Jurnal TAM (Technology Acceptance Model)
Jurnal TAM, Volume 14, Number 2, December 2023
E ISSN: 2579-4221; P ISSN: 2339-1103, pp. 253-258
Accredited SINTA 4 Number 225/E/KPT/2022
https://jurnal.ftikomibn.ac.id/index.php/JurnalTam/index

APPLICATION OF DATA MINING IN PREDICTING THE AMOUNT OF RESTAURANT TAX REVENUE USING C4.5

Roby Afriza[1], Nurmayanti[2], Merri Parida[3], Sidik Rahmatullah[4]
[1,3,4]Program Studi Sistem Informasi, Institut Teknologi Bisnis dan Bahasa Dian Cipta Cendikia
[2]Program Studi Teknologi Komputer, Institut Teknologi Bisnis dan Bahasa Dian Cipta Cendikia
[1,2,3,4]Jl. Negara No. 03 Candimas, Kotabumi, Lampung Utara, Indonesia
E-mail: robyaja0708@gmail.com[1*], nurmayanti89@gmail.com[2], merriparida27@gmail.com[3], sidik@dcc.ac.id[4]

Article history:
Received: November 16, 2023
Revised: December 12, 2023
Accepted: December 30, 2023

Corresponding authors
*robyaja0708@gmail.com

Keywords:
Restaurant Tax;
Prediction;
Data Mining;
C4.5 Algorithm;
Google Colab.

Abstract
Regional taxes are one of the important sources of regional income to finance the implementation of regional government in the context of serving the community and realizing regional independence. Restaurant tax is one of the regional taxes collected by the Way Kanan Regency Regional Revenue Agency and one of the determinants of the increase in Way Kanan District Original Revenue (PAD). This research raises the problem of not achieving the Restaurant Tax target in 2023. In this study using Data Mining there are various methods in data mining including the C4.5 algorithm. The C4.5 approach can forecast an increase in restaurant taxes, and the computation of the C4.5 algorithm yields the following results. From the results of calculating the 2018-2022 data above using Microsoft Excel, it is known that Class Recommendations total 185 are classified as Yes and No, 0 are classified as Yes but No, Next Class No total 65 is classified as No, and 0 Yes is classified as No, with a total data of 250. Microsoft Excel and Google Colab programs have been used to implement the C4.5 algorithm. Implementation of the C4.5 Algorithm has been carried out using Microsoft Excel and Google Colab applications. The result is that the description of all formulas and predictive results is simpler than the results of manual calculations through Microsoft excel using the C4.5 Algorithm which has an accuracy of 75%, and then proven by Google Colab with results of 100% accuracy.

I. INTRODUCTION

Regional taxes are one of the important sources of regional income to finance the implementation of regional government in the context of serving the community and realizing regional independence. Seeing from this phenomenon, it can be seen that the importance of taxes and fees for a region, especially in supporting the development of the area itself. In relation to taxes, the regional government must explore sources of regional original income (PAD), which are part of the sources of income which can be freely used by each region to organize regional government and development. Based on the potential of each region, an increase in the receipt of Regional Original Revenue will increase the financial capacity of the region. In line with developments, optimizing the utilization of regional original revenue sources is very important. The greater the acceptance and percentage of Regional Original Income to the total regional revenue, it shows that the area is increasingly independent [1].

Restaurant tax is one of the regional taxes collected by the Way Kanan Regency Regional Revenue Agency and one of the determinants of the increase in Way Kanan District Original Revenue (PAD). Restaurant Tax is levied at 10% of the turnover or cost of eating and drinking issued by the Taxpayer/Institution. Some of the causes of not achieving restaurant tax realization are the need to optimize potential sources of restaurant taxes, so actions or solutions are needed to optimize restaurant tax revenue. By socializing the importance of paying

payak to develop the Way Kanan Regency area and increase PAD. The resurgence of economic growth after the pandemic has not had a significant impact on the restaurant tax revenue set by the local government until 2022. The Regional Revenue Agency has never made predictions for restaurant tax revenue to increase tax revenue during that period.

