/ijmi.v1i3.201>



Introduction

Retail companies are businesses that directly interact with end consumers, individuals who purchase products or services for their personal use. Retail companies act as intermediaries between distributors and end consumers, providing a wide variety of products and brands for daily needs. One example of a store from this retail company is a supermarket.

A supermarket is a shopping place that offers a wide range of products and brands for daily needs, with many shoppers coming to fulfill their daily needs such as consumable goods that are required on a daily basis. With a diverse sales model, supermarkets experience rapid growth and high market competition. (Shafarindu et al., 2021)

In terms of ordering goods, retail companies often face issues related to supply or inventory. There are several types of problems, namely: insufficient stock, excess stock, and unsold goods. Ordering goods will be the main factor in reducing those problems, so the accuracy and effectiveness of the ordering process need to be improved. There is one important thing that companies often overlook before making forecasts, which is the analysis of goods to determine the classification of items with low, medium, high, or no sales. According to Dhika's opinion in his journal, he states: The company lacks analysis of product sales and inventory stock in the warehouse, and if the company never conducts sales analysis, it will not know whether this month's product sales are higher or lower than the previous month and will not know which products are the best sellers. (Faisal et al., 2021).

Based on research conducted by Dewanti (2022) on sales at pharmacies, there are similar issues related to inventory stock. Purchasing items without conducting an analysis process will impact future sales, with some problems being stock shortages or excess stock. This is caused by the lack of analysis related to the items, leading to errors in stock provision. (Dewanti et al., 2022).

The predicate will become a new attribute that influences the forecast of product orders. By knowing the predicates of these items, the company can make additional considerations regarding the forecast to be made.

The classification of goods can enhance the level of accuracy and effectiveness before placing an order. Accuracy and effectiveness in ordering goods mean that the ordered items are neither excessive nor insufficient. Data analysis can use the data mining method of decision tree classification or the C4.5 algorithm. The data used for the C4.5 decision tree classification is sales data. By utilizing sales data from the previous months, an accurate sales pattern can be determined to be used as one of the attributes in deciding on the ordering of goods.

1. Purpose

The purpose of this research is to establish a new attribute that will influence the process of ordering goods using the Decision Tree classification method of the C4.5 algorithm in data mining. This attribute will serve as a determinant of the quantity of goods ordered.

2. Previous Research

The research by Faisal et al. (2021) used a similar approach with a focus on predicting restocking of items based on historical sales data. The results of the study indicate that the use of the decision tree algorithm can provide a high level of accuracy in predicting restock needs, with an accuracy rate of 88.89%. The research conducted by Shafarindu et al. (2021) titled "Sales Data Classification in Supermarkets Using the Decision Tree Method" aims to find the sales index from a dataset containing supermarket data with various information related to the supermarket. This research uses the C4.5 decision tree algorithm classification method, MinMax normalization, and data splitting using K-Fold Cross Validation and Holdout Validation methods. The test results show that with the Hold Out data splitting method, the regression accuracy value is 0.501 and the Decision Tree accuracy is 0.5. Meanwhile, with the K-Fold data splitting method, the regression accuracy value is 0.466 and the decision tree accuracy is 0.492. The highest accuracy in

the branch class is 1.0, while in the customer type class, the highest accuracy is obtained from the decision tree using the Hold Out method at 0.5.

Research by Dewanti et al., 2022 titled "Prediction of Drug Inventory for Sales Process Using Decision Tree Method in Pharmacies" has issues in predicting item needs, resulting in inventory that does not match consumer demand. The result of this research is the development of a system capable of predicting sales. The results of the validity test show an accuracy of 89%, indicating that the system has good performance.

Methods

1. Research Framework

The research framework is a stage of the research process that serves as a guideline and reference to ensure that the research can be focused, structured, and systematic. In this research, data mining becomes a relevant method. Data mining is a series of processes to extract added value from a dataset in the form of knowledge that has not been manually known until now. (Rosela, 2019).

By utilizing data mining techniques, research can be more effective in identifying hidden patterns and gaining deeper insights from the existing data. This process enables researchers to discover valuable information and support research objectives in a more efficient and targeted manner.

There are six stages in this research, here are the stages carried out for the data mining process.

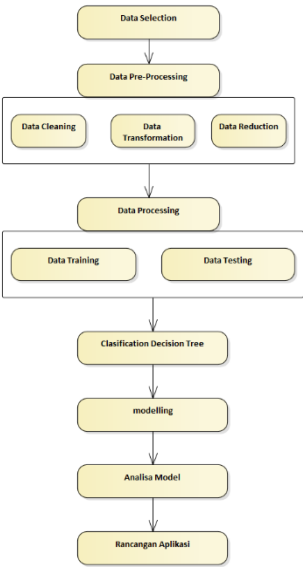


Figure 1. Research Framework

a. Data Selection

The researcher obtained sales data from the server through access to the company's Application Programming Interface (API), with the selected data ranging from November 2023 to April 2024. The column attribute that will be used for decision-making is the Suggestion attribute.

b. Data Pre-Processing

1) Data Cleaning

The data cleaning process involves filling in empty attributes and removing duplicate rows.

