Journal of Systems Engineering and Information Technology
Vol. 1 No. 1 (2022) Page 1 – 9 | e-ISSN: 2829-310X
Journal Homepage: www.jurnal.iaii.or.id/index.php/JOSEIT

Implementation of Decision Tree for Making Decision of Claim Product from Steel Production

Anissa Lestari[1*], Saruni Dwiasnati[2]
[1,2]Faculty of Computer Science, Universitas Mercu Buana, Indonesia

*Email corresponding author: anissalestari22@gmail.com

Abstract

Product Claims is requests from consumers for products purchased from suppliers in accordance with agreements agreed by both parties. Products that have been claimed from consumers produce historical data sets that can be used as evaluations for producers to produce higher quality products. This study aims to process production data and shipment data then classify the types of products claimed based on the results of claim report from consumers. Data mining can be extracted information from a very large amount of data with specific methods to obtain information or new science. The method used in this study is the C4.5 algorithm method using the production code attribute as a claim or non-claim label attribute. This study produced a decision tree of 4 variables, there are thick of product, width of product, weight of product, destination of product, and type of product claim as label. This decision tree concept collects data which then calculates the value of entropy and gain to determine the rule. The conclusion from this study is the C4.5 algorithm helps classify the product claims and form a decision tree that can provide information about production results and can ensure with consumers related to product limits that may be claimed according to the agreed agreement. Evaluation of the results obtained that the algorithm C4.5 is 99.9% accuracy.

Keywords: data mining, algorithm C4.5, decision tree, claim product

1. Introduction

Steel is the main element that determines the quality of construction or building products. PT. KS is one of the steel industries in Banten province. Currently, the steel industry is struggling from imported steel strikes that are still difficult to control. That is, PT. KS not only competes with the local industry but also the steel industry abroad. With so many steel mill competitors, more and more competitors must be faced by PT. ks. To strengthen this, PT. KS must produce good and high quality products as expected by consumers. In the process of buying and selling, PT. KS as a producer or supplier for consumers has several agreements regarding product claims. It is possible that the results of inspections from Quality Control assesses that the product is good so that it passes the inspection, while the goods sent to consumers have defects so that consumers make claims on the product. Claim Product is a request from the consumer for a product purchased from a supplier in accordance with an agreement agreed by both parties. Products that have been claimed from consumers produce historical data that can be used as useful knowledge for producer as a supplier.

To predict product claims by consumers to producers, a classification method is needed. Where this method can classify data classes. Data Mining is one of the business solutions in the field of Information Technology that can help analyze the decision making of a lot of data that must be processed.

Based on the problems in PT. KS, it is necessary to make a decision to determine the product claims and not claims of production using the C4.5 Algorithm, because the C4.5 Algorithm can be used to build decision trees or decision making. The decision tree is an answer to a system that humans have developed to help find and make decisions for these problems and take into account the various factors that are within the scope of the problem. With a decision tree, humans can easily identify and see the relationships between the factors that influence a problem and can find the best solution by taking into account these factors. [1]

There are various methods for doing data mining such as Naive Bayes, K-NN (K-Nearest Neigbour) and C4.5. Based on research conducted by Riza Alamsyah, Ade Davy Wiranata, and Rafie [2] concluded that a model built using the K-Nearest Neighbor method can improve accuracy for detecting defects in ceramic tiles with an accuracy value of 98.947% for K=3. Research conducted by Choirul Anam and Harry Budi [3] testing the classification model shows that the accuracy rate of the C4.5 algorithm mode is 96.4% with 0s processing time and the accuracy of the Naive Bayes algorithm model is 95.11% with 0s processing time. Comparison of accuracy is in line with

Accepted by editor: 30-07-2020 | Final revision: 4-10-2020 | Online publication : 25-03-2022

the results of similar studies in other studies where the accuracy of the C4.5 algorithm model is higher than the Naive Bayes algorithm model.

Previous research conducted by Dian Ardiansyah and Walim [4] has a conclusion that the accuracy of the C4.5 algorithm in determining 33 students data classification of prospective quiz participants is 81.81%.

To support research in determining decision making, information on thick of product, width of product, weight of product, and destination of product. Then the data is classified and produce a product decision label is claim or no claim using the C4.5 Algorithm and testing using WEKA application.

2. Method

2.1.Theoritical

A. Data Mining

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically [20].