To boost the efficiency of tax collection from the restaurant industry, data mining is being used to forecast the amount of tax revenue from restaurants using the C4.5 algorithm. The use of data mining technology can help overcome obstacles by processing large and complex data to predict the amount of restaurant tax revenue in the future. The application of data mining is also expected to strengthen the performance of the Way Kanan District Revenue Agency in collecting restaurant taxes by optimizing the use of data to identify patterns or trends in restaurant tax revenue data. Thus, the use of data mining in predicting the amount of restaurant tax revenue using the C4.5 algorithm is expected to provide significant benefits to the Way Kanan District Revenue Agency.

II. LITERATURE

2.1. Data Mining

Larose claims that the phrase "data mining" refers to the process of discovering hidden knowledge in databases. Data mining is a semi-automatic process that draws knowledge-related data from vast databases and identifies it using mathematical, statistical, artificial intelligence, and machine learning approaches[2]. The Gatrner Group claims that data mining is the process of identifying significant connections, patterns, and trends through examination of big data sets kept in storage utilizing pattern recognition strategies including statistical and mathematical methodologies[3].

2.2. Data Mining Grouping

Based on the job or task Data mining is divided into several groups, namely [4][5]:

1. an explanation
   Sometimes, analysts and researchers just want to try to describe patterns and trends in data. Possible justifications for patterns and trends are frequently included in descriptions of patterns and trends.
2. A rough estimate
   The only difference between estimation and classification is that during estimation, the target variable is treated more quantitatively than categorically. The target variable's value is used as a predictive value to build the model utilizing a complete record. Additionally, based on the value of the anticipated variable, the estimated value of the target variable is made in the subsequent review.
3. Prognoses
   Classification and estimation are very similar to prediction, with the exception that prediction forecasts the value of the outcome in the future. In the right situations, some of the approaches and methodologies utilized for estimate and classification can also be applied to prediction.
4. Classification
   Classification There is a goal categorical variable in classification. The classification of income, for instance, can be divided into three groups: high income, middle income, and low income.
5. Clustering
   Clustering is the process of establishing classes of related things through data, observations, or paying attention. A cluster is a group of records that are distinct from records in other clusters but comparable to one another. In contrast to classification, clustering doesn't have a goal variable. No classification, estimation, or prediction of the target variable's value are made throughout the clustering process. The clustering algorithm seeks to partition the data into homogenous groups, where records in one group will be most similar to each other and least similar to records in other groups.
6. Association
   In data mining, the association job is to identify attributes that repeatedly emerge.

2.3. Decision Tree Algorithm C4.5

Decision Tree Algorithm C4.5 is a predictive modelling technique that can be used for task classification and prediction. Decision Tree uses a "divide and conquer" technique to divide the problem search space into sets of problems[6], [7][8]. The C4.5 algorithm is an extension of the ID3 (Interative Dichotomiser) algorithm developed by J. Ross Quinlan. The idea is to create a tree with the initial branch as the most important attribute, and then divide it into branches until the rules are satisfied [9]. The decision tree itself is defined as a method of dividing a set of data into smaller sets by applying a set of rules, or decision rules [10]. Algorithm C4.5 basically just repeats the partitioning step so that a situation is obtained where all samples at a node belong to the same class. Each path from root to leaf will represent a decision rule that will be used as a predictor for the next data class.

The information entropy is calculated in this algorithm in order to determine which attribute will occupy a node. The minimum value is then chosen. This algorithm's attribute selection is predicated on the idea that a decision tree's complexity and the amount of information it can convey through its attribute values are directly correlated. In other words, attributes are selected based on the highest information gain to produce subtrees [11].

Some of the criteria possessed by the C4.5 inductive algorithm are[12][13]–[16]:

a. Attribute-value description the data set used for analysis must be represented in the form of an attribute set. Each attribute can have a discrete or continuous value.
b. Predefined class the category that will be given to each sample must be determined first.
c. Discrete class, a case or sample must include or not belong to a certain class and the number of samples must be far more than the number of existing classes.
d. Sufficient amount of data the amount of data required is influenced by the number of attributes and classes as well as the complexity of the classification model used.
e. Logical classification models an inductive approach is used to build a classifier that can be expressed as a decision tree or decision rule.