2) Data Transformation

The application in this research involves changing the content of attributes from numerical or integer data to text or string form, as well as generalizing overly complex attribute content to make it simpler.

a) Store Code and Store Name: These attributes indicate the name of a store.

Functionally, these attributes serve the same purpose as the store's identity, but each store has a different likelihood of selling goods, so their labeling is generalized to be easily distinguished by the C4.5 decision tree algorithm.

b) Sales Qty, Sales Amt, Vat Amt, Total Cost, and Margin:

These attributes contain numerical data. In the classification using the C4.5 Decision Tree Algorithm, the attributes used must be in string form, so the data, which is originally numerical, must be transformed into text to be recognized by the C4.5 decision tree algorithm.

3) Data Reduction

Unnecessary attributes can become complicated and inaccurate due to containing noise (unnecessary information) and can cause overfitting. (Model terlalu kompleks).

a) Month and Sales Date: Data such as dates have unique characters that are diverse (not duplicated) with the same function, so this type of data can cause the model to overfit.

b) Sub Family Code and Sub Family Name: This attribute indicates the category of an item. The level of detail required is at the item or product level, so category data is not needed for modeling.

c) Vendor Code and Vendor Name: This attribute contains supplier data that supplies to the store, this data is not necessary because the assessment will be based on the store's sales of goods.

ITEM_CODE	ITEM_NAME	PRIORITY	STORE	MANAGEMENT	INCOME	SUGGESTION
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Normal	Highest	High	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Low	Highest	Normal	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Low	High	Normal	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Normal	Highest	High	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Normal	Highest	High	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Normal	Highest	High	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Low	Highest	Normal	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Normal	Highest	High	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Low	Highest	Normal	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Normal	Highest	High	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Low	Low	Medium	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Low	High	Low	
10010189001	TEH PUCUK HARUM PET 1300ML	Middle	Low	High	Normal	

Figure 2. Sales Attributes After Pre-Processing

c. Data Processing

After pre-processing, the data will be entered into the open-source software RapidMiner. The data will undergo a split process into Training data and Testing data. This splitting process will be performed using the data split operator, with a ratio of 70% : 30% to be used.



Figure 3. Split Operator in RapidMiner

d. Classification Decision Tree

A tree is a data structure consisting of nodes and edges. Nodes in a tree are divided into three types: root nodes, branch/internal nodes, and leaf nodes. A decision tree is a simple representation of a classification scheme for a subset of classes, where the root and internal nodes represent attribute names, the edges are labeled to represent possible attribute values, and the leaf nodes represent different classes. (Nasrullah, 2021).

e. Modeling

The model is the final result of the decision tree process, containing rules related to the classification of product sales predicates. These rules can be used to provide labels in the form of product sales predicates, which serve as a reference in the product ordering process.

f. Model Analysis

The model formed from the Decision Tree process will be analyzed by checking accuracy and classification errors using the Performance operator.

2. Research Instruments

To determine product predicates based on sales, the researcher chose to use the C4.5 Decision Tree algorithm for classification. The type of data used is primary data obtained from the company's server sales archives. Data adjustments are necessary to meet the requirements of the Decision Tree classification, so several attribute determinations and changes were made. In this research, the researcher will use RapidMiner and Excel 365 software to analyze and manage data.

a. Product Data Attributes

Product data consists of two attributes, namely Item_Code and Item_Name.

Table 1. List of Top 10 Beverages

NO	ITEM CODE	ITEM NAME
1	10081175001	LE MINERALE BTL 600ML
2	10081006001	AQUA WATER BOTTLE 330 ML
3	10081197001	LE MINERALE BTL 330ML
4	10081089001	NESTLE PURE LIFE WATER 600 ML
5	10051024001	MARJAN SYRUP COCOPANDAN BTL 460ML
6	10082055001	PRIM A WATER BOTTLE 1500ML
7	10082005001	AQUA WATER BOTTLE 1500 ML
8	10083018001	AQUA WATER GALLON 19 LT
9	10010189001	TEH PUCUK HARUM PET 1300ML
10	10050000001	ABC SQUASH ORANGE BOTTLE 450 ML

Attribute Priority Store

Each store has a different sales level based on its sales figures. There are three levels of priority: Potential, indicating that the store has a high potential to generate good sales. Middle, indicating that the store has a good sales record. High, indicating that the store has an excellent sales record. Attribute Management Attribute Management is a label that indicates the sales level of an item at the store level, determined based on the average

sales over six months by calculating the cumulative sales quantity divided by 180 days, with one month counted as 30 days.

$$AVG = (\text{Total sales per item})/(\text{Total Days})$$

The process produced the following parameter data, which was approved by the company as the determinant of the Management attributes. Income Attribute The income attribute contains labels that indicate the level of profit from the sale of goods based on the sales margin. This attribute is determined based on the percentage level of the sales margin of the goods. The researcher presented their calculations in the following manner:

$$MRG = (\text{Monto de Margen})/(\text{Total Amount})$$

Explanation:

MRG = Margin in percent

Margin Amount = Amount of margin for each item

Total Amount = Total sales for each item

Based on the results of the analysis and discussions with the company, the following parameters have been established as the reference labels for the Income attribute.