Many uses can be used in processing data mining, which can help get useful information and increase knowledge from various data that can be solved by various algorithms in data mining. The definition of data mining itself is Data mining is a step in carrying out Knowledge Discovery in Databases (KDD). Knowledge discovery as a process consists of data cleaning (data cleaning), data integration (data integration), data selection (data selection), data transformation (data transformation), data mining, pattern evaluation (pattern evaluation) and presentation of knowledge (knowledge presentation) [15].

B. Classification

Classification method is a method that is included in the most common supervised process that can be used in data mining. Business issues such as Churn Analysis, and Risk Management usually involve a Classification method to facilitate the resolution of the problem. Classification is the act of giving groups to every situation. Each state contains a group of attributes / indicators, one of which is a class attribute. This method requires a model that can explain the class attribute as a function of the input attribute. The advantage of the classification method is that the dataset used in absolute classification must display the class / target attribute and the knowledge generated by the classification method in the form of clusters (can be Decision Tree, Ruleset, Weight on BackPropagation, etc.) [11].

C. Algorithm C4.5

C4.5 algorithm is an algorithm used to produce a decision tree developed by Ross qiunlan. The basic idea of this algorithm is to make a decision tree based on the selection of attributes that have the highest priority or can be called the highest gain value based on the value of the attribute entropy as the axis of the attribute classification. At the stage C4.5 algorithm has 2 working principles, namely: Making a decision tree, and making rules (rule model). Rules that are formed from the decision tree will form a condition in the form of if then. There are four steps in the decision tree making process C4.5 algorithm, namely:

1. Choose the attribute as the root, based on the highest gain value of the existing attributes.
2. Make a branch for each value, which means make a branch in accordance with the highest number of variable values.
3. Dividing each case in a branch, based on the calculation of the highest gain value and the calculation carried out after the calculation of the highest initial gain value and then the process of calculating the highest gain again without including the initial gain variable value.
4. Repeating the process in each branch so that all cases in the branch have the same class, repeating all the processes of calculating the highest gain for each branch of the case until the calculation process can no longer be done.

C4.5 algorithm recursively visits each decision node, choosing the optimal division, until it cannot be subdivided. Of the three researchers who have done it, the classification with C4.5 Algorithm is used by researchers as a solution for making decisions that are expected to help in making decisions more easily and quickly [12].

In the application and use of the C4.5 algorithm, it can be used for make predictions and classifications of the wheather dataset having 14 instances and 4 attributes, to make a decision whether or not to play a particular game on the basis of the conditions [13]. Besides this algorithm is used to predict motorcycle sales at a PT. The variable that has the first priority on the predicate of selling motorcycle sales is the selling variable via, meaning that the distribution of motorcycles from various places greatly affects the motorcycle sales [6].

D. Decision Tree

Decision tree is a structure that can be used to convert data into decision trees that will produce large decision rules into smaller record sets by applying a series of decision rules. The decision tree produced by the C4.5 algorithm can be used for classification. Decision tree is one of the most popular data mining techniques for knowledge discovery. Systematically analyzing and extracting rules for the purpose of classification / prediction [12].

Data in decision trees is usually stated in tabular form with attributes and records. Attributes state a parameter called criteria for tree formation. The main benefit of using a decision tree is its ability to break down complex decision-making processes to be simpler so that decision makers will better interpret the solution of the problem (Elmande, 2012). For example to determine playing tennis, the criteria to consider are weather, wind, and temperature. One attribute is an attribute that states the solution data per data item called the result attribute [21]. In the decision tree there are 3 types of nodes, namely:

1. Root Node, is the top node, at this node there is no input and can have no output or have more than one output.
2. Internal Node, is a branching node, at this node there is only one input and has a minimum output of two.
3. Leaf node or terminal node, is the final node, at this node there is only one input and has no output.

E. WEKA

WEKA (Waikato Environment for Knowledge Analysis) is a software that has many machine learning algorithms for data mining purposes. Weka also has many tools for data processing, ranging from pre-processing, classification, regression, clustering, association rules, and visualization. Weka is an open source Java based software and we can use it directly or through our Java program. Weka can also be implemented into a python program.