2.4. Google Colab Software

According to Marlindawati Google Colab or Google Colboratory is an executable document that can be used to store, write, and share programs that have been written via Google Drive [17]. Some of the benefits of Google Colab, namely that as the name itself is Collaborate, can collaborate with other users through various coding online, Free GPU Google Colab makes it easier for users to run computer programs with high specs (GPU Tesla, RAM-12GB, DISK-300GB) Flexible because it can easily run deep learning programs [18]. The advantages of Google Colab are that you don't need to do any configuration because you already use cloud computing technology, free access for high-speed machines (GPU) and it's very easy to connect to Google Drive and GitHub [19]. For the world of Data Science, Google Colab can be used as a Python library to analyzd and visualize data. As for machine learning, Google Colab can be used to do many things such as importing image datasets, training image classifiers, and evaluating models, all of which can be done with just a few codes [20].

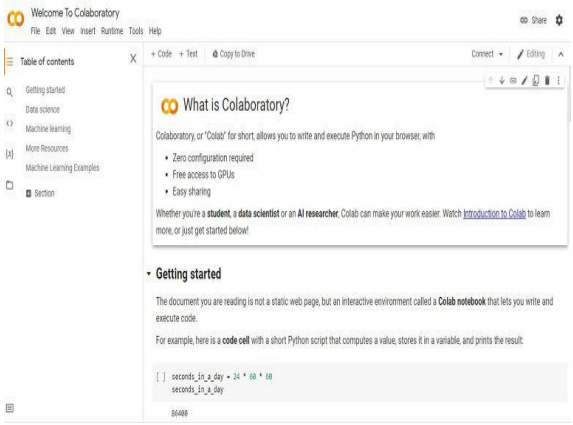

Figure 1. Google Colab Initial Display [20]

2.5. Knowledge Discovery in Database (KDD)

KDD is defined as the discovery or pursuit of knowledge (added value) in a database. Since data mining is a collection of operations, it may be divided into a number of stages, including [21] :

1. Data cleansing (to eliminate erroneous and noisy data).
2. Integration of data (the combining of data from several sources)
3. Data transformation (converting data into a format appropriate for data mining).
4. The use of data mining methods.
5. Analyze the patterns discovered to identify the most intriguing or valuable ones.
6. Knowledge presentation using visualization approaches.
7. Figure 2 provides an illustration of the stages.

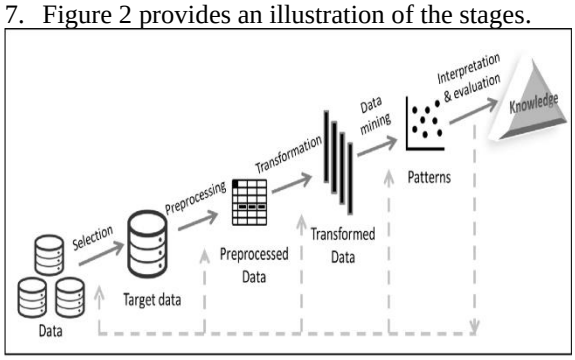

Figure 2. Stages of Knowledge Discovery in Database [4]

III. RESEARCH METHODS

3.1. Stages of the Data Mining Process

The method used in this study is the C4.5 method. The C4.5 research method serves to form a decision tree and is a calcification and prediction method. In making a decision tree using the C4.5 algorithm the steps that must be taken are as follows:

The following is the formula for finding entropy and gain values:

Entropy formula: $Eenropy(S) = \sum_{i=1}^{n} -pi * log2\ pi$

Information:
S: The case
n: Amount of partitions is S
pi: a measure of the Si to S ratio

Gain Formula:

$$Gain\ (S, A) = Entropy\ (s) - \sum_{i=1}^{n} * Entropy\ (Si)$$

Information:
S = The case
A = Attribute
n = Amount
S= in number cases in S
Si = ratio of SI to