Table 2. Income Attribute Parameter

No	RULE	DESC IN %	INCOME
1	MRG > 25	Lebih dari 25	Highest
2	15 <= MRG <= 25	15 s/d 25	High
3	10 <= MRG <15	10 s/d 14	Normal
4	1 <= MRG < 10	1 s/d 9	Low
5	0	No Data	Bad

3. Data Collection Methods, Population, and Sample

a. Data Collection Methods

1) Data Extraction, the direct retrieval of data available on the company's server, often through the use of APIs provided by the company or through other means such as direct database queries.

2) Interviews, data collection by researchers or interviewers interacting directly with respondents or informants to obtain the necessary information. In this case, the researcher conducted interviews with the Order Management division and the Territory Director of Jabodetabek for the company.

b. Population and Sample

1) Population, According to Sugiyono, "Population is defined as the area of generalization consisting of objects/subjects that have certain qualities and characteristics determined by the researcher to be studied and then concluded." (Suriani et al., 2023) In this study, the population consists of sales data from stores with the highest sales levels over a six-month period, starting from November 2023 to April 2024.

2) Sample, Hibberts states, "A sample is a group of elements selected from a larger group with the hope that studying this smaller group (the sample) will reveal important information about the larger group (the population)." (Firmansyah & Dede, 2022). In this study, to determine the sample, the Cluster Sampling (Area Sampling) technique was used to select samples based on regions with the highest transaction levels. "Cluster Sampling (Area Sampling) is a sampling technique used to determine samples when the object to be studied or the data source is very extensive." (Amin et al., 2023).

c. Data Analysis Method

In this research, a primary method is used where data is directly taken from the company's server. The initial stage is data extraction from the server, based on the initial data from the server which has 15 attributes.

Table 3. Initial Data Attributes

N O	ATRIBUT	DESCRIPTION
1	SALES_MONT H	Transaction month
2	SALES_DATE	Date of transaction
3	STORE_CODE	Shop code consisting of a 5-digit number
4	STORE_NAME	Shop name
5	SUBFAMILY_C ODE	Item category code
6	SUBFAMILY_ NAME	Item category
7	ITEM_CODE	Item code consisting of 11 digits
8	ITEM_NAME	Item name
9	VENDOR_CO DE	Supplier code
10	VENDOR_NA ME	Supplier name
11	SALES_QTY	Quantity of goods sold
12	SALES_AMOU NT_INC_VAT	Total price of goods sold including tax
13	VAT_AMOUN T	Sales tax amount
14	TOTAL_COST	Total price of goods sold

15	MARGIN	Profit earned from sales
----	--------	--------------------------

Next, the interview method was used to obtain explanations related to sales data from the server, as well as data analysis for the process of data cleaning and transformation to meet the needs of decision tree classification. Based on the interview results, a data set was formed consisting of 6 attributes that will be used during data mining.

Table 4. Test Data Attributes

NO	ATRIBUT	DESCRIPTION
1	ITEM_CODE	Item code consisting of an 11-digit number
2	ITEM_NAME	Item name
3	PRIORITY STORE	Store category by sales level
4	MANAGEMENT	Item category by sales level
5	INCOME	Item sales category based on sales margin
6	SUGGESTION	Decision attributes for modeling

Results and Discussion

1. Research Phase Results

This research aims to determine the product sales level at a retail company using the C4.5 decision tree algorithm classification technique. The result of this research is a model that determines new attributes of a product to be used as a benchmark in forecasting order quantities.

2. C4.5 Algorithm Calculation

In performing the C4.5 algorithm calculation, there are several stages that need to be carried out as follows:

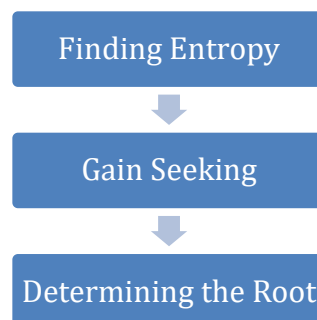


Figure 4. C4.5 Algorithm Workflow

To determine a root, the gain value must first be known; the highest gain value will become a root. To find the gain value, the entropy value must first be determined.

a. Entropy

Entropy is a measure of uncertainty. In the context of the C4.5 decision tree algorithm, entropy is used to measure how pure or mixed a dataset is. The higher the entropy value, the more disordered the data is. The formula for determining the value of entropy is as follows:

$$Entropy(S) = \sum_{i=1}^n -p_i \cdot \log_2(p_i)$$

Explanation:

S = Set of cases

n = Number of partitions of S

p_i = Proportion of class i relative to S

The dataset has a total of 264 rows. In the testing, the data will be divided with a ratio of 70% : 30%, with 70% of the data becoming the training data and 30% becoming the testing data. Out of the 264 rows, 185 rows will be used as training data and 75 rows as testing data. For the dataset table as training data, which is divided by the data split operator in the RapidMiner application.

