Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 9, No. 2, June 2023, pp. 268-283
ISSN: 2338-3070, DOI: 10.26555/jiteki.v9i2.25970 

# Classification and Clustering of Internet Quota Sales Data Using C4.5 Algorithm and K-Means

Eriska Vivian Astuti, Asep Afandi, Dwi Marisa Efendi
ITBA Dian Cipta Cendikia Kotabumi, State of St. 34518, Indonesia

**ARTICLE INFORMATION**

**Article history:**

Received March 19, 2023
Revised April 18, 2023
Published April 27, 2023

**Keywords:**

C4.5 Algorithm;
K-Means;
Data mining;
Quota

**ABSTRACT**

The number of restrictions or limits on internet use is known as the internet quota. When you use internet data for a short time, you usually run out of bandwidth. In the Candimas South Abung area, many quotas have been sold in various variants. Visitors to quota outlets have access to various kinds of quota references that they can buy. Apart from guaranteeing the quality of the quotas sold, sales always increase every year, especially in the various quota variants. Based on quota data for 2019 to 2022. This study aims to analyze internet quota sales statistics in the Candimas area between 2019 and 2022. In 2021-2022 the classification produces an accuracy of up to 100% where the best-selling data dominates while clustering remains at the same figure, namely 19 data are very salable, 43 data are lacking sold, and 178 data did not sell. We use the C4.5 classification algorithm and K-Means clustering to identify patterns in the data and provide insight into which brand quotas are the most popular. Our findings can help Xena Cell counter owners make informed decisions about which quota to add or remove to optimize sales and minimize losses.



**Corresponding Author**:

Eriska Vivian Astuti, ITBA-Dian Cipta Cendikia Kotabumi, Negara St., Kotabumi 34518, Indonesia
Email:eriskavivianastuti@gmail.com

## 1. INTRODUCTION

The number of restrictions or limits on internet use is known as the internet quota. Currently internet quota is a basic need for everyone around the world [1],[2].

This study uses quota data samples from the South Abung area, Candimas. Apart from guaranteeing the quality of the quotas sold, quota sales always increase every year, especially in the various quota variants. Every day, the Counter in the South Abung Candimas area encounters problems selling available goods that do not meet demand. Counters also do not have predictors or plans for selling quotas, resulting in confusion of goods and losses when the quota has not been sold in the South Abung Candimas area [3]. The counter in the Abung Selatan area of Candimas has developed into an important trade facilitator over time, influencing not only the local economy but also national and regional economic functions. Physical Internet (PI) is a new, all-encompassing, and long-term vision of the future global internet system that aims to significantly increase its sustainability and effectiveness as a mobility community for the Abung Selatan Candimas area [4]. From this research, the quota sales data for the Abung Selatan Candimas area that we have processed will contain a business model where now you can easily reach potential customers like never before with access to unprecedented information and the development of new technologies [5]. In this study, the writer will use 2 calculation methods, namely classification using the C4.5 algorithm and clustering using K-Means.

The ID3 algorithm developed by J. Rose Quinlan forms the basis of the widely used C4.5 Algorithm, which is a development of the ID3 Algorithm [6]. The decision tree includes the C4.5 algorithm. The structure of a decision tree is similar to a flowchart in that each internal node (also known as a non-leaf node) tests an attribute, each branch represents a result set, and each leaf node (also known as a terminal node) becomes a label for a class [7]. The C4.5 algorithm can be used to group or classify data sets. An algorithm known as C4.5

can be used to group or classify data sets [7],[8]. The C4.5 algorithm can handle discrete or continuous numeric data. This missing value attribute can be cleaned with the average value of the variable in question if the data set has several missing observation values or a small number of observations [9]. Decision trees are a well-known classification and prediction technique [10]. C4.5 turns large amounts of data into decision trees that represent the rules [9],[8],[11]. Finding a model or function that describes or differentiates a class or data concept is a classification calculation. The goal is to be able to create classes for objects whose labels are unknown [11],[12]. The method known as a decision tree has the shape of a tree, with each leaf representing a decision made and each branch representing a choice from several other options [1],[2]. The C4.5 algorithm is intended to help classify vehicle test results according to the factors that influence them [14],[15]. By using the C4.5 (Decision Tree) method and data mining techniques, the author can make the right decisions to predict sales quotas that meet the needs of outlet owners and help sales [6],[16].

Next the author connects with another method namely K-Means or RCM type method, a clustering scheme inspired by the collection of courses, taking into account three types of membership: the lower, upper, and boundary areas of each cluster, each of which represents belonging to a particular object, probable, and not certain [17]. Next, choosing k points as cluster centers or k centroids is the first step in K-means clustering. The next step is to assign the center of mass to each point in the data set. The sum of the squared clusters is then minimized to update the centroid (WCSS) [18]. The classification of highly salable, undersold, and unsold internet quotas and the complexity of k-grouping is the subject of our investigation in this paper. The k-Clustering complexity landscape seems to be a more useful and complex outcome than this new "dimensionality". For more details, we will discuss the following issues [19]. Convex grouping, k-means and k-means++ are all used in it [20]. It has an iterative algorithm for clustering analysis of K-means clustering method [21]. Because of its simplicity, ease of implementation, and good interpretation, the k-means clustering method is one of the most widely used methods for clustering problems [22]. However, little is known about its relationship with routine Internet-related behaviors such as signal instability and problematic Internet usage, and the roles played by highly in demand, underperforming, and under-selling [23]. We will reduce non-conforming inter-cluster deployments [24]. Therefore, the authors hope that customers will have greater access to finance to buy certain quota products that do not hinder their needs [25].

This study aims to compare the prediction accuracy of the C4.5 and k-Means algorithms [26],[6]. The reason for using this method is because it is in accordance with the topic we are taking, namely, Internet Quota data [27]. In addition, it has several advantages, one of which is that C4.5 is an efficient decision tree classification algorithm for discrete and numeric type attributes [28]. K-Means clustering and C4.5 classification are combined in the group [29]. So that research is easily understood by every reader.

Prior to losing significant profits the owner consistently needs past sales statistics for all their marketing and sales quotas [30]. This study uses a partition-based clustering algorithm, MacQueen J first proposed the K-means algorithm in 1967 [31]. One popular technique for automatically dividing a data set into k-groups is K-means clustering [32]. In the conventional k-means algorithm, k cluster centers are initialized randomly in space, the distance between each point and the cluster center is calculated, and the location of the cluster center is updated repeatedly until the best cluster center is found [33]. Despite the widespread use of internet data in our daily lives, little is known about how consumers shop for data quota directly [34]. There is data mining, compression, probability density estimation and many other important tasks clustering is the best tool [35]. We focus on selling internet quota as a factor in our research. It has been found that internet use is increasing in a number of ways, including as a result of telework and online shopping (Abigail), and the use of the internet as a substitute for travel has received a lot of attention [36]. From the description of the problem background, the contribution of this research has been found. That is, it can be used as a sales benchmark in the field of internet data quota for quota sellers and internet usage by providing a practical application of the method in the context of selling quotas.

### 1.1. Data Mining

Data discovery analysis known as data seeks to discover unexpected relationships and provide the data owner with an understandable and useful summary of the data [37],[38],[39].

### 1.2. Knowledge Discovery in Databases

Knowledge Discovery in Databases, or KDD, is a comprehensive, labor-intensive method for determining the validity, novelty, usability, and understanding of data. Data mining can be broken down into stages, as shown in Fig. 1, as a series of processes. Users can participate directly in this stage or through the knowledge base [40],[41]. Fig. 1 depicts the KDD process levels that follow step by step or through the knowledge base [41]. Fig. 1 illustrates the KDD process level.

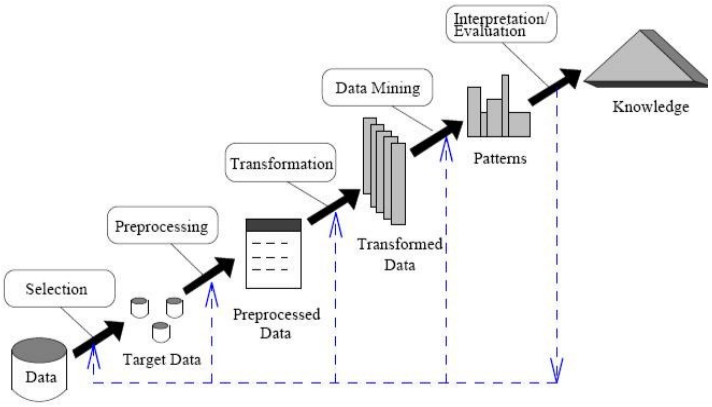


**Fig. 1.** KDD process

The stages of the KDD process consist of:

**1. Data Selection**

Before the information review stage in KDD begins, data must be selected (selected) from a set of operational data. Data from search results can be used in the data mining process and stored separately from the operating database in a file [41].