Then from the table above to determine the limits achieved or not achieved, it can be calculated by finding the average value of each variable:

$$A1 = \frac{\text{Number of Variable Range Values}}{\text{Many Range Data}}$$

Table 1. Variable Average Value

No	Attribute Evaluation		Assessment Range		Average
	Code	Information	Range	Mark	
1.	A1	turnover	37.6 million - 50 million	4	2.5
			25.1 million - 37.5 million	3	
			12.6 million - 25 million	2	
			560 thousand -12.5 million	1	
2.	A2	reception	3.76 million - 5 million	4	2.5
			2.6 million - 3.75 million	3	
			1.26 million - 2.5 million	2	
			56 thousand -1.25 million	1	
3.	A3	information	Achieved	2	1.5
			No achieved	1	
4.	A4	paid off / no paid off	Paid off	2	1.5
			Not paid off	1	
					8

Table 2. Data Analysis Process

No	Category	Classification
1.	achieved	> =8
2.	No achieved	< 8

Table 3. Calculation results in Microsoft Excel

Knot	Amount	Recommendation		Entrophy	Gain Information
		Yes	No		
Amount Recommendation	12	3	9	**0.811278124**	
A1 Turnover					**0**
560 thousand -12.5 million	9	0	9	**0**	
12.6 million - 25 million	3	3	0	**0**	
25.1 million - 37.5 million	0	0	0	**0**	
37.6 million - 50 million	0	0	0	**0**	
A2 Reception					**0**
56 thousand - 1.25 million	9	0	9	**0**	
1.26 million - 2.5 million	3	3	0	**0**	
2.6 million - 3.75 million	0	0	0	**0**	
3.76 million - 5 million	0	0	0	**0**	
A3 Information					**0.174988789**
Achieved	8	3	5	**0.954434003**	
No achieved	4	0	4	**0**	
A4 Paid/Not Paid					
Paid off	11	3	8	**0.845350937**	**0.036373099**
Not paid off	1	0	1	0	

Based on the calculation above, it shows that the highest gain is in the Description attribute, then the Information is made the Root Node, the Information has 2 criteria, namely Achieved and Not Achieved, for the Description attribute "Not Achieved" Has got its class, namely "No" while for the Description attribute "Achieved" it has not gotten the class, it must be recalculated.

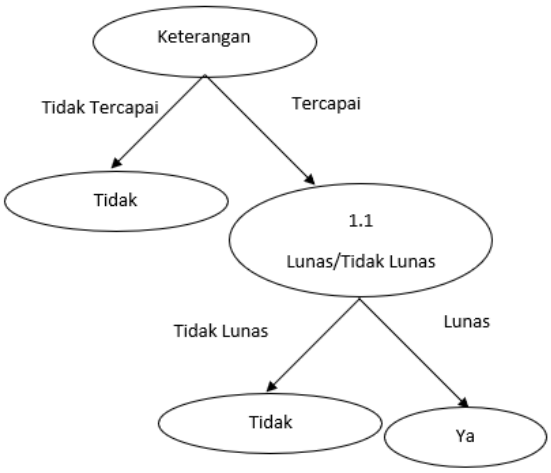

Figure 3. decision-making node 1

Accuracy
The following formula is used to calculate the processed data's accuracy percentage:

$$\text{Accuracy Percentage} = \frac{\text{Number of Correct Prediction Results Data}}{\text{Number of Predictions Made}} \text{ X } 100\%$$

Following are the results of calculations based on data testing with a total of 12 data sales quotas:

$$\text{Accuracy Percentage} = \frac{9}{12} \text{ X } 100\% = 75\%$$

Table 4 Confusion Table

	Class	
Recommendation	YES	NO
YES	4	5
NO	1	2

Based on the calculations above, it can be said that the accuracy of the testing data, which consists of 12 data, has a 75% accuracy rate. The decision tree on the testing data above shows that the information gain on Criterion A4 (Paid/Not Paid) is 0 greater than the other criteria.