Then the form of the entropy formula is as follows:
Entropy (Total) = ((-S₁/S)*Log₂ (S₁/S))+ ((-S₂/S)*Log₂ (S₂/S))+ ((-S₅/S)*Log₂ (S₅/S))

Based on the formula above, the researcher will implement it into an Excel formula for calculations. Here is the formula for calculating total entropy using Excel:
Entropy (Total) = ((-S₁/S)*IMLOG2(S₁/S)) + ((-S₅/S)*IMLOG2(S₅/S))

Explanation:

IMLOG2 = Excel formula to return the base 2 logarithm of a complex number.

Table 5. Total Entropy Calculation Results

ALL		SUGGEST						Entropy
		Jumlah	Bad Performance	Low	Normal	Medium	High	
		S	S ₁	S ₂	S ₃	S ₄	S ₅	
TOTAL		185	12	24	37	43	69	2.12259229

The entropy calculation was also performed on the attributes Item Name, Priority Store, Management, and Income using the same method. Here are the entropy results for all attributes:

Table 6. Entropy Results of All Attributes

		SUGGEST						Entropy
		Jumlah	Bad Performance	Low	Normal	Medium	High	
		S	S ₁	S ₂	S ₃	S ₄	S ₅	
TOTAL		185	12	24	37	43	69	2.12259229
ITEM NAME								
	ABC SQUASH ORANGE BOTTLE 450 ML	15	1	0	2	6	6	0
	AQUA WATER BOTTLE 1500 ML	23	2	3	5	4	9	2.136872696
	AQUA WATER BOTTLE 330 ML	21	2	4	2	6	7	2.146543163
	AQUA WATER GALLON 19 LT	20	3	6	6	2	3	2.195461844
	LE MINERALE BTL 330ML	16	0	0	0	4	12	0
	LE MINERALE BTL 600ML	12	0	0	0	2	10	0
	MARJAN SYRUP COCOPANDAN BTL 460ML	28	2	4	9	8	5	2.159534726
	NESTLE PURE LIFE WATER 600 ML	15	0	2	2	3	8	0
	PRIM A WATER BOTTLE 1500ML	16	0	2	5	3	6	0
	TEH PUCUK HARUM PET 1300ML	19	2	3	6	5	3	2.214812225
Priority Store								
	High	67	5	9	3	15	35	1.841878481
	Middle	63	6	9	16	10	22	2.177821565
	Potential	55	1	6	18	18	12	1.987784712
Management								
	Highest	26	0	0	0	2	24	0
	High	43	0	0	1	16	26	0
	Normal	59	0	1	14	25	19	0
	Low	51	10	23	18	0	0	0
	Lowest	6	2	0	4	0	0	0
Income								
	Highest	58	0	4	12	7	35	0
	High	48	0	5	9	5	29	0
	Mid	34	1	7	6	17	3	1.869727098
	Low	29	6	7	3	11	2	2.100398417
	Bad	16	5	1	7	3	0	0

b. Gain and Root

Gain or Information gain is a measure used in decision tree algorithms to determine how well an attribute can split the dataset into the desired classes. Information gain measures the reduction in uncertainty (Entropy) of a dataset after the dataset is split based on certain attributes. The attribute with the highest information gain will be chosen as the root. Here is the formula to find information gain:

$$Gain(S, A) = entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} entropy(S_i)$$

Explanation:

Gain S = Set of Cases

Gain A = Feature

n = Number of Partitions of Attribute A

|S_i| = Proportion of S_i to S

|S| = Number of cases in S

Based on the gain formula above, the breakdown for finding gain information is as follows:

$$Gain(S, A) = Entropy Total - \left(\left(\frac{S_i}{S}\right) * Entropy(S_i)\right) + \dots \left(\left(\frac{S_n}{S}\right) * Entropy(S_n)\right)$$

Explanation:

Gain S = Set of Cases

Gain A = Feature

n = Number of Partitions of Attribute A

S_i = Proportion of S_i to S

S = Number of cases in S

To simplify the calculations, the researcher will implement the formula into an Excel formula, then divide it into 2 parts to obtain the following formula:

1) Calc Column

The Calc Column or Calculation will contain the formula to calculate the following part of the gain formula:

$$\sum_{i=1}^n \frac{|s_i|}{|s|} \text{entropy}(S_i)$$

So the formula in Excel is as follows:

$$\text{Calc} = \frac{\text{Jumlah Data}}{\text{Total Data}} \times \text{Entropi Proporsi}$$

2) Gain Column

The gain column will combine the gain formula with the formula in the Calc column, thus providing information on the gain of an attribute. The formula in this column is as follows:

$$\text{Gain}(S, A) = \text{entropy}(S) - \text{SUM}(\text{Calc})$$

Explanation:

Entropy(S) = Total Entropy

SUM = Excel formula used to sum data in a cell

Table 7. Gain Calculation Results

	Jumlah	SUGGEST					Entropy	Calc	Gain
		Bad Performance	Low	Normal	Medium	High			
TOTAL	185	5	12	24	37	43	69	2.12291224	
ITEM NAME									0.821602453
ABC SQUASH ORANGE BOTTLE 450 ML	15	1	0	2	6	6	0	0	
AQUA WATER BOTTLE 1500 ML	23	2	3	5	4	9	2.136872696	0.265665254	
AQUA WATER BOTTLE 330 ML	21	2	4	2	6	7	2.146543163	0.243661656	
AQUA WATER GALLON 19 LT	20	3	6	6	2	3	2.195461844	0.237347226	
LE MINERALE BTL 330ML	16	0	0	0	4	12	0	0	
LE MINERALE BTL 500ML	12	0	0	0	2	10	0	0	
MARJAN SYRUP COCOPOANDAN BTL 450ML	28	2	4	9	8	5	2.159534726	0.326848499	
NESTLE PURE LIFE WATER 600 ML	15	0	2	2	3	8	0	0	
PRIMA WATER BOTTLE 1500ML	16	0	2	5	3	6	0	0	
TEH PUCUK HARUM PET 1300ML	19	2	3	6	5	3	2.214812225	0.227467201	
Priority Score									0.122934041
High	67	5	9	3	15	35	1.841878481	0.667058615	
Middle	63	6	9	16	10	22	2.177821565	0.741636533	
Potential	55	1	6	18	18	12	1.987784712	0.590963023	
Management									2.12291224
Highest	26	0	0	0	2	24	0	0	
High	43	0	0	1	16	26	0	0	
Normal	59	0	1	14	25	19	0	0	
Low	51	10	23	18	0	0	0	0	
Lowest	6	2	0	4	0	0	0	0	
Income									1.449715126
Highest	58	0	4	12	7	35	0	0	
High	48	0	5	9	5	29	0	0	
Mid	34	1	7	6	17	3	1.849727098	0.343625521	
Low	29	6	7	3	11	2	2.100398417	0.329251644	
Bad	16	5	1	7	3	0	0	0	

The highest information gain, then the Management attribute will become the root. This process continues to be repeated for other attributes until a Decision Tree is formed.

3. RapidMiner

a. RapidMiner Design


The following is the design used by the researcher in conducting trials on the RapidMiner application in Figure 2:



Figure 5. RapidMiner Design

1) Data Set

The dataset contains all sales data for a period of 6 months starting from November 2023 – April 2024. The attributes present in this dataset can be seen in RapidMiner in Figure 3:



Row No.	Suggestion	ITEM_CODE	ITEM_NAME	Priority Store	Management	Income
1	Medium	10010189001	TEH PUCUK ...	High	High	Bad
2	High	10010189001	TEH PUCUK ...	High	High	High
3	High	10010189001	TEH PUCUK ...	High	High	Highest
4	Medium	10010189001	TEH PUCUK ...	High	High	Low

Figure 6. Dataset in RapidMiner

Based on Figure 3, it is known that the dataset has a total of 264 records, has six attributes, and one of the attributes is a label. Decision tree classification is a machine learning method based on supervised learning, so labels are needed as teachers to determine whether the data is correct or not. From the data in Figure 3, the label is in the Suggestion attribute. Here are the detailed attributes information from the dataset:



Figure 7. RapidMiner Dataset Attribute Statistics

2) Data Split Operator

To divide the dataset into training and testing data, the researcher uses a 70% : 30% split. For this division, the researcher chooses the Stratified Sampling technique as the sampling type, which ensures that each randomly taken subset from the data has the same class distribution as the entire dataset.

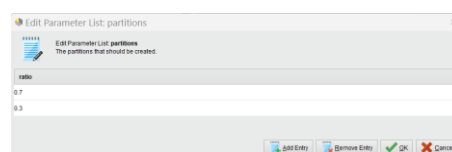


Figure 8. Ratio Split Data

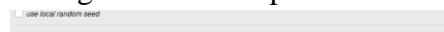


Figure 9. Parameter Split Data

3) Árbol de Decisión del Operador

The tool used to create decision tree models in the process of data analysis and machine learning.

4) Operador Aplicar Modelo

Apply Model will apply the new model to the test data.



Figure 10. Operator Apply Model

5) Operator Performance

To measure or test the accuracy level of the formed model, this operator will display the confusion matrix and the accuracy level of the newly formed model.

b. Decision Tree Model

Based on the sales dataset, a decision tree model was obtained from the RapidMiner application.