**2. Data Pre-Processing and Cleaning**

Duplicating data, checking for inconsistent data, and correcting data errors such as typos are all part of the cleanup process [41].

**3. Transformation**

It is the process of transforming selected data so that it is suitable for data mining processes and creative processes which depend on the type or pattern of information in the database that needs to be searched [41].

**4. Data Mining**

The process of finding interesting patterns or information on selected data using certain techniques is called data mining. In data mining, there are many different algorithms, techniques or methods. The goal and overall KDD process determines which method or algorithm is best [41].

**5. Interpretation/Evaluation**

This stage of the KDD process involves determining whether the pattern or information found contradicts previous facts or hypotheses [41].

### 1.3. Application

The definition of implementation is an activity that tests the data and puts the system obtained from the selection activity into action. The author will then put the tools designed to solve the problems that have been identified into action once the series has been successfully created [40],[41],[42].

### 1.4. C4.5 Algorithm

Since the C4.5 algorithm is required to generate the decision tree, the two models—the C4.5 algorithm and the decision tree—are inseparable. By using the C4.5 algorithm, you can create a Decision Tree in several stages, including:

1. Compilation of information contains information and reality that has occurred and has been arranged into certain classes.
2. The Gain value of the selected attribute will be used to determine the roots of the root tree; the attribute with the highest Gain value will be the first root. Use the formula in (1) to determine the gain [43]

### 1.5. K-Means

K-means clustering is the method used in this study. A non-hierarchical data grouping technique known as K-means clustering is used to group data into one or more clusters or groups.

### 1.6. Sale

All company activities end with a sale. Assistance will advance the meaning of the agreement approved by the authorities. What is meant by "sale" is an amount of money charged to the buyer for the sale of goods and services, both on credit and in cash, called sales [44],[45].

## 2. METHOD

Flowchart of the method that the author uses, namely the C4.5 and K-Means Algorithm methods. The following are the stages of the C4.5 Algorithm in Fig. 2, starting from the first, which is data collection, then calculating the entropy value, then calculating the profit value, calculating the value of each separate info, calculating the profit ratio, creating a branch for each value, repeating each process until all partitioned nodes.

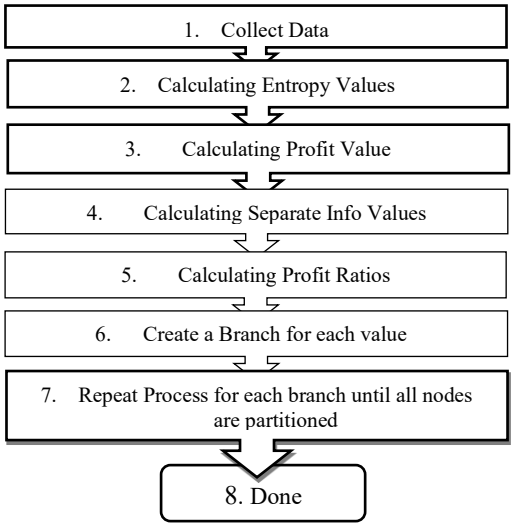


**Fig. 2.** C4.5 Algorithm Stages

The following are the stages of calculation in the Fig. 3. K-Means algorithm. Starting from the start then enter each data that has been transformed, determine the number of each cluster, determine the cluster at the center point, then calculate the data distance to the cluster center, group the data based on the minimum distance to the cluster center, then the center cluster will be obtained, then return to the calculation determine the center point until the same value is obtained, then it's done.

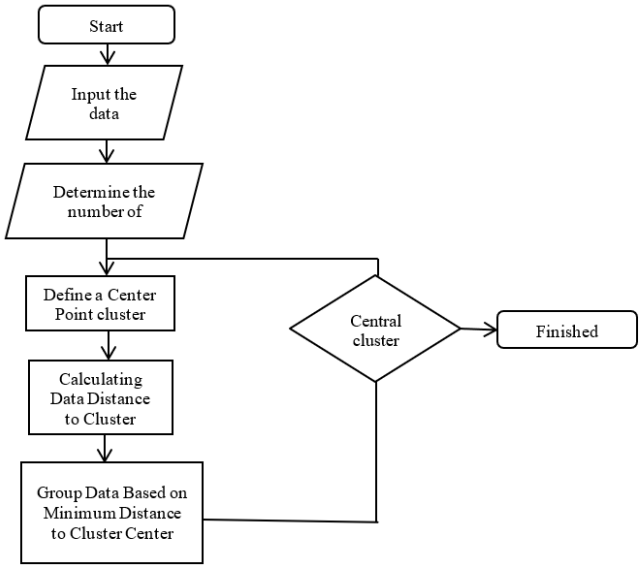


**Fig. 3**. K-Means Algorithm Flowchart.

### 2.1. Research Resources

The research source that the author took came from the South Abung Candimas area. Where the author takes the example of the Xena Cell Counter. The main reason we took this research is because of a very strategic location in urban areas and get findings that are more in line with the problem of quota sales.

### 2.2. Sample population

Sales data from January to December 2019-2022 constitutes the data sample population. There are 240 testing data used in this research process. The goal is to get precise and accurate results.

### 2.3. Data collection technique

Until here, the author obtained sales data for 2019-2022 directly from the Xena Cell Counter, Abung Selatan Candimas in Kotabumi, North Lampung. We use observation, interview, and literature study techniques. Starting from the observation technique, namely visiting all quota sellers in the Abung Selatan Candimas area and checking every available location. Next, we conducted an interview with the counter owner and asked several questions relevant to our research. Finally, conducting a literature study from journals, books, and problems that existed at the time of the research.

### 2.4. Mathematical Formulation

#### 2.4.1. C4.5 Algorithm

The steps for determining the Entropy and Gain values for each criterion with High and Low information are listed. The C4.5 algorithm is intended to help classify vehicle test results according to the factors that influence them [29],[30]. By using the C4.5 (Decision Tree) method and data mining techniques, the author can make the right decisions to predict sales quotas that meet the needs of outlet owners and help sales [9],[31].

**Entropy Calculation**

The initial step of the C4.5 algorithm is to find the entropy value. First determine the total Entropy value in the case. With the following formula:

$$Entropy\ (S) = i = 1n - \ pi * \log_2 pi \tag{1}$$

$S$ is set of cases, $A$ is attribute, $n$ is number of partitions $S$, $pi$ is the proportion of $Si$ to $S$. $S$ is a series of case sets to be studied ranging from simple to complex cases, $A$ is the required attribute in each case, $n$ is the number of $S$ partitions and $pi$ is the proportion of $Si$ to $S$.

The next step after calculating the entropy value is calculating the gain value to determine the root of the decision tree decision tree with the following formula:

$$Gain\ S, A = entropyS - in = 1SiS \tag{2}$$

$S$ is set of cases, $A$ is attribute, $n$ is the number of partitions attribute $A$, $|Si|$ is the number of cases on the i-th partition, $|S|$ is number of cases in $S$. $S$ is a series of case sets to be studied ranging from simple to complex cases, $A$ is the required attribute in each case, n is the number of attribute $A$ partitions, $|\boldsymbol{Si}|$ is the number of cases on each node's I-th partition, and $|S|$ is the number of each case in $S$.

**Determine the Decision Tree (decision tree)**

Explanation of making a decision tree is done after calculating the entropy and gain, namely in Microsoft Excel. So the results are obtained is 7 columns that explain nodes, amount, selling and not selling, entropy and gain values. The highest value lies with Telkomsel, namely 0.276434. This value tends to be greater than the others, although the results are not necessarily true. But seen from the acquisition of Entropy and gain values can be the basis of this research to equate the precision and final destination of the numbers that will be needed by the author. The author will provide a more detailed explanation at the time of making the decision tree. How to determine a decision tree can be started by looking for Entropy and Gain values. Furthermore, it is determined that all branches of the node can be properly partitioned. After that, make a decision tree branch by either using an insert shape or also using a vector. The process of finding the entropy and gain values first to determine the decision tree (Table 1).

