Jurnal CoreIT, Vol.9, No.1, June 2023
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Violation Types Determination of the Whistleblowing System Using the C4.5 Algorithm

Dwi Vernanda[*], Rian Piarna[1], Helfira Lustiana[2], Tri Herdiawan Apandi[3]
[*,1,2,3] Manejemen Informatika, Politeknik Negeri Subang, Indonesia
nanda@polsub.ac.id, piarna@polsub.ac.id, helfira@student.polsub.ac.id, tri@polsub.ac.id

Abstract. Whistleblowing is a complaint system and follow-up management of each violation report. The problem that arises is when determining the follow-up, namely determining the severity or severity of the violation and the sanctions given are only based on the superior's assessment without adhering to standard guidelines or rules. This results in the sanctions given not in accordance with the violations committed. The purpose of this study is to classify the types of violations so as to facilitate the determination of sanctions on the whistleblowing system using the C4.5 Algorithm. The partition was performed three times with the highest additional value of 0.8516 and a decision tree was obtained. Based on the decision tree, the final node that has been generated is then extracted into 27 rules. The classification results from the C4.5 Algorithm can be used to classify the types of violations with an accuracy rate of more than 80%. The first validation with 15 tests obtained an accuracy rate of 86.66%. The second validation is the combination of data on 125 cases of combination data and obtained an accuracy rate of 84.8%. The decision tree generated from three partitions consists of 27 rules that can be used as a pattern to classify the types of violations.

Keywords: C4.5 Algorithm, Classification, Violation, WBS.

Received May 2023 / **Revised** May 2023 / **Accepted** June 2023

INTRODUCTION

The Indonesian government has promoted an honest and clean government, this applies not only to government institutions but also to State-Owned Enterprises (BUMN) and other private companies. Information technology governance, especially in companies under the auspices of State-Owned Enterprises (BUMN), is something that needs to be done so that companies have a systematic mechanism to meet the rules and needs for their business. Based on the principles of Good Corporate Governance (GCG) a company can achieve the vision, mission [1]. In addition to better corporate governance, WBS plays an important role as a warning or sanction for someone who commits fraud [2].

PT X is one of the state-owned companies that has implemented steps to realize the principles of GCG in accordance with the SOE Circular Letter Number: SE-05/MBU/2013 Regarding the Road Map Towards a CLEAN BUMN, namely by creating a reporting system or complaint against fraud, illegal practices, and immoral or unlawful activities carried out by members of the organization that occur within the organization where they work. However, in its implementation the system is still in the form of a manual using written media in the form of a complaint box.

In addition, some parties also feel uncomfortable when they want to complain about a violation because the written media provided is not sufficient to maintain the confidentiality of the reporter as well as the contents of the report. The decisions taken when determining the severity or severity of the violation and the appropriate sanctions are still based on the superior's assessment without adhering to the applicable guidelines or rules and a definite system in determining the decision, so that sometimes the types of violations and sanctions given are not in accordance with the violations committed. This can lead to injustice among employees [3].

Exposure to some of the statements above, we need a data mining approach with the application of the C4.5 Algorithm that can be done to determine the classification of the type of violation according to the description of the violation committed [4]. The use of a simple and easy to interpret structure allows this algorithm to help solve problems [5]. The chance of compatibility between sanctions and the type of violation committed will be greater because the prediction results from the classification using the C4.5

DOI: 10.24014/coreit.v9i1.22897

Algorithm [6]. So it is hoped that the C4.5 Algorithm will be able to become the right decision support tool in determining the classification of types of violations [7].

Based on the description of some of the problems above, an analysis will be developed for the classification of types of violations that will facilitate the determination of sanctions for violations that occur. This study uses data mining techniques by applying the C4.5 Algorithm to find patterns of types of violations and existing sanctions [8], then used as the basis for classifying the next type of violation [9].

METHODS

The data used in this study are data that obtained from the PKB PT X document. The document contains all regulations, violation sanctions, and types of violations that apply at PT X. As well as data on violations and sanctions that occurred at PT X from 2010 to 2020. The data that has been collected is then processed following the steps in the calculation of the C4.5 Algorithm [10]. The settlement method is made as a travel path to facilitate the author in conducting research, the completion method can be seen in Figure 1.