Figure 1. WEKA Logo

2.2 Data Collector Methodology

A. Processing and Transformation of data

Not all attributes in the production claim data database can be used in research, attributes such as: coil number, Customer ID, Production Order, Customer Name, Sample Test Number, spec raw material, raw material weight, and production date are not needed in this study. Processing only requires five attributes, such as: thick of product, width of product, weight of product, destination of product, and claim categories as the objective attributes of this study, so that the data transformation process will be processed, after the data transformation is done later will be processed using the C4.5 algorithm. Samples of data that have been transformed can be seen in table 1 below:

Table 1. Data Transformasi Product Claim of Steel Production

No	Thick	Width	Weight	Destination	Claim
1	2.1	1219	20640	domestik	Yes
2	12	650	5120	domestik	No
3	2.3	670	5220	blank	No
4	8.36	750	12640	export	No
5	2	775	14440	domestik	Yes
6	11.8	800	12640	blank	No
7	8.36	750	12650	export	No
8	1.8	1200	11000	domestik	Yes
9	2.3	700	12650	domestik	No
10	2.25	1025	12680	domestik	No
...	...	...	...	...	...
77.393	4.4	1340	25510	export	No

B. Data Processing

Data processing starts with finding total entropy of all attributes and then determine the gain the highest. To get the gain value in decision tree formation, need to calculate first information value in units of bits of a collection of objects. The form of calculation for entropy is as the following:

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

Where:
S = The Set of cases
N = number of pastitions
P_i = proportion of S_i to S

The value of Entropy (X) indicates that X is random attribute. The entropy value reaches a value minimum 0, when all other pj = 0 or are at same class. At the construction of the C4.5 tree, at each tree node is filled by the attribute with the gain ratio value highest, with the following formula:

$$\text{Gain(S)=Entropy(S)} - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy\ (S_i)$$

Where:
S = the set of cases
n = number of partitions
$|S_i|$ = number of cases in the partition i
$|\ S\ |$ = number of cases in S

The process of finding total entropy and gain is done by grouping the data correctly, then calculating the data and using the entropy and gain search formula for each data attribute.

3. Result and Discussion

Testing of the analysis is very important to determine and ascertain whether the results of the analysis are in accordance with the expected decision. To test the truth of the results of data processing done manually, it can use one of the WEKA 3.9.0 application software.

The steps in processing data using WEKA are as follows:
All variable data and decisions to be processed by WEKA are stored in Microsoft Excel in the .xls format first. Then change the data format and Save as type look for the format .csv. Please note that CSV (MS-DOS) for Window, CSV (Macintosh) for Apple users and CSV (Comma Delimited) the difference only lies in the comma [6].

1. Open the WEKA 3.9.0 software, double click on the shortcut or search through My Computer and the WEKA display will appear.
2. There are 4 Application selection buttons namely Explorer, Experiment Knowledge Flow and Simple CLI. Select Explorer, then select Open file, find where the sales.csv file is located, select it and click Open.
3. The next step is to select the variables that affect the data to be processed.
4. Click the Classify menu, in Classifier click Choose, for the C4.5 algorithm select trees and click J48.
5. Select Use training set then click the Start button.
    a. Classifier output will appear in the right area which is the result of processing from WEKA. The output classifier contains the inputted running data which are the attributes that will form the decision tree. Information
    b. which is displayed in the form of the number of cases and their decisions and the number of branches of the decision tree.
6. Finally, to see the decision tree formed by WEKA processing is to right-click on the Result list, select Visualize tree.

The process of training data using applications in WEKA is Data Claim production using C4.5 algorithm with test options using menu the "Use Training Set".