**Table 1.** Entropy and Gains

| knot | Amount | Collectability | | Entropy | Gains Information |
|---|---|---|---|---|---|
| | | Bestseller | Not Selling | | |
| Total Sell/Not | 3 | 1 | 2 | 0.276434 | |
| A1Axis | | | | | 0 |
| PAXIS 1.5GB | 0 | 0 | 0 | 0 | |
| v-axis 4 mini 5D | 0 | 0 | 0 | 0 | |
| v axis 3gb a month | 0 | 0 | 0 | 0 | |
| Vaxis 5gb a month | 3 | 1 | 2 | 0.276434 | |
| A2XL | | | | | 0 |
| PXL 3GB | 0 | 0 | 0 | 0 | |
| VXL L16 | 0 | 0 | 0 | 0 | |
| VXL M8 | 0 | 0 | 0 | 0 | |
| VXL 3GB-8GB MINI | 3 | 1 | 2 | 0.276434 | |
| A3Telkomsel | | | | | 0.276434 |
| P-TSEL 3GB | 1 | 0 | 1 | 0 | |
| V-TSEL 2GB 5D | 1 | 0 | 1 | 0 | |
| V-TSEL 3GB 7D | 1 | 1 | 1 | 0 | |
| A4Indosat | | | | | 0 |
| PM3 2GB | 3 | 1 | 2 | 0.276434 | |

After the elaboration of making entropy and gain tables is done, then we enter them into the decision tree consisting of A1 axis, A3 Telkomsel. Between these two variables in Fig. 4. can be used as a determinant of goods that are clearly in demand and not in demand. The explanation from axis are Paxis 1,5 gb best-selling, v axis 4 mini 5D best-selling, and v axis 3 gb month not selling. Next telkomsel PTSEL 3GB not selling, V-TSEL 2 GB not selling, and V-TSEL 3GB 7D best-selling.

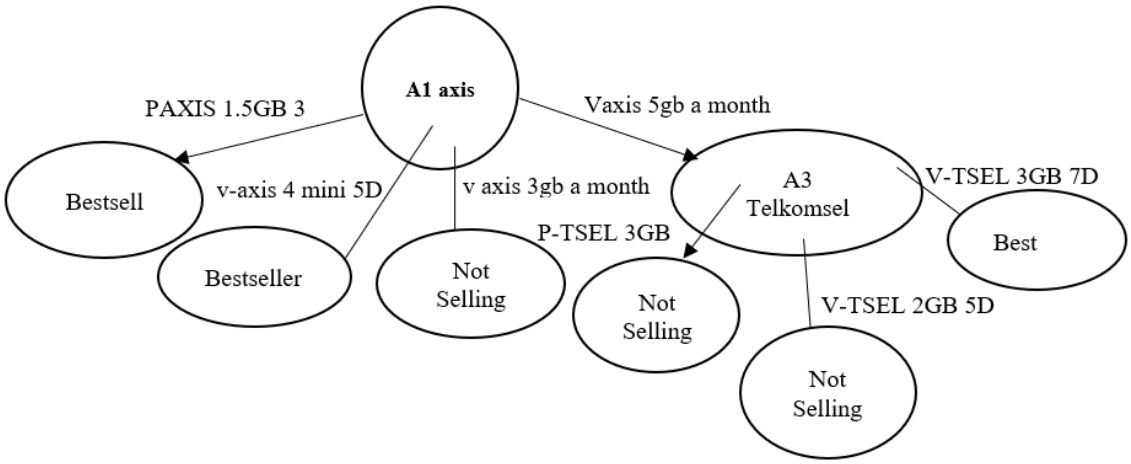


**Fig. 4.** Decision Tree

**Accuracy**

Following the calculation of the entropy value, the gain value is calculated to determine the root of the decision tree using the formula (3)-(5).

- Accuracy Percentage

Accuracy percentage is a form of statement using a large data model divided by the number of predictions from the testing data multiplied by 100% accuracy. Formula (3) is an overview of the formulas in the accuracy presentation.

$$= The\ data\ on\ the\ number\ of\ correctly\ predicted\ outcomes The\ number\ of\ predictions\ made \times 100\% \quad (3)$$

Next, we enter the formula used to determine the percentage of accurate processed data as follows:

- Percentage of Accuracy

The amount of data in the correct table from the testing data is divided by the predicted data table multiplied by 100% as shown (4).

$$= Number\ of\ Correct\ Predicted\ Data\ Tabel\ Table\ prediction \times 100\% \quad (4)$$

Calculation of Quota Sales Data with a total of 240 data items sold based on data samples in 4 years and the presentation accuracy is as follows:

$$Percentage\ of\ Accuracy = 237240 \times 100\% = 98{,}75\% \quad (5)$$

The percentage of accuracy is 237 correct data divided by 240 the number of predicted data multiplied by 100%, the result is 98.75%. Where this figure already shows that the accuracy obtained during manual calculations is the way it is. This shows that the calculation of the C4.5 algorithm is included in an almost perfect calculation.

Furthermore, the calculation uses a confused table where in Table 2 shows that there are 3 columns where column 1 contains predictions of the best or unsold sales, the second column has 156 best sales and 1 does not sell, then the third column does not sell 3 and does not sell 80.

**Table 2.** Confusion Table

| | **Class** | |
|---|---|---|
| predictions | **Best Selling** | **Not Selling** |
| Best selling | 156 | 3 |
| Not Selling | 1 | 80 |

Based on the calculation on the training data, the feasibility accuracy is 98.75% and it is known that 156 best-selling classes and classified as best-selling Predictions are not selling as much as 1 is classified as best-selling but not sold, then the non-selling class is group 80 classified as not selling, and 3 selling is classified as not selling with a total of 240 Quota Sales Data 2021 and the most influential features affecting sales. it can be interpreted that from testing the Data Testing and Training "XL" criteria can affect the results of Sales Quota Sales in determining Best Selling or Not Sold to determine the addition of the number of quotas.

### 2.4.2. K-Means

In the conventional k-means algorithm, k cluster centers are initialized randomly in space, the distance between each point and the cluster center is calculated, and the location of the cluster center is updated repeatedly until the best cluster center is found [14].

To achieve the desired data grouping results, the sample data will undergo a grouping procedure using the K-Means algorithm. The steps of the K-means algorithm clustering are as follows:

1. Determine the number of clusters for quota sales data in the Xena cell counter, namely the clusters used in:

C1 = Best selling product
C2 = Hot sale product
C3 = The product is not selling

2. Find out the center of mass first. The initial center of the cluster, or centroid, can be chosen randomly or from previously collected data. Cluster 1 values come from the second row, Cluster 2 values come from the 119th row, and Cluster 3 values come from the 239th row.
3. Data sales quota.

The explanation in Table 3 is that there are 10 sample data taken based on actual data that has gone through a data transformation process. There are 7 different columns in each attribute. There are monthly columns, axis cards, xl cards, Telkomsel cards, Indosat cards, initial stock and sold columns. Each attribute has 4 different types of data quota. This data will be processed by the author in Microsoft Excel and then the application will use Colab.Research.Google.

The explanation in Table 4 is that there are 10 sample data taken based on actual data that has gone through a data transformation process. There are 7 different columns in each attribute. There are month

columns, axis cards, xl cards, Telkomsel cards, Indosat cards, C1 and C2 calculations. Each attribute has 4 different types of data quota. This data will be processed by the author based on the K-means formula which has been attached to this paper.