Figure 11. Decision Tree Model

From Figure 8, we can see that the management attribute is the root attribute, in accordance with the manual calculation test with Excel, which shows that the management attribute has the highest information gain value. Here is the description or text tree of the decision tree model formed from the RapidMiner application:

```

Tree
Management = High
| Income = Bad: Medium (Medium=4, High=0, Normal=1, bad performance=0, Low=0)
| Income = High: High (Medium=0, High=11, Normal=0, bad performance=0, Low=0)
| Income = Highest: High (Medium=0, High=14, Normal=0, bad performance=0, Low=0)
| Income = Low: Medium (Medium=8, High=0, Normal=0, bad performance=0, Low=0)
| Income = Mid: Medium (Medium=7, High=0, Normal=0, bad performance=0, Low=0)
Management = Highest: High (Medium=2, High=24, Normal=0, bad performance=0, Low=0)
Management = Low
| Income = Bad: bad performance (Medium=0, High=0, Normal=1, bad performance=4, Low=0)
| Income = High
| | Priority Store = High: Low (Medium=0, High=0, Normal=0, bad performance=0, Low=4)
| | Priority Store = Middle: Normal (Medium=0, High=0, Normal=6, bad performance=0, Low=0)
| | Priority Store = Potential: Normal (Medium=0, High=0, Normal=3, bad performance=0, Low=0)
| Income = Highest
| | Priority Store = High: Low (Medium=0, High=0, Normal=0, bad performance=0, Low=5)
| | Priority Store = Middle: Normal (Medium=0, High=0, Normal=5, bad performance=0, Low=0)
| | Priority Store = Potential: Normal (Medium=0, High=0, Normal=4, bad performance=0, Low=0)
| Income = Low
| | Priority Store = High: bad performance (Medium=0, High=0, Normal=0, bad performance=2, Low=0)
| | Priority Store = Middle: Low (Medium=0, High=0, Normal=0, bad performance=0, Low=6)
| | Priority Store = Potential: Low (Medium=0, High=0, Normal=0, bad performance=0, Low=2)
| Income = Mid: Low (Medium=0, High=0, Normal=0, bad performance=3, Low=5)
Management = Lowest: Normal (Medium=0, High=0, Normal=6, bad performance=2, Low=1)
Management = Normal
| Income = Bad
| | Priority Store = Middle: Normal (Medium=0, High=0, Normal=2, bad performance=0, Low=0)
| | Priority Store = Potential: Low (Medium=0, High=0, Normal=0, bad performance=0, Low=2)
| Income = High
| | Priority Store = High: High (Medium=0, High=3, Normal=0, bad performance=0, Low=0)
| | Priority Store = Middle: High (Medium=0, High=6, Normal=0, bad performance=0, Low=0)
| | Priority Store = Potential: Medium (Medium=5, High=0, Normal=0, bad performance=0, Low=0)
| Income = Highest
| | Priority Store = High: High (Medium=0, High=2, Normal=0, bad performance=0, Low=0)
| | Priority Store = Middle: High (Medium=0, High=7, Normal=0, bad performance=0, Low=0)
| | Priority Store = Potential: Medium (Medium=8, High=0, Normal=0, bad performance=0, Low=0)
| Income = Low
| | Priority Store = High: Medium (Medium=3, High=0, Normal=0, bad performance=0, Low=0)
| | Priority Store = Middle: Medium (Medium=3, High=0, Normal=0, bad performance=0, Low=0)
| | Priority Store = Potential: Normal (Medium=0, High=0, Normal=4, bad performance=0, Low=0)
| Income = Mid
| | Priority Store = High: Medium (Medium=2, High=0, Normal=0, bad performance=0, Low=0)
| | Priority Store = Middle: Medium (Medium=4, High=0, Normal=0, bad performance=0, Low=0)
| | Priority Store = Potential: Normal (Medium=0, High=0, Normal=4, bad performance=0, Low=0)

```

Figure 12. Description Tree RapidMiner

From the results of the decision tree above, the rules can be made so that they can be implemented in the application, here are the rules that are formed:

- 1) If Management = Highest, then Suggest = High
- 2) If Management = High and Income = Highest or High, then Suggest = High
- 3) If Management = High and Income = bad or Low or Mid, then Suggest = Medium

- 4) If Management = Normal and Income = High or Highest and Priority Store = High or Middle, then Suggest = High
- 5) If Management = Normal and Income = High or Highest and Priority Store = Potential, then Suggest = Medium
- 6) If Management = Normal and Income = Mid or Low and Priority Store = High or Middle, then Suggest = Medium
- 7) If Management = Normal and Income = Mid or Low and Priority Store = Potential, then Suggest = Normal
- 8) If Management = Normal and Income = Bad and Priority Store = Middle, then Suggest = Normal
- 9) If Management = Normal and Income = Bad and Priority Store = Potential, then Suggest = Low
- 10) If Management = Low and Income = Highest or High and Priority Store = High, then Suggest = Low
- 11) If Management = Low and Income = Highest or High and Priority Store = Middle or Potential, then Suggest = Normal
- 12) If Management = Low and Income = Mid, then Suggest = Low
- 13) If Management = Low and Income = Low and Priority Store = High, then Suggest = Bad Performance
- 14) If Management = Low and Income = Low and Priority Store = Middle or Potential, then Suggest = Low
- 15) If Management = Low and Income = Bad, then Suggest = Bad Performance
- 16) If Management = Lowest, then Suggest = Normal

4. Model Analysis

Based on the testing, the following is the confusion matrix that was obtained, which can be seen in Table 6:

Table 13. Confusion Matrix Model Decision Tree

	true High	true Medium	true Normal	true Low	true bad performance
pred. High	28	1	0	0	
pred. Medium	0	19	1	0	
pred. Normal	0	0	15	0	
pred. Low	0	0	0	10	
pred. bad performance	0	0	0	0	

a. Accuracy Analysis

The accuracy level in a classification model measures how often the model makes correct predictions. Here is the accuracy formula:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Prediction}}$$

Therefore, the accuracy level can be seen in the calculations below:

$$Accuracy = \frac{76}{79} \approx 0.962 \text{ or } 96.20\%$$

b. Class Precision Analysis

Class precision is an evaluation metric used in classification model analysis to measure the accuracy of predictions for each specific class. Precision calculates the

proportion of correct predictions (True positive) against all predictions made for that class (True positive + False positive). Precision for the predicted class is calculated using the following formula:

$$Precision = \frac{TP}{TP + FP}$$

Description:

TP = True Positive

FP = False Positive

	true Medium	true High	true Normal	true bad performance	true Low	class precision
pred. Medium	19	0	1	0	0	95.00%
pred. High	1	28	0	0	0	95.55%
pred. Normal	0	0	15	1	0	93.75%
pred. bad performance	0	0	0	4	0	100.00%
pred. Low	0	0	0	0	10	100.00%

Figure 14. Hasil Class Precision RapidMiner

c. Class Recall Analysis

Class recall is used to measure how well the model can recognize or detect instances of a class that actually exist in the dataset. The higher the recall value, the better the model is at recognizing positive instances of the intended class. The formula to find the class recall value for the true class is as follows:

$$Recall = \frac{TP}{TP + FN}$$

Description:

TP = True Positive

FP = False Positive

	true Medium	true High	true Normal	true bad performance	true Low
pred. Medium	19	0	1	0	0
pred. High	1	28	0	0	0
pred. Normal	0	0	15	1	0
pred. bad performance	0	0	0	4	0
pred. Low	0	0	0	0	10
class recall	95.00%	100.00%	93.75%	80.00%	100.00%

Figure 15. RapidMiner Class Recall Results RapidMiner Class Recall Results

5. Application Design

The proposed application design by the researcher includes user interface design, application workflow, and the integration of the C4.5 Decision Tree algorithm with the goods processing system.

a. Diagrama de Casos de Uso

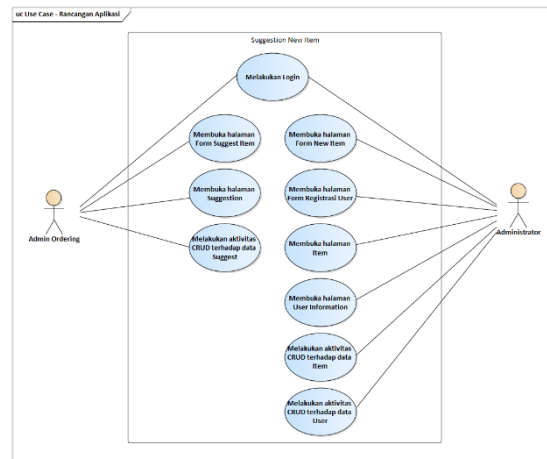


Figure 16. Use Case Application Design

Here are the interactions of each actor with the application:

1. Admin Ordering Actor

- Logging in
- Opening the Suggest Item Form page
- Opening the Suggestion page
- Performing CRUD (Create, Read, Update, Delete) activities on Suggestion data.

2. Administrator Actor

- Logging in
 - Opening the New Item Form page
 - Opening the User Registration Form page
 - Opening the Item page
 - Opening the User Information page
 - Performing CRUD (Create, Read, Update, Delete) activities on Item data
 - Performing CRUD (Create, Read, Update, Delete) activities on user data.
- b. Diagrama de Actividad

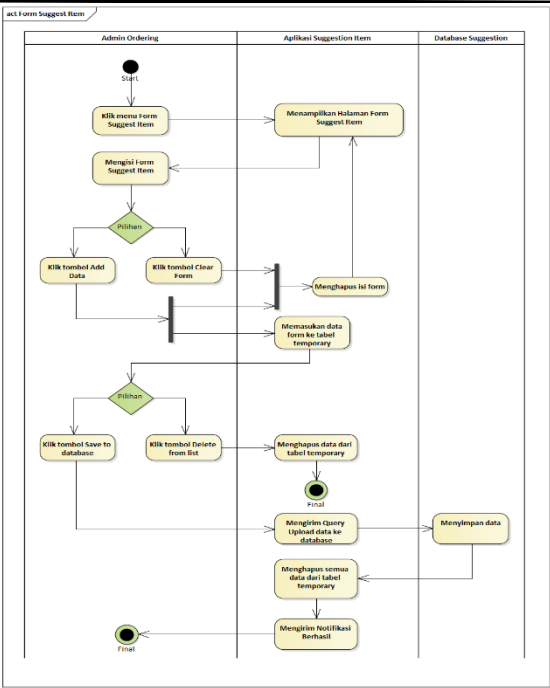


Figure 17. Activity Diagram Form Suggest Item

c. Entity Relationship Diagram

The ERD will illustrate the main entities involved and the relationships between those entities.

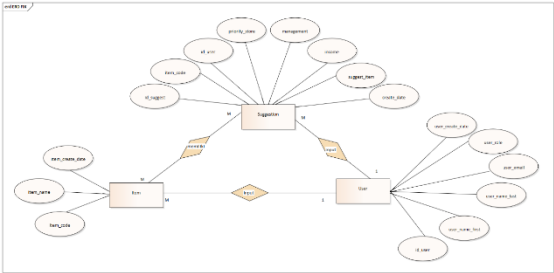


Figure 18. Entity Relationship Diagram Web Order

a. Mockup App

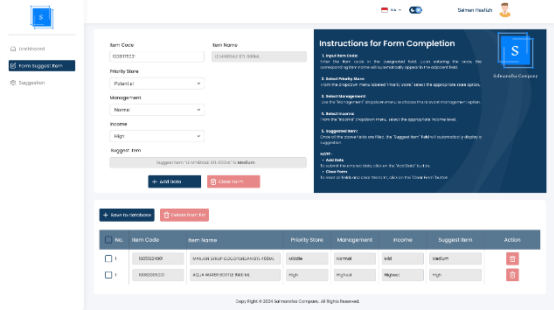


Figure 19. Mockup - Suggest Item Form Page

Conclusion

Based on the research conducted, several conclusions can be drawn regarding the performance of the decision tree classification model used in this study. These conclusions include model performance evaluation, precision and recall, as well as confusion matrix analysis.