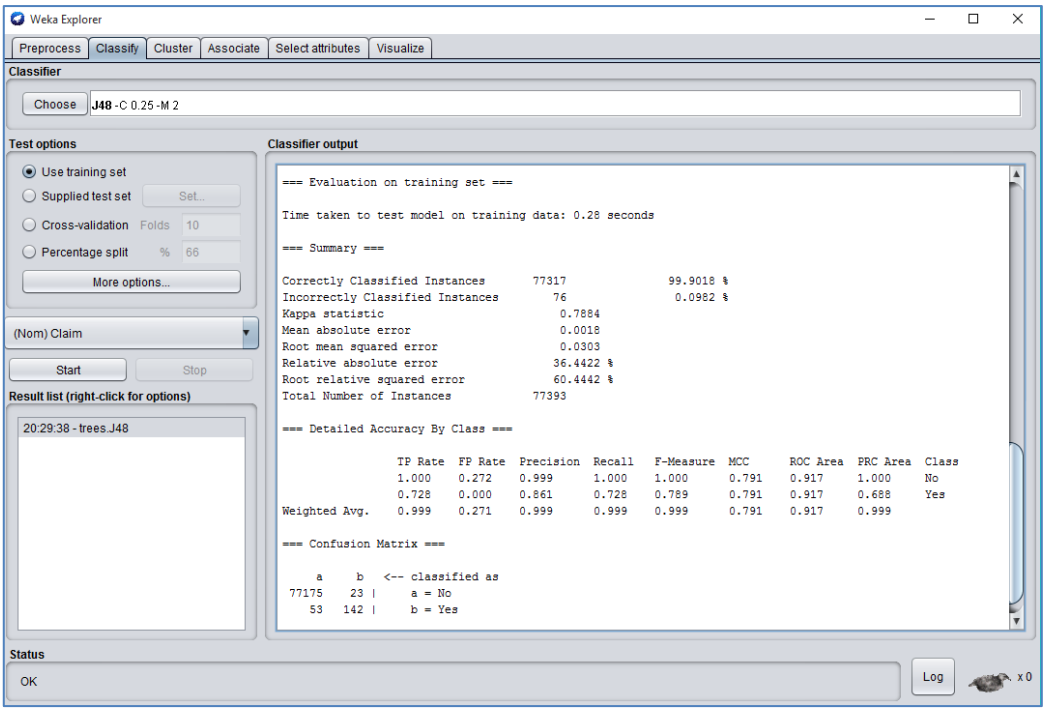

Figure 2. Process data J48 in WEKA

Here is the visualization of the decision tree produced by WEKA processing:

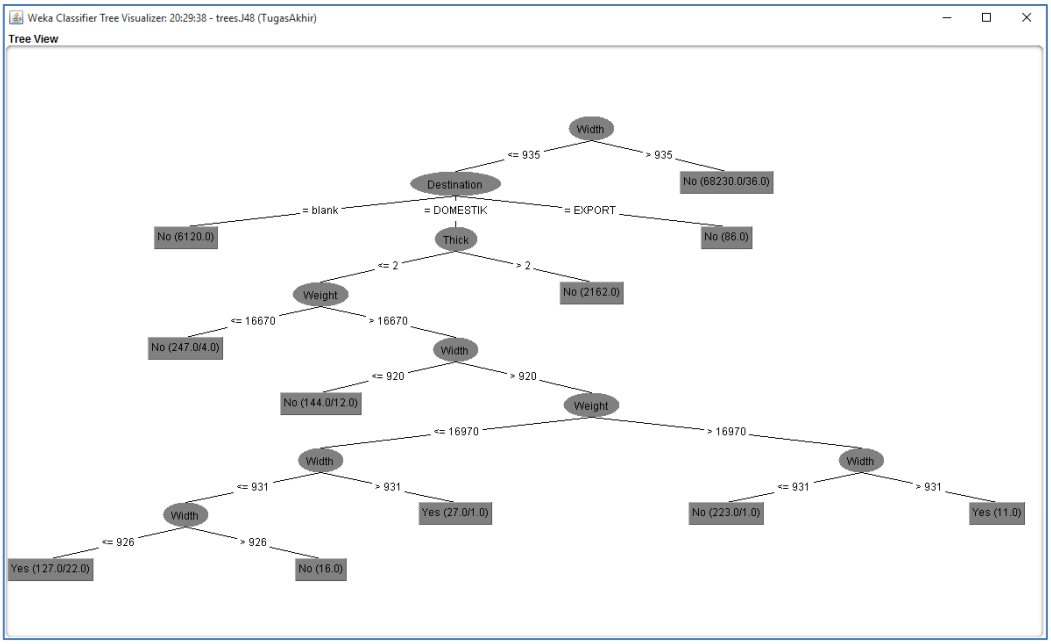

Figure 3. Decision Tree in WEKA

And here is the zoom in of the decision tree:

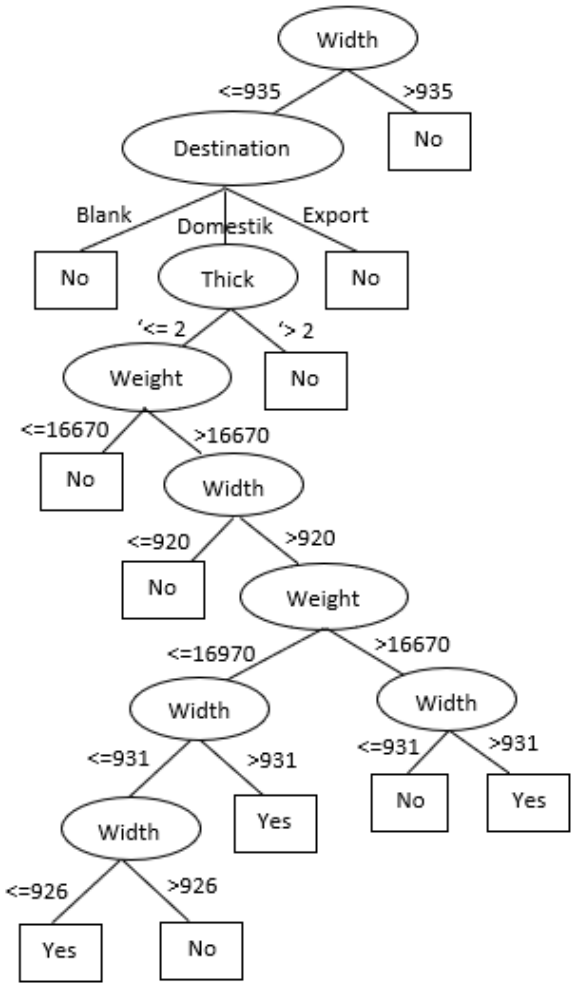

Figure 4. Detail of Decision Tree

In the 5 attributes a decision tree is formed as shown below:
Number of Leaves :11
Size of the tree :20

From the decision tree formed in Figure 3 & 4 above, obtained rules - model rules in determining acceptance recommendations sales partner. There are 11 rules that are formed, can be seen as follows:

1) *If* Width > 935 *Then* no claim.
2) *If* Width < = 935 *And* Destination blank *Then* no claim.
3) *If* Width < = 935 *And* Destination export *Then* no claim.
4) *If* Width < = 935 *And* Destination domestik *And* Thick > 2 *Then* no claim.
5) *If* Width < = 935 *And* Destination domestik *And* Thick < = 2 *And* Weight < = 16670 *Then* no claim.
6) *If* Width < = 935 *And* Destination domestik *And* Thick < = 2 *And* Weight > 16670 *And* Width < = 920 *Then* no claim.
7) *If* Width < = 935 *And* Destination domestik *And* Thick < = 2 *And* Weight > 16670 *And* Width > 920 *And* Weight > 16970 *And* Width < = 931 *Then* no claim.
8) *If* Width < = 935 *And* Destination domestik *And* Thick < = 2 *And* Weight > 16670 *And* Width > 920 *And* Weight > 16970 *And* Width > 931 *Then* yes claim.
9) *If* Width < = 935 *And* Destination domestik *And* Thick < = 2 *And* Weight < = 16670 *And* Width > 931 *Then* yes claim.
10) *If* Width < = 935 *And* Destination domestik *And* Thick < = 2 *And* Weight < = 16670 *And* Width < = 931 *And* Width > 926 *Then* no claim.
11) *If* Width < = 935 *And* Destination domestik *And* Thick < = 2 *And* Weight < = 16670 *And* Width < = 931 *And* Width < = 926 *Then* yes claim.

The results of production claim data testing are done shown in the Table below:

Table 2. Test result of *Use Training Set* in WEKA application

Use Training Set		
Summary		
Correctly Classified Instance	77317	99.9018 %
Incorrectly Classified Instance	76	0.0982 %
Kappa statistic	0.7884	
Mean absolute error	0.0018	
Root mean squared error	0.0303	
Relative absolute error	36.4422 %	
Root relative squared error	60.4442 %	
Total Number of Instance	77393	

The Weka application also has a Confusion Matrix result like the table below:

Table 3. The Result of Confusion Matrix

	No	Yes
Prediction of No claim	77175	23
Prediction of Yes claim	53	142

Table 4. The Result of Confusion Matrix

		True Value	
		True	False
Prediction of Value	True	TP	FP
	False	FN	TN

Where:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Furthermore, the decision tree algorithm is calculated by calculating Accuracy, Precision and Recall.

a. Accuracy is the ratio of true predictions (positive and negative) to the overall data. This accuracy answers the question "What percentage of consumers are correctly predicted not to claim and claim from the total production of steel factory PT.KS?"