**Table 3.** Xena Cell counter quota data

| NO | Month | Axis | XL | Telkomsel | Indosat | First stock | Sold |
|---|---|---|---|---|---|---|---|
| 1 | December | v axis 4 mini 5D | VXL 3GB MINI | V-TSEL 3GB 7D | PM3 2GB | 250 | 250 |
| 2 | December | axis v 1.5gb Mini 3D | VXL M8 | V-TSEL 3GB 7D | PM3 2GB | 250 | 200 |
| 3 | may | v axis 1.5gb a month | VXL L16 | V-TSEL 3GB 7D | PM3 2GB | 300 | 200 |
| 4 | November | v axis 1.5gb a month | VXL L16 | V-TSEL 3GB 7D | PM3 2GB | 250 | 200 |
| 5 | November | v-axis 4 mini 5D | VXL 3GB MINI | V-TSEL 3GB 7D | PM3 2GB | 200 | 150 |
| 6 | December | v axis 3gb a month | VXL 4GB MINI | P-TSEL 3GB | PM3 2GB | 200 | 150 |
| 7 | August | v-axis 4 mini 5D | VXL 3GB MINI | V-TSEL 3GB 7D | PM3 2GB | 150 | 120 |
| 8 | November | axis v 1.5gb Mini 3D | VXL M8 | V-TSEL 3GB 7D | PM3 2GB | 200 | 100 |
| 9 | November | v-axis 4 mini 5D | VXL 3GB MINI | P-TSEL 3GB | PM3 2GB | 150 | 100 |
| 10 | December | Vaxis 5Gb Mini | VXL MINI 14GB | V-TSEL 2GB 5D | PM3 2GB | 250 | 100 |

**Table 4.** Calculation of K-Means

| NO | month | Axis | XL | Telkomsel | Indosat | C1 | C2 |
|---|---|---|---|---|---|---|---|
| 1 | December | v-axis 4 mini 5D | VXL 3GB MINI | V-TSEL 3GB 7D | PM3 2GB | 126.40.00 | 312.92 |
| 2 | December | axis v 1.5gb Mini 3D | VXL M8 | V-TSEL 3GB 7D | PM3 2GB | 81515 | 278.86 |
| 3 | may | v axis 1.5gb a month | VXL L16 | V-TSEL 3GB 7D | PM3 2GB | 116.57.00 | 319.13.00 |
| 4 | November | v axis 1.5gb a month | VXL L16 | V-TSEL 3GB 7D | PM3 2GB | 81515 | 278.86 |
| 5 | November | v-axis 4 mini 5D | VXL 3GB MINI | V-TSEL 3GB 7D | PM3 2GB | 19.156 | 208.62 |
| 6 | December | v axis 3gb a month | VXL 4GB MINI | P-TSEL 3GB | PM3 2GB | 19.156 | 208.62 |
| 7 | August | v-axis 4 mini 5D | VXL 3GB MINI | V-TSEL 3GB 7D | PM3 2GB | 56,767 | 150.81 |
| 8 | November | axis v 1.5gb Mini 3D | VXL M8 | V-TSEL 3GB 7D | PM3 2GB | 32,150 | 182.65 |
| 9 | November | v-axis 4 mini 5D | VXL 3GB MINI | P-TSEL 3GB | PM3 2GB | 63,946 | 138.85 |
| 10 | December | Vaxis 5Gb Mini | VXL MINI 14GB | V-TSEL 2GB 5D | PM3 2GB | 54,571 | 229.01.00 |

The explanation in Table 5 is that there are 3 centeroid literacy centers taken based on the K-Means cluster centers, namely Cluster 1, Cluster 2, and Cluster 3. In the first stock in cluster 1 there were 250 initial items and 200 data were sold. In cluster 2 there were 20 initial data and 10 data were sold. In cluster 3 there are 5 initial data and 1 data sold.

**Table 5.** Early Literacy Centeroid Point 1

| C | First stock | Sold Products |
|---|---|---|
| CLUSTER 1 | 250 | 200 |
| CLUSTER 2 | 20 | 10 |
| CLUSTER 3 | 5 | 1 |

Using the Euclidean distance, determine the distance from the centroid and the distance between the centroid points and those of each object. Regarding the initial manual centroid calculation. The following is thecalculation:

$$D(,) = xy\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (3)$$

Explanation D, symbolizes the main dimension to be searched for. $xy$ are variables 1 and 2 and some are used to determine dimensions based on existing formulas. The calculation will continue according to the existing formula until it meets the same centeroid number.

**literacy 1**

Members have chosen from the smallest of the three Clusters: if the smallest cluster is in section C1, it is included as a member of C1 with up to 18 data, if the smallest clustering is in section C2, it has included as a member of C2 with up to 127 data, and if the smallest cluster is in section C3, it has included as a member of C3 with up to 95 data.

**literacy 2**

Members have chosen from the three smaller Clusters: if the smallest cluster is in section C1, it has been included as a member of C1 with as many as 19 data. If the smallest clustering is in section C2, it has been included as a member of C2 with as many as 68 data, and if the smallest cluster is in section C3, it has been included as a member of C3 with as many as 153 data.

It is necessary to re-do the calculations for the third iteration so that until you get the same results because the results of the second iteration are not the same as those of the first iteration.

**literacy 3**

Members are chosen from the smallest of the three Clusters. If the smallest cluster is in section C1, it is included as a member of C1, which has up to 19 data, if the smallest cluster is in section C2, it is included as a member of C2, which has up to 54 data, and if the smallest cluster is in section C3, it is included as a member of C3, which has up to 167 data.

It is necessary to re-do the calculations for the fourth iteration and so on until you get the same results because the results of the third iteration are not the same as those of the second iteration.

**literacy 4**

Members have chosen from the smallest of the three clusters. If the smallest cluster is in section C1, it is included as a member of C1, which contains 19 data; if the smallest clustering is in section C2, it has included as a member of C2, which contains 44 data; and if the smallest cluster is in section C3, it has included as a member of C3, which contains 177 data.

Because the results of the fourth iteration are not the same as those of the third iteration, you will need to repeat the calculations for the fifth iteration until you reach the same conclusion.

**literacy 5**

Members have chosen from the smallest of the three clusters. If the smallest cluster is in section C1, it has included as a member of C1, which contains 19 data; if the smallest clustering is in section C2, it has included as a member of C2, which contains 43 data; and if the smallest cluster is in section C3, it is including as a member of C3, which contains 178 data.

Because the results of the fifth iteration are not the same as those of the fourth iteration, you will need to repeat the calculations for the sixth iteration until you reach the same conclusion.

**literacy 6**

Members have chosen from the smallest of the three clusters. If the smallest cluster is in section C1, it has included as a member of C1, which contains 19 data; if the smallest clustering is in section C2, it has included as a member of C2, which contains 43 data; and if the smallest cluster is in section C3, it has included as a member of C3, which contains 178 data.

There is no need to proceed to the seventh literacy or simply stop at the sixth literacy because the results of the sixth literacy and the fifth literacy are identical.

## 3. RESULTS AND DISCUSSION

### 3.1. C4.5 Algorithm

We can see that the collab.research.google application's numbers appear to be selected, ranging from 0 to 9. From lowest to highest, this is the order. So that the number sequence looks like 1, 2, 3, etc. contract, each number in the Python programming language follows the command line. There are 9 columns consisting of index, number, month, axis, xl, telkomsel, indosat, sold, and description. Everything is still in the form of real data and has not been transformed. After that, provide a numerical value for each imported data as the class. Display data shown in Fig. 5.

The trick is to show the description of the attribute in the code in Fig. 6, there is a dataset class "description".