First, the decision tree classification model shows a high accuracy rate of 96.20%, indicating that this model is very effective in classifying sales data into the correct categories.

Second, the precision and recall for each class yield good results, especially for the "High" class with a precision of 96.55% and a recall of 100.00%. The "Low" and "Bad Performance" classes also achieved perfect precision (100.00%) and adequate recall. Finally, based on the confusion matrix, the number of prediction errors is very small. For example, there is only one instance that was incorrectly predicted for the "High," "Medium," and "Normal" classes, indicating that the model is capable of minimizing errors in predicting different classes.

References

- [1] N. F. Amin, S. Garancang, and K. Abunawas, "Konsep Umum Populasi dan Sampel dalam Penelitian," *Jurnal Pilar: Jurnal Kajian Islam Kontemporer*, vol. 14, pp. 15–31, 2023. [Online]. Available: <https://www.bing.com/ck/a?!&&p=d0130d833a5583ccJmltdHM9MTcxNjQyMjQwMjZpZ3VpZD0yZmIyN2ZkYS01MDA1LTZyYTYtMmZmNS02ZTY2NTE1MzYyYTImaW5zaWQ9NTE5Nw&ptn=3&ver=2&hsh=3&fclid=2fb27fda-5005-63a6-2ff5-6e66515362a2&psq=jurnal+populasi+adalah&u=a1aHR0cHM6Ly9qb3VybmFsLnVuaXNtdWguYWMuaWQvaW5kZXgucGhwL3BpbGFyL2FydGljbGUvZG93bmxvYWQvMTA2MjQvNTk0Nw&ntb=1>
- [2] F. P. Dewanti, Setiyowati, and S. Harjanto, "Prediksi Persediaan Obat untuk Proses Penjualan Menggunakan Metode Decision Tree pada Apotek," *Jurnal Teknologi Informasi dan Komunikasi (TIKOMSiN)*, vol. 10, no. 1, pp. 25–33, 2022. doi: 10.30646/tikomsin.v10i1.604.
- [3] Dhika Faisal, H. Dhika, and H. Veris, "Penerapan Algoritma Decision Tree dalam Penjualan Handphone," *JRKT (Jurnal Rekayasa Komputasi Terapan)*, vol. 1, no. 4, pp. 239–246, 2021.
- [4] D. Firmansyah and Dede, "Teknik Pengambilan Sampel Umum dalam Metodologi Penelitian: Literature Review," *Jurnal Ilmiah Pendidikan Holistik (JIPH)*, vol. 1, no. 2, pp. 85–114, 2022. doi: 10.55927.
- [5] A. I. Shafarindu et al., "Klasifikasi Data Penjualan pada Supermarket dengan Metode Decision Tree," in *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, 2021, pp. 660–667.
- [6] J. T. Kumalasari and A. Merdekawati, "Analisis Sentimen Terhadap Program Kampus Merdeka pada Twitter Menggunakan Metode Naïve Bayes, Union dan Synthetic Minority Over Sampling Technique (SMOTE)," *SATIN - Sains dan Teknologi Informasi*, vol. 9, no. 1, pp. 01–12, 2023. doi: 10.33372/stn.v9i1.894.
- [7] A. H. Nasrullah, "Implementasi Algoritma Decision Tree untuk Klasifikasi Produk Laris," vol. 7, no. 2, 2021. [Online]. Available: <http://ejournal.fikom-unasman.ac.id>
- [8] Y. Rosela, "Implementasi Klasifikasi Decision Tree Menganalisa Status Penjualan Barang Menggunakan C4.5 (Studi Kasus: PT. Matahari Department Store Medan Mall)," *Jurnal Pelita Informatika*, vol. 7, no. 3, pp. 404–411, 2019.

- [9] D. R. Sitorus et al., “Penerapan Klasifikasi C4.5 dalam Meningkatkan Sistem Pembelajaran Mahasiswa,” in *Komik (Konferensi Nasional Teknologi Informasi dan Komputer)*, vol. 3, no. 1, pp. 593–597, 2019. doi: 10.30865/komik.v3i1.1665.
- [10] N. Suriani, Risnita, and M. S. Jailani, “Konsep Populasi dan Sampling serta Pemilihan Partisipan Ditinjau dari Penelitian Ilmiah Pendidikan,” *IHSAN: Jurnal Pendidikan Islam*, vol. 1, no. 2, pp. 24–36, 2023. [Online]. Available: <http://ejournal.yayasanpendidikandzurriyatulquran.id/index.php/ihsan>