IV. RESULTS

After carrying out several stages of processing the data that has been imported into the Python programming language, the prediction results from the Python programming language are known in the image above. From the above code A pdf file with the name Tree_Kelarang.dot, which denotes the file I made, will appear Yes/No Recommendation on Restaurant Tax with 100% accuracy.

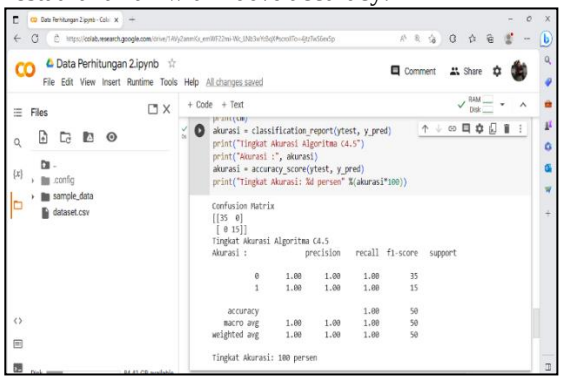

Figure 4. Python Accuracy Results

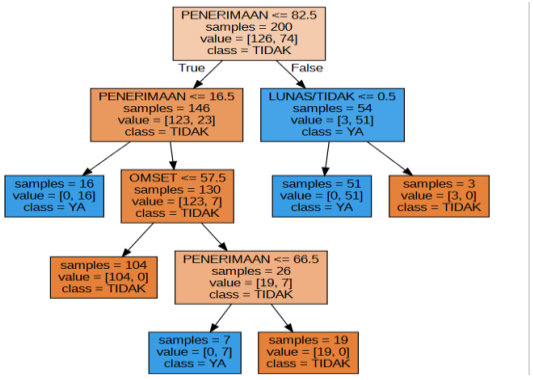

Figure 5. Choice Tree

The C4.5 algorithm's knowledge tree will be used to illustrate the data that we process. It states that the restaurant tax is less than 8 (<=82.5). A selection of 200 out of 250 data from Colab Research Google is automatically integrated. False Menu Paid/Unpaid has 54 data stamps with 51 data stamps Yes and 3 data stamps No, True menu Acceptance has 146 data stamps with 16 stamps Yes, turnover has 130 data stamps with 104 stamps No, Acceptance has 26 data

stamps with 7 stamps data Yes and 19 data semepel No. Additionally, calculations made using Python in the Colab Research Google application are up to 100% accurate.

V. CONCLUSION

The following conclusions can be derived from the debate in earlier chapters the Google Colab and Microsoft Excel apps have been used to develop the C4.5 algorithm. As a result, the complete formula is explained in more detail and with ease. It is possible to use the information system created using the C4.5 algorithm's theory and application right now, but it should also employ Google Colab and Microsoft Excel. Google Colab Research is software that enables all calculations to be performed online with more accurate outcomes. C4.5 implementation Algorithm has been carried out using Microsoft Excel and Googlel Colab applications. The result is that the description of all formulas and predictive results is simpler than the results of manual calculations through Microsoft excel using the C4.5 Algorithm which has an accuracy of 75%, and then proven by Google Colab with results of 100% accuracy.

REFERENCES

[1] A. Biringkanae and R. G. Tammu, "Terhadap Pendapatan Asli Daerah Kabupaten Tana Toraja," *Public Adm. J.*, vol. 4, no. 1, 2021.

[2] A. I. Waspah *et al.*, "Expectation Maximization Algorithm Memprediksi Penjualan Susu Murni Pada Pt . Sewu Primatama Indonesia Lampung," *JUTIM (Jurnal Tek. Inform. Musirawas)*, vol. 7, no. 1, pp. 27–38, 2022.

[3] N. A. Hasibuan *et al.*, "Implementasi Data Mining Untuk Pengaturan Layout," vol. 4, no. 4, pp. 6–11, 2017.

[4] Yuli Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika," *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2019.

[5] dan S. I. Rulin Swastika, Siti Mukodimah, Ferry Susanto, Muhamad Muslihudin, *Implementasi Data Mining (Clastering, Association, Prediction, Estimation, Classification)*, 1st ed. Indramayu: CV. Adanu Abimata, 2023.