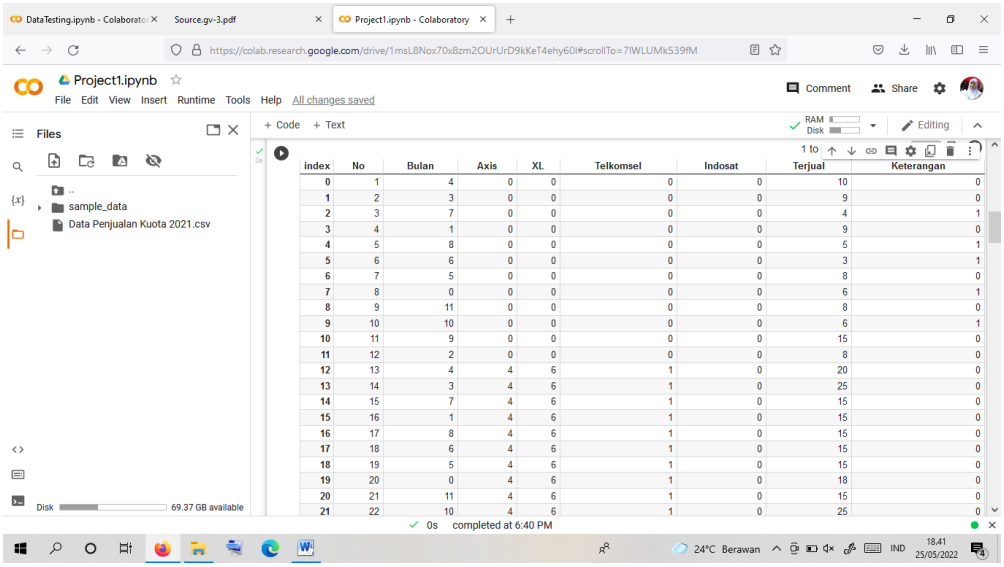


**Fig. 5.** Display data

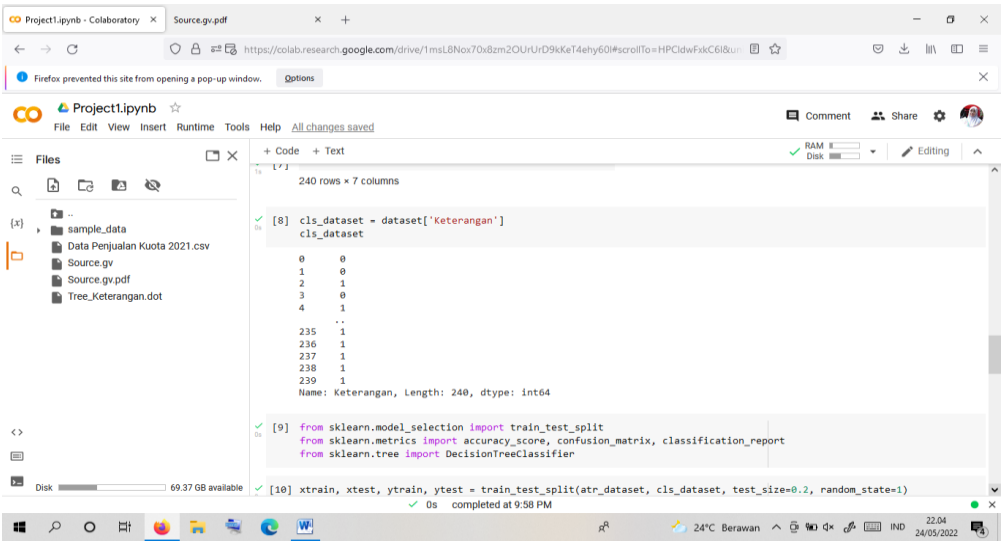


**Fig. 6.** Classification

The "remarks" class was created to help researchers determine the main benchmarks in this study. Best Selling and Not Selling are the two attributes in the class description that will be displayed in the knowledge tree at the picture's conclusion.

Enter the code in Fig. 7 automatically perform calculations based on the original C4.5 algorithm formula, which can be understood that each code is instructed to make predictions on the confusion matrix and determine the number of decisions at the end.

Prediction is in the form of a decision tree, then a confusion matrix appears, showing the level of accuracy of the C4.5 algorithm, then Accuracy, and displays accuracy in percent form. Pay attention to the confusion matrix in the picture explaining the numbers [31,0], [0,17] which indicate that there are 31 selling data and 17 unselling sample data that are only understood by machine language or python language.

After that type, the code in Fig. 8 determine the knowledge tree image using python language. Next, ensure that each accuracy is an overall quotient that includes the entire formula or formula. Each formula is based on the confusion matrix table, the accuracy of the sample data where only 31 correct data and 17 wrong data are used.

After finding correct and incorrect data, accuracy will be determined to reach a limit of 100% if it is reached. But this is not a significant problem because the data that the author processes is real data that can be implemented in the sale of quota case studies. So that the information that can be captured is a real result in the

calculation of the C4.5 algorithm formula. This method can minimize the error rate or manual method calculations.

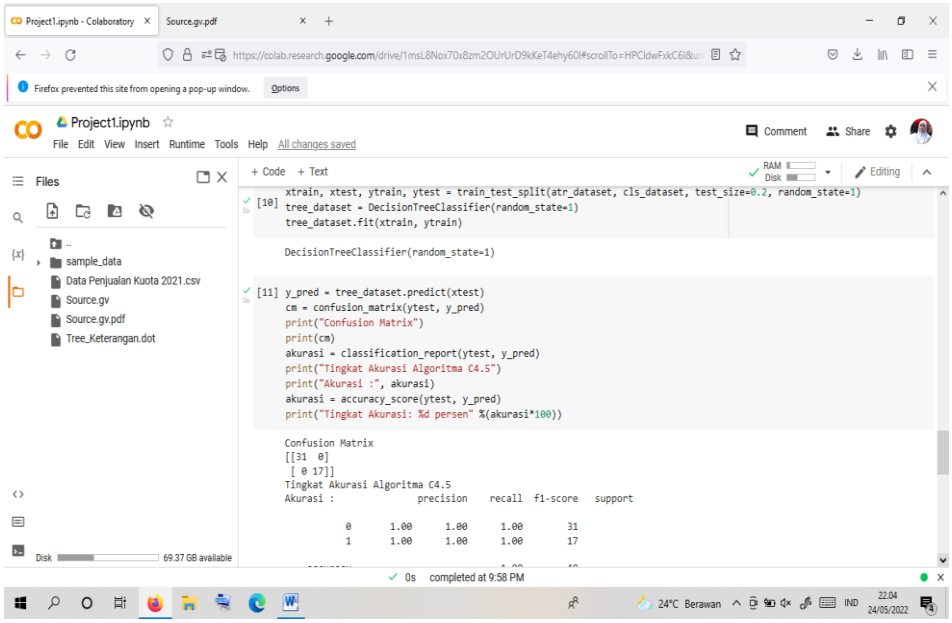


**Fig. 7.** Formulas

Fig. 8 is an import graphic that will show an image of a decision tree that describes the findings in the form of a statistical picture of the best-selling and not-selling decision tree branches in the image Fig. 8. The graph is named "tree_keterangan_dot". The name can be changed as needed but must be the same as when determining the confusion marix so as not to get confusing and misleading results. Please pay attention to the picture in Fig. 8 so that it can be used as the final result of the research that the researcher has done.

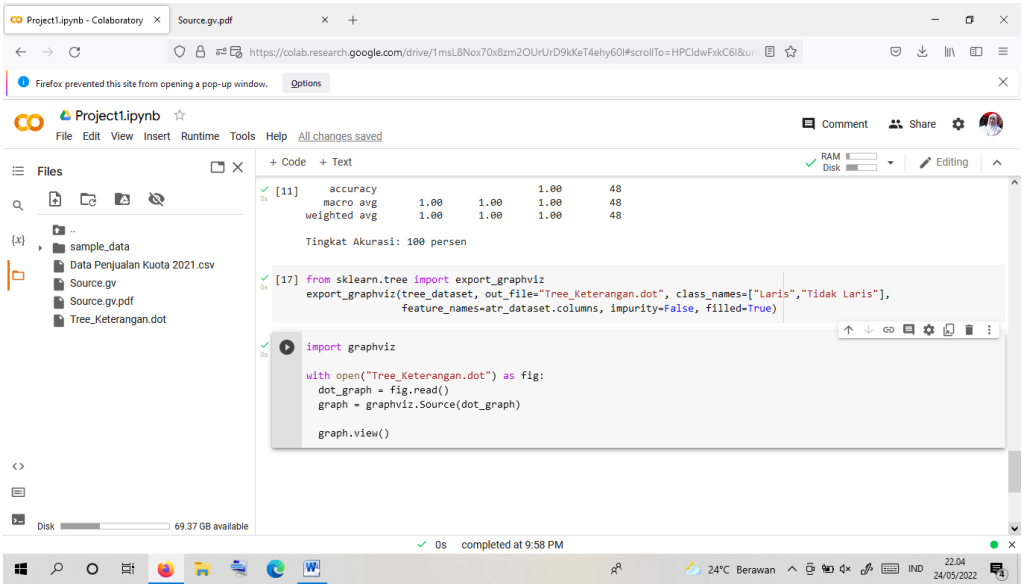


**Fig. 8.** Display image

From the code in Fig. 8, a pdf file will appear, and the file that I created is called Tree_Keterangan.dot indicates the best seller/not a best seller for the quota sold at the Xena Cell Counter with 100% purity. Fig. 9 is a presentation of the knowledge tree from the python language in the collab.research.google app.