   Accuracy $\left(\left(\frac{TP+\mathrm{TN}}{TP+TN+FP+FN}\right)\times 100\%\right)$

   The accuracy calculation is as follows:

   Accuracy $\left(\left(\frac{77175+142}{77175+142+23+53}\right)\times 100\%\right) = 99{,}90\ \%$

   So, the results of the Accuracy in calculating the number of products that are not claimed and claimed by consumers divided by the total amount of PT.KS production are 99.90%.

b. Precision is the level of accuracy between the requested data and the predictive results given by the model or the ratio of positive true predictions compared to the overall positive predicted results. Precission answers the question "What percentage of non-claim consumers of all consumers do not actually claim?"

   $$\text{Precision} = \left(\left(\frac{TP}{TP+FP}\right)\times 100\%\right)$$

   The calculation of precision is as follows:

Precision = $\left(\left(\frac{77175}{77175+23}\right) \times 100\%\right) = 99{,}97\ \%$

So, the Precision results in calculating the yield of products that are not claimed by consumers divided by the actual number of non-claimed products is 99.97%

c. Recall is the ratio of true positive predictions to the overall true positive data. Recall answers the question "What percentage of consumers who do not claim of the total consumers who are predicted not to claim?"

Recall = $\left(\left(\frac{TP}{TP+FN}\right) \times 100\%\right)$

The calculation of recall is as follows:

Recall = $\left(\left(\frac{77175}{77175+53}\right) \times 100\%\right) = 99{,}93\ \%$

So, the recall result calculates the actual product yield that is not claimed by the consumer divided by the predicted number of the non-claimed product is 99.93%.

From the above calculation, it can be concluded that the results of the calculation of accuracy, precision, and recall are the same as the results of the calculations shown in table 3. Based on testing and analysis of the results of tests performed, with an accuracy rate of 99.90%, precision 99.97%, and recall 99.93% showed a value of almost one hundred percent accurate, which is still in the good category and concluded that researchers succeeded in implementing the C4.5 classification algorithm so that it can help in making a decision, whether steel products that have been sent later will be claimed or not by consumers.

4. Conclusion

C4.5 algorithm test results that have been done, there are some conclusions as well as suggestions related to the previous discussion. There are 5 attributes that influence claim prediction modeling, namely thick, width, weight, destination, and claim status. The accuracy of prediction modeling is not a claim in this study, namely 99.90%.C4.5 algorithm with the decision tree method can provide predictive rule information to describe the process related to making decision of claim product from Steel Production.

Algorithm C4.5 with the decision tree method can provide rule information by modeling the claim prediction results that consumers who have a width less than 926 mm have a higher claim rate than the steel production of PT. KS.

References

[1] Kurniah Reni, 2020. Analisa dan Penerimaan Metode Klasifikasi dalam Data Mining Untuk Penerimaan Siswa Jalur Non-Tulis. *Jurnal Ilmiah Informatika (JIF).* Vol.8 No.1 ISSN (Online) 2615-1049 / ISSN (Print) 2337-8379.

[2] Alamsyah. Riza, D.W. Ade and Rafie, 2019. Deteksi Cacat Ubin Keramik dengan Metode K-Nearest Neighbor. *Techno.com* Vol.18 No.3 : 245-250.

[3] Anam Choirul, Budi Santoso. Harry. 2018. Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerimaan Beasiswa. Jurnal ENERGY. Vol.8 No.1 ISSN 2088-4591.

[4] Ardiansyah. Dian, 2018. Algoritma C4.5 Untuk Klasifikasi Calon Peserta omba Cerdas Cermat siswa SMP dengan Menggunakan Aplikasi Rapid Miner. *Jurnal Infokar*. Vol.1 No.2 ISSN (Print): 2615-3645 / ISSN (online) 2581-2920

[5] Siringoringo. Rimbun, 2016. Kajian Kinerja Metode Fuzzy k-Nearest Neighbour pada Prediks Cacat Software". *JURNAL methodika.* Vol.2 No.2 ISSN: 2442-7861

[6] Azwanti. Nurul, 2018. "Analisa Algoritma C4.5 untuk memprediksi Penjualan Motor pada PT. Capella Dinamik Nusantara Cabang Muka Kuning". *Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer.* Vol.13. No.1. ISSN (Print) 1858-4853 / ISSN (Online) 2597-4963.