[6] J. Han and Kamber, *Data Mining Concepts and Techniques*. San Francisco: Morgan, 2006.

[7] H. Yulianton, "Data Mining untuk Dunia Bisnis Keputusan Informasi," *J. Teknol. Inf. Din.*, vol. XIII, no. 1, pp. 9–15, 2008.

[8] S. Ipnuwati, "Sistem Pendukung Keputusan Perencanaan Promosi Kampus Berbasis Data Mining Dengan Metode Klasifikasi Pada Stmik Pringsewu Lampung," 2013.

[9] O. Kinerja, E. C. U. Study, K. Mobil, and A. Dan, "Implementasi Algoritma K-Means Dan Algoritma Apriori," vol. 1, no. 2, pp. 81–88, 2021.

[10] C. Algorithm *et al.*, "Classification and Clustering of Internet Quota Sales Data Using," vol. 9, no. 2, pp. 268–283, 2023.

[11] I. Junaedi, N. Nuswantari, and V. Yasin, "Perancangan Dan Implementasi Algoritma C4 . 5 Untuk Data Mining," *J. Inf. Syst. Informatics Comput.*, vol. 3, no. 1, pp. 29–44, 2019.

[12] A. Afandi, D. Nurdianah, P. C. Rejo, N. Bayes, and K. A. Dominan, "Naive Bayes Method and C4 . 5 in Classification of Birth Data," vol. 16, no. 4, pp. 435–446, 2022.

[13] A. Zakir, Y. Ndruru, and E. Hadinata, "Penerapan Data Mining Untuk Klasifikasi Data Penjualan Makanan Terlaris Dengan Algoritma C45," *JIFTI - J. Ilm. Teknol. Inf. dan Robot.*, vol. 2, no. 2, pp. 7–12, 2020.

[14] N. N. Nandang Iriadi, "Kajian Penerapan Metode Klasifikasi Data Mining Algoritma C4.5 Untuk Prediksi Kelayakan Kredit Pada Bank Mayapada Jakarta," *J. Tek. Komput. AMIK BSI*, vol. 2, no. 1, pp. 132–137, 2016.

[15] K. Rismayanti, Fera Damayanti, "Penerapan Data Mining Algoritma C4.55 Dalam Menentukan Rekam Jejak Kinerja Dosen STT Harapan Medan," *J. Penelit. Tek. Inform.*, vol. 3, no. 1, pp. 99–104, 2018.

[16] A. Dhond, "Data Mining Techniques for Optimizing Inventories for Electronic Commerce," in *Knowledge discovery and data mining*, 2020, pp. 480–486.

[17] A. Sitio, A. Sindar, M. Marbun, D. Tiara, and A. Aswin, "Pengenalan Data Scientist Pada Peserta PKBM AL HABIB Melalui Belajar Dasar Coding Python," *J. Pengabdi. Pada Masy.*, vol. 7, no. 1, pp. 194–200, 2022.

[18] S. Mintoro and A. Afandi, "Implementasi Algoritma K-Means Dan Algoritma Apriori Optimasi Kinerja Ecu (Study Kasus Mobil Avanza Dan Xenia)," *J. Inf. dan Komput.*, vol. 9, no. 2, pp. 81–88, 2021.

[19] F. Rahmadani, A. M. H. Pardede, and Nurhayati, "Jaringan Syaraf Tiruan Prediksi Jumlah Pengiriman Barang Menggunakan Metode Backpropagation," *J. Tek. Inform. Kaputama*, vol. 5, no. 1, pp. 100–106, 2021.

[20] R. O. Felani, W. Cholil, and H. Syaputra, "Analisa Prilaku Pengguna E-Learning Menggunakan," vol. 7, no. 1, pp. 61–73, 2022.

[21] F. Sari and D. Saro, "Implementasi Algoritma C4.5 Dalam Menentukan Lokasi Prioritas Penyuluhan Program Keluarga berencana di kecamatan dumai timur," *J. Penelit. Pos dan Inform.*, vol. 8, no. 1, p. 63, 2018.