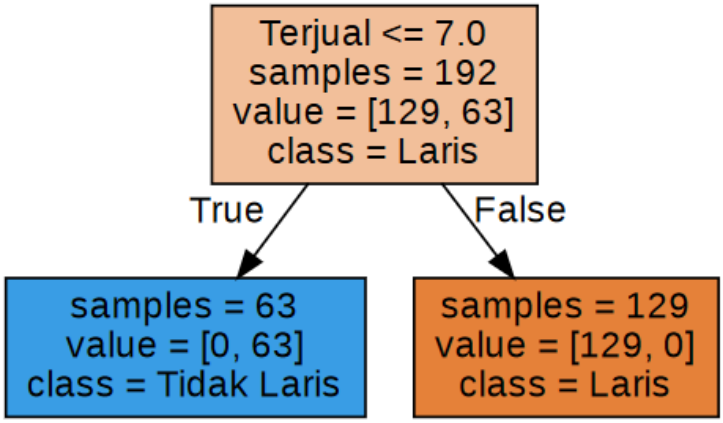


**Fig. 9.** Decision Tree

The data that we have processed will be presented through the knowledge tree in the C4.5 algorithm. Items sold have listed as less than equal to 7 (<= 7.0) samples that have automatically connected from the collab.research.google is 192 data out of 240 data. The 48th data has undefined in the sense that the Python language performs data sampling starting from 0, so the 48th data is defined as 0. Then it will be concluded that 63 samples are not selling well or 129 that are selling well. This means that the sales quota at the Xena Cell Counter has a fairly high selling rate. And calculations in the collab.research.google application using python has an accuracy of up to 100%.

### 3.2. K-Means

Fig. 9 shows the point of distribution of quota sales starting from the number 0 to the very end, namely the number 240 where each point determines the sales that follow the graph increasing every year that is divided into 6 clusters according to calculations using Microsoft Excel. K-Means graphic image shown in Fig. 10. K-Means Graphic Image shown in Fig. 11.

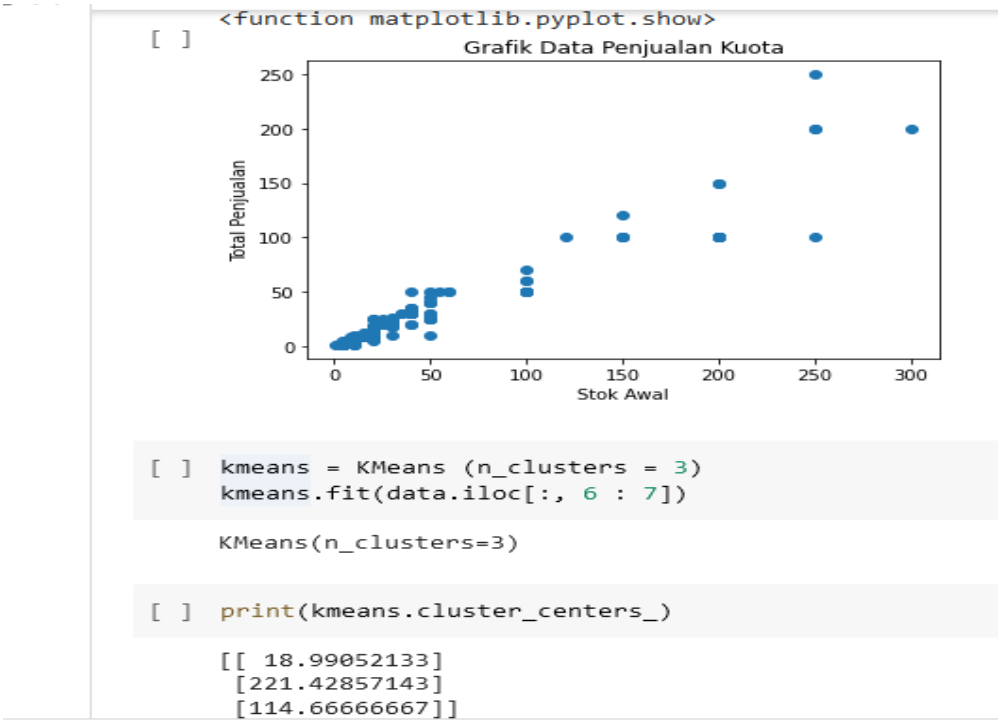


**Fig. 10.** K-Means graphic image

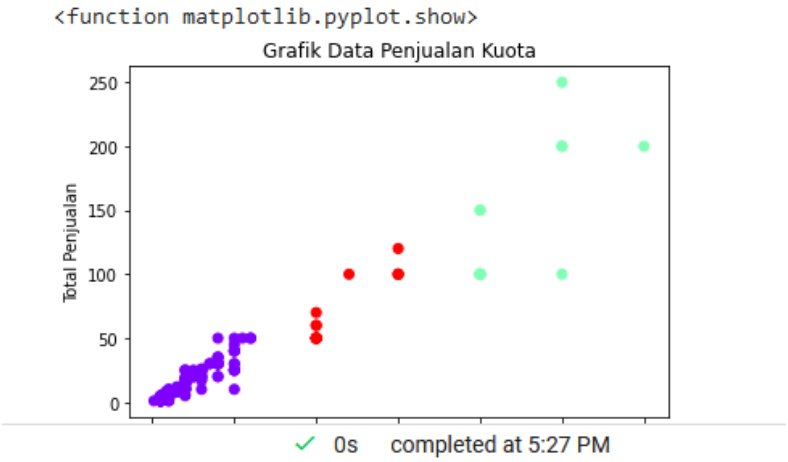


**Fig. 11.** K-Means Graphic Image

The graph that the class has been divided into 3, namely Best Selling, Hot Best Selling, and less-selling. The purple color indicates that the quota is not in demand, the red color is the best seller, and the blue color indicates the best seller so that these results can help the counter owner determine the next sale.

### 3.3. Algorithm Comparison of C4.5 and K-Means

From the explanation and also the calculations, the writer can analyze the findings in the form of real accuracy and truth as shown in the Table 6. Where there are 4 columns which explain that there are data sample years from 2019 to 2022. Next there is a classification from the C4.5 algorithm which states the accuracy in manual calculations, namely in 2019-2020 there are 58% best-selling and 42% not selling, but there are also in 2021-2022 the quota has sold 100% best-selling. After that, in the K-Means clustering there is an explanation of some real data where 2019-2022 has the same sales, namely 19 best-selling, 43 selling-well, and 178 less-selling. Finally, there is the use of Colab.Research.Google where every accuracy from 2019-2022 has a best-selling accuracy of 100%.

**Table 6.** The Results of Comparative

| year | Algorithm Classification C4.5 | K-Means Clustering | Use Google Research Colab |
|---|---|---|---|
| 2019 | produce best-selling healing of 58% and not selling 42%. In the C4.5 algorithm, it is only determined that it is selling and not selling. | Producing the best-selling figures of 19 types of quotas, 43 best-selling members of the quota, and 178 less-selling data. In K-means, only 3 cluster centers were determined, namely very in demand, less in demand, and not in demand. | Produces a best-selling Accuracy of 100% |
| 2020 | produce best-selling healing of 58% and not selling 42%. only determined selling and not selling. | Producing the best-selling figures of 19 types of quotas, 43 best-selling members of the quota, and 178 less-selling data. In K-means, only 3 cluster centers were determined, namely very in demand, less in demand, and not in demand. | |
| 2021-2022 | Produces a best-selling Accuracy of 100%. only determined selling and not selling. | Producing the best-selling figures of 19 types of quotas, 43 best-selling members of the quota, and 178 less-selling data. In K-means, only 3 cluster centers were determined, namely very in demand, less in demand, and not in demand. | |

From the equation, the authors can describe that there is a difference between the results of the C4.5 algorithm and K-Means. Where the presentation produced is clearly very different where in 2019 around 58% were in demand and 42% were not in demand in the C4.5 algorithm. then on K-means almost 178 unsold data quota. How did it happen? This is because the data obtained determines a significant difference seen from the

depth of data mining. Because classification is certainly much different from clustering. Classification determines the attributes used while clustering classifies data according to demand.