[7] Elisa. Erkin, 2018. "Prediksi Profit ada Perusahaan dengan Klasifikasi Algoritma C4.5". *Kumpulan Jurnal Imu Komputer (KLIK)* Vol.05. No.02. ISSN 2406-7857

[8] Oktana. Ivan, Hansun. Seng, 2015. "Penerapan Algoritma C4.5 pada Analisis Kerusakan Barang Jadi". *ULTIMA Computing*, Vol.VII, NO.1. ISSN 2335-3286

[9] Saleh Alfa, Maryam Meilinda, 2019. "Pemanfaatan Teknik Data Mining dalam Menentukan Standar Mutu Jagung". Cogito Smart Journal Vol.5 No.2 ISSN (Print) 2541-2221 / ISSN (Online) 2477-8079

[10] Aditya F. Yudha, Defit Sajon, Yuhandri. 2017. Penerapan Algoritma C4.5 untuk Klasifikasi Data Rekam Medis berdasarkan International Classification Disease (ICD-10). *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, Vol. 1 No.2 (2017) 82-89

[11] Dwiasnati Saruni, Devianto Yudo. 2019. "Classification of Flood Disaster Predictions using the C5.0 and SVM Algorithms based on Flood Disaster Prone Areas" Vol. 67 Issue 7 – July, International Journal of Computer Technique (IJCT), ISSN:2394 – 2231, www.ijctjournal.org
[12] Fauzul A. Muhammad, Fitrianah Devi, 2018 "Penerapan Algoritma Klasifikasi C4.5 dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus: PT Atria Artha Persada", *IncomTech, Jurnal Telekomunikasi dan Komputer*, Vol. 8 No.2, ISSN (Print) 2085-4811/ ISSN (Online) 2579-6089
[13] Chauhan Harvinder, Chauhan Anu, 2013. Implementation of Decision Tree C4.5 Algorithm. *International Journal of Science and Research Publications*, Vol. 3 (10), ISSN 2250-3153
[14] G. L. Pritalia, 2018. Penerapan Algoritma C4.5 untuk Penentuan Ketersediaan Barang E-Commerce. *Indonesia Journal of Information Systems (IJIS)*, Vol.1. No.1 ISSN (Online) 2623-2308 / ISSN (Print) 2623-0119.
[15] Firdaus Diky, 2017. Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer. *Journal Format* Vol.6 No.2. NISN:2089-5615
[16] Selva. Jumeilah, Fithri, 2017. Penerapan Support Vector Machine (SVM) untuk pengkategorian Penelitian. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* ISSN:2580-0760
[17] Hamza Sahriar, 2019. Implementasi Data Mining Untuk Memprediksi Jumlah Peredaran Kendaraan Roda Empat di Kota Ternate Menggunakan Metode C4.5. *Jurnal DINTEK*. VOL 12. Nomor 1 Maret 2019. ISSN (Online) 2508-889 / ISSN (Print) 1979-3855.
[18] Angelia S.Yolanda, 2019. Implementasi Algoritma C4.5 Untuk Pengolahan Data Hutan Lindung (Studi Kasus Dinas Kehutanan Sumatera Utara). KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer). Vol.3 Nomor 1. ISSN (Online) 2597-4645 / ISSN (Print) 2597-4610
[19] Naufal. Muhammad Alfareza, Ayu. Tifa Praditya, Mawarni Kintan, 2019. Implementasi Fuzzy Klasifikasi sebagai pengambilan Keputusan Produk Cacat (Studi Kasus pada UMKM X). *Seminar Nasional* IENACO Yogyakarta: Universitas Islam Indonesia. ISSN: 2337-4349
[20] Han, J. Kamber, M & Jian, Pei. 2011. Data Mining: Concepts and techniques, Third Edition. America: Morgan Kauffman, San Francisco.
[21] Yulia, Y., & Azwanti, N. 2018. Penerapan Algoritma C4. 5 Untuk Memprediksi Besarnya Penggunaan Listrik Rumah Tangga di Kota Batam. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *2*(2), 584-590.