In 2019 it has the same accuracy, namely 58% selling and 42% not selling, while 2020 has the same accuracy 58% selling and 42% not selling either. In the same year, K-means also had 178 unsolicited data, 19 very salable data, and 43 unsolicited data. In 2021-2022 the classification produces an accuracy of up to 100% where the best-selling data dominates while clustering remains at the same number, namely 19 highly salable data, 43 less salable data, and 178 unsolicited data. However, after searching using the application, all sales have 100% accuracy. This research is real and not manipulated so it can be concluded that the difference is that K-means has very perfect accuracy because it is able to reach data that has not been detected as unsold.

### 3.4. Discussion of Implementation Results of colab.research.google

Based on the results obtained from the results of the Prediction Data Training test, which totals 240 data, it shows that the calculation of the C4.5 Algorithm and K-means method using the collab.research.google application has an accuracy of 100% and the "Axis & XL" criteria can affect the results of Sales success in the Xena Cell Counter in determining the Selling or Not Selling quota sales quota.

## 4. CONCLUSION

After making several calculations from the C4.5 and K-Means algorithms, a score that has a high accuracy of up to 100% is obtained. The picture shows that the number of unsold products in the K-Means method is higher than in the C4.5 algorithm. In the C4.5 algorithm, data is centered when calculating gain and entropy. All of these calculations obtain very high accuracy, namely products that are selling higher than those that are not selling well.

Therefore, this research is one way to determine the target for the promotion of goods that are not in demand so that the calculation can be used as a reference for making decisions in the future. The counter owner makes improvements to his marketing department so that the sales quota can increase as expected. Furthermore, this research will help customers get the quota as expected and have the best-selling price so that customers feel satisfied in terms of service and competitive prices.

The researcher realizes that this research is still limited and we suggest that further research be carried out even better. This research only focuses on one area, namely Abung Selatan Candimas, precisely at the xena cell counter. From these findings can be generalized to counters or other stores. In addition, other factors that can affect sales performance are demographics where the closer to residential areas, the sales are increasingly in demand according to customer preferences. These findings can also be implemented in other areas. Therefore, future research can broaden the scope of the study to include a larger sample size.

## Acknowledgments

We are authors from this article would say thank you very much to ITBA-PSDKU Dian Cipta Cendikia Kotabumi who supported us in arranging the best article

## REFERENCES

[1] M. Moutaib, T. Ahajjam, M. Fattah, Y. Farhaoui, B. Aghoutane, and M. El Bekkali, "Application of Internet of Things in the Health Sector: Toward Minimizing Energy Consumption," *Big Data Mining and Analytics*, vol. 5, no. 4, pp. 302–308, 2022, https://doi.org/10.26599/BDMA.2021.9020031.

[2] M. Azrour, J. Mabrouki, A. Guezzaz, and Y. Farhaoui, "New Enhanced Authentication Protocol for Internet of Things," *Big Data Mining and Analytics,* vol. 4, no. 1, pp. 1–9, 2021, https://doi.org/10.26599/BDMA.2020.9020010.

[3] P. Ehin, M. Solvak, J. Willemson, and P. Vinkel, "Internet voting in Estonia 2005 – 2019: Evidence from eleven elections," *Gov. inf. Q*, vol. 39, no. 4, p. 101718, 2022, https://doi.org/10.1016/j.giq.2022.101718.

[4] PBM Fahim*et al.*, "On the evolution of maritime ports towards the Physical Internet," *Futures*, vol. 134, p. 102834, 2021, https://doi.org/10.1016/j.futures.2021.102834.

[5] S. Zhang, C. Bi, M. Zhang, S. Zhang, C. Bi, and M. Zhang, "Logistics service supply chain allocationmixed K-Means and Qos order matching Qos matching," *Procedia Comput. Sc*., vol. 188, 2019, pp. 121–129, 2021, https://doi.org/10.1016/j.procs.2021.05.060.

[6] C. K. Lo, H. C. Chen, P. Y. Lee, M. C. Ku, L Ogiela, and C. H. Chuang, "Smart Dynamic Resource Allocation Model for Patient-Driven Mobile Medical Information System Using C4. 5 Algorithm," *J.Electron. sci. Technol.*, vol. 17, no. 3, pp. 231–241, 2019, https://doi.org/ 10.11989/JEST.1674-862X.71018117.

[7] S. Lee, Z. Xu, T. Li, and Y. Yang, "A bagging novel C4. 5 algorithm based on wrapper feature selection for supporting wise clinical decision making," *J. Biomed. inform.*, vol. 78, pp. 144–155, 2018, https://doi.org/10.1016/j.jbi.2017.11.005.

[8] K. R. Pradeep and N. C. Naveen, "Lung Cancer Survivability Prediction based on Performance Using Lung Cancer Survivability on Performance Using Classification Techniques Prediction of Support based Vector Machines, C4.5

and Naive Bayes Algorithms for Healthcare Analytics," *Procedia Comput. sci.*, vol. 132, pp. 412–420, 2018, https://doi.org/ 10.1016/j.procs.2018.05.162.
[9] C. Rajeswari, B. Sathiyabhama, S. Devendiran, and K. Manivannan, "A Gear fault identification using wavelet transform, rough set based GA, ANN and C4. 5 algorithm," *Procedia Eng.*, vol. 97, pp. 1831–1841, 2014, https://doi.org/ 10.1016/j.proeng.2014.12.337.
[10] I. Satisfaction, P. Informa, A. Febriyani, GK Prayoga, and O. Nurdiawan, "Information Customer Satisfaction Index Using the C.45 Algorithm," vol. 8, no. 6, pp. 330–335, 2021, https://doi.org/ 10.30865/jurikom.v8i6.3686.
[11] A. P. Muniyandi, R. Rajeswari, and R. Rajaram, "Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm," *Procedia Engineering*, vol. 30, no. 2011, pp. 174–182, 2012, https://doi.org/ 10.1016/j.proeng.2012.01.849.
[12] K. Bouchard, B. Bouchard, and A. Bouzouane, "A new qualitative spatial recognition model based on Egenhofer topological approach using C4. 5 algorithms: experiments and results," *Procedia Computer Science,* vol. 5, pp. 497–504, 2011, https://doi.org/10.1016/j.procs.2011.07.064.
[13] X. Wang, C. Zhou, X. Wang, C. Zhou, and X. Xu, "Application of C4.5 decision tree for scholarship evaluations," *Procedia Computer Science*, vol. 151, pp. 179-184, 2019, https://doi.org/10.1016/j.procs.2019.04.027.
[14] S. K. Maurya, X. Liu, T. Murata, "Feature selection: Key to enhance node classification with graph neural networks," *The Institution of Engineering and Technology,* vol. 8, no. 1, pp. 14–28, 2023, https://doi.org/ 10.1049/cit2.12166.
[15] B. N. Lakshmi, T. S. Indumathi, and N. Ravi, "A Study on C.5 Decision Tree Classification Algorithm for Risk Predictions During Pregnancy," *Procedia Technol.*, vol. 24, pp. 1542–1549, 2016, https://doi.org/ 10.1016/j.protcy.2016.05.128.
[16] C. Shi, D. Liao, T. Zhang, and L. Wang, "Hyperspectral image classification based on 3D coordination attention mechanism network," *Remote Sensing*, vol. 14, no. 3, p. 608, 2022, https://doi.org/10.3390/rs14030608.
[17] S. Ubukata, A. Notsu, and K. Honda, "Objective function-based rough membership C-means clustering," *Information Sciences,* vol. 548, pp. 479–496, 2021, https://doi.org/ 10.1016/j.ins.2020.10.037.
[18] Y. Hozumi, R. Wang, C. Yin, and G. Wei, "UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets," *Comput. Bio. med.*, vol. 131, p. 104264, 2021, https://doi.org/10.1016/j.compbiomed.2021.104264.
[19] F. V. Fomin, P. A. Golovach, and K. Simonov, "Parameterized k-Clustering: Tractability island," *Journal of Computer and System Sciences,* vol. 117, pp. 50–74, 2021, https://doi.org/10.1016/j.jcss.2020.10.005.
[20] T. Zhang and G. Lin, "Generalized k-means in GLMs with applications to the outbreak of COVID-19 in the United States," *Comput. Stats. Anal Data.*, vol. 159, p. 107217, 2021, https://doi.org/10.1016/j.csda.2021.107217.
[21] G. Niu, Y. Ji, Z. Zhang, W. Wang, J. Chen, and P. Yu, "Clustering analysis of typical scenarios of island power supply system by using cohesive hierarchical clustering based on K-Means clustering method," *Energy Reports,* vol. 7, pp. 250–256, 2021, https://doi.org/10.1016/j.egyr.2021.08.049.
[22] C. Chandrashekar, P. Agrawal, P. Chatterjee, and D. S. Pawar, "Development of E-rickshaw driving cycle (ERDC) based on micro-trip segments using random selection and K-means clustering techniques," *IATSS Res.*, vol. 45, no. 4, pp. 551–560, 2021, https://doi.org/10.1016/j.iatssr.2021.07.001.
[23] A. Cebollero-salinas, S. Orejudo, and J. Cano-escoriaza, "Cybergossip and Problematic Internet Use in cyberaggression and cybervictimization among adolescents," *Computers in Human Behavior,* vol. 131, p. 107230, 2022, https://doi.org/10.1016/j.chb.2022.107230.
[24] S. Mullin *et al.*, "Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes," *J. Biomed. Inform.*, vol. 122, p. 103889, 2021, https://doi.org/10.1016/j.jbi.2021.103889.
[25] M. Sekolovska, "Internet business models for e-insurance and conditions in the republic of macedonia," *Procedia-Social and Behavioral Sciences,* vol. 44, pp. 163–168, 2012, https://doi.org/10.1016/j.sbspro.2012.05.016.
[26] D. Das, P. Kayal, and M. Maiti, "A K-means clustering model for analyzing the Bitcoin extreme value return," *Decision Analytics Journal*, vol. 6, p. 100152, 2023, https://doi.org/10.1016/j.dajour.2022.100152.
[27] W. Liu, P. Zou, D. Jiang, X. Quan, and H. Dai, "Zoning of reservoir water temperature field based on K-means clustering algorithm," *J. Hydrol. Reg. Studs.*, vol. 44, p. 101239, 2022, https://doi.org/10.1016/j.ejrh.2022.101239.
[28] G. Dinu and L. Dinu, "Using Internet as a Commercial Tool: a Case Study of E-Commerce in Resita," *Procedia Eng.*, vol. 69, pp. 469–476, 2014, https://doi.org/10.1016/j.proeng.2014.03.014.
[29] D. Liu, F. Yang, and S. Liu, "Estimating wheat fractional vegetation cover using a density peak k-means algorithm based on hyperspectral image data," *J. Integr. Agric.*, vol. 20, no. 11, pp. 2880–2891, 2021, https://doi.org/10.1016/S2095-3119(20)63556-0.
[30] G. Bighiu, A. Manolic, and C. T. Roman, "Compulsive buying behavior on the internet*," Procedia Economics and Finance,* vol. 20, no. 15, pp. 72–79, 2015, https://doi.org/10.1016/S2212-5671(15)00049-0.
[31] D. Liu, F. Yang, and S. Liu, "Estimating wheat fractional vegetation cover using a density peak k-means algorithm based on hyperspectral image data," *J. Integr. Agric.*, vol. 20, no. 11, pp. 2880–2891, 2021, https://doi.org/10.1016/S2095-3119(20)63556-0.
[32] G. Bighiu, A. Manolic, and C. T. Roman, "Compulsive buying behavior on the internet," *Procedia Economics and Finance*, vol. 20, no. 15, pp. 72–79, 2015, https://doi.org/10.1016/S2212-5671(15)00049-0.
[33] S. R. Vadyala, S. N. Betgeri, E. A. Sherer, and A. Amritphale, "Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM," *Array*, vol. 11, p. 100085, 2021, https://doi.org/10.1016/j.array.2021.100085.
[34] N. Yabe, T. Hanibuchi, H. M. Adachi, S. Nagata, and T. Nakaya, "Transportation Research Interdisciplinary Perspectives Relationship between Internet use and out-of-home activities during the first wave of the COVID-19

outbreak in Japan," *Transp. Res. Interdiscip. Perspect.*, vol. 10, p. 100343, 2021, https://doi.org/10.1016/j.trip.2021.100343.
[35] D. Mining, "Guest Editorial: Rough Sets and Data Mining," *CAAI Transactions on Intelligence Technology*, vol. 4, no. 4, pp. 201–202, 2019, https://doi.org/10.1049/trit.2019.0063.
[36] H. Sakai and M. Nakata, "Rough set-based rule generation and Apriori-based rule generation from table data sets: a survey and a combination," *CAAI Transactions on Intelligence Technology*, vol. 4, pp. 203–213, 2019, https://doi.org/10.1049/trit.2019.0001.
[37] H. Sakai and Z. Jian, "Rough set-based rule generation and Apriori-based rule generation from table data sets II: SQL-based environment for rule generation and decision support," *CAAI Transactions on Intelligence Technology,* vol. 4, pp. 214–222, 2019, https://doi.org/10.1049/trit.2019.0016.
[38] W. A. Castillo and C. J. Meneses, "Graphical Representation and Exploratory Visualization for Decision Trees in the KDD Process," *Procedia - Soc. Behav. Sci.*, vol. 73, pp. 136–144, 2013, https://doi.org/10.1016/j.sbspro.2013.02.033.
[39] C. Wagner, C. Wagner, P. Saalmann, and B. Hellingrath, "Machine Condition Monitoring and Fault Diagnostics with Imbalanced Data Sets based on the KDD Process," *IFAC-PapersOnLine*, vol. 49, no. 30, pp. 296–301, 2016, https://doi.org/10.1016/j.ifacol.2016.11.151.
[40] J. D. Smith and M. Hasan, "Quantitative approaches for the evaluation of implementation research studies," *Psychiatry Res.*, vol. 283, p. 112521, 2020, https://doi.org/10.1016/j.psychres.2019.112521.
[41] M. S. Bauer *et al.*, "Implementation science: What is it and why should I care?," *Psychiatry research*, vol. 283, 2020, https://doi.org/10.1016/j.psychres.2019.04.025.
[42] S. Shamsuddin, "Resilience resistance: The challenges and implications of urban resilience implementation," *Cities*, vol. 103, p. 102763, 2020, https://doi.org/10.1016/j.cities.2020.102763.
[43] M. Li, L. Zhao, S. Jin, D. Li, J. Huang, and J. Liu, "Heliyon Process schemes of ethanol coupling to C4 olefins based on a genetic algorithm for back propagation neural network optimization," *Heliyon*, vol. 8, no. September, p. e12301, 2022, https://doi.org/10.1016/j.heliyon.2022.e12301.
[44] J. Chen, X. Feng, G. Kou, and M. Mu, "Multiproduct newsvendor with cross-selling and narrow-bracketing behavior using data mining methods," *Transp. Res. Part E*, vol. 169, p. 102985, 2023, https://doi.org/10.1016/j.tre.2022.102985.
[45] S. Almeria and A. Shipley, "Detection of Cyclospora cayetanensis on bagged pre-cut salad mixes within their shelf-life and after sell by date by the U.S. food and drug administration validated method," *Food Microbiol.*, vol. 98, p. 103802, 2021, https://doi.org/10.1016/j.fm.2021.103802.

## BIOGRAPHY OF AUTHORS

**Ericka Vivian Astuti,** I am currently pursuing a Bachelor of Computer (S1) education at ITBA-Dian Cipta Cendikia Kotabumi and have expected to graduate in 2023. This is my second experience in writing articles after being published in the International Journal of Information Systems and Computer Science (IJISCS). E-mail: eriskavivianastuti@gmail.com.

**Asep Afandi,** Completed his undergraduate studies at Gunadarma University in 2010 in the field of Informatics Systems. Then in 2017 he completed his Master's Degree in Information Systems Management at Gunadarma University. Currently a lecturer at the Faculty of Computer Science and Head of study program based on a certification issued by DIKTI. The Ministry of National Education. E-mail: asepafandi189@gmail.com.

**Dwi Marissa Effendi,** Completed his undergraduate studies at STMIK Dian Cipta Cendikia in 2012 in the field of Information Systems. Then in 2018 she completed her master's degree in Informatics Engineering at IIB Darmajaya. Currently she is a lecturer at the Faculty of Computer Science and Vice Chancellor 3 based on a certificate issued by DIKTI. The Ministry of National Education. E-mail: Dwimarisa89@gmail.com.