Initial data processing is data from PT X's PKB document which consists of several attributes or assessment factors which are then processed using RapidMiner 9.7.001 to obtain a decision tree pattern. At this stage, attribute determination is carried out, where these attributes will produce a model to form a decision tree for determining the type of violation [11]. Attributes are determined based on statements that are adjusted to the provisions of PT X. The following are attributes with several statements [12].

The data used in this study is data obtained from the PKB document of PT X. Documents contain all regulations, violation sanctions, and types of violations that apply in PT X. As well as data on violations and sanctions that occurred at PT X from 2010 to 2020. The data that has been collected is further processed following the steps in the calculation of the C4.5 Algorithm. The settlement method was created as a journey flow to make it easier for the author to conduct research, the settlement method can be seen in Figure 1.

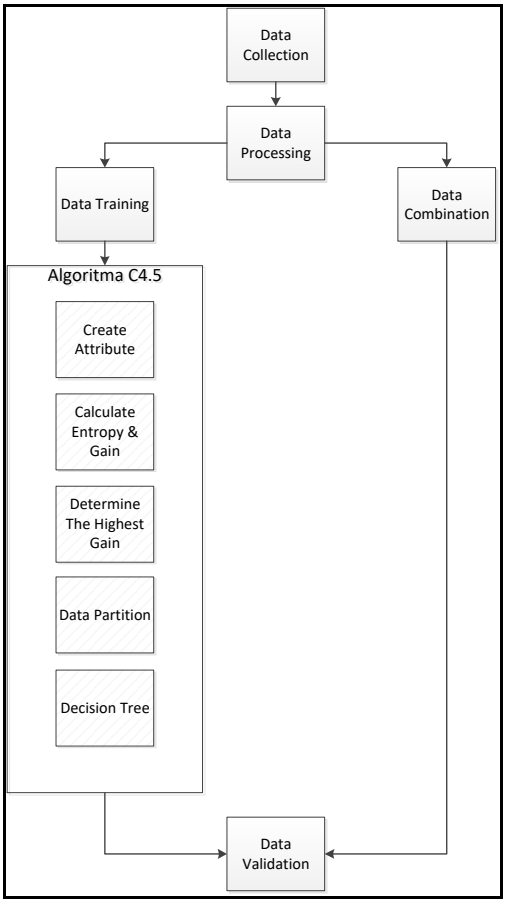

Figure1. Settlement Methods

The initial data processing is data from PT X's PKB document which consists of several attributes or assessment factors which are then processed using RapidMiner 9.7.001 to obtain a decision tree pattern. At this stage, attribute determination is carried out, where these attributes will produce a model to form a decision tree for determining the type of violation [11]. Attributes are determined based on statements that are adjusted to the provisions of PT X. Here are attributes with multiple statements.

Table 1. Statements and attributes

Statement 1	
A	None
B	Perform actions that hinder work activities such as absenteeism without a valid reason or acceptable to the company from and up to 3 – 4 consecutive days
C	Late for work, leaving work early without permission or during working hours leaving the workplace without the permission of the superior
D	Manipulating absenteeism
E	Neglecting work for its own sake without taking responsibility
Statement 2	
A	None
B	Misusing company-owned goods/facilities for personal interests or other activities that harm the company
C	Providing false information, falsifying data and/or signatures and other staffing documents that can harm the company
D	Abusing (possessing, selling, buying, pawning, renting, lending or giving away) confidential/valuable goods, documents or papers belonging to the government or belonging to a legitimate company
E	Carelessly or intentionally damage or leave in a state of danger the company's property that causes losses to the company
Statement 3	
A	None
B	Give preferential treatment to anyone who may result in personal gains or losses to the Company
C	Committing fraud, theft, embezzlement of goods/money belonging to others or belonging to companies
D	Committing corruption, collusion and nepotism so as to cause losses to the company
E	Committing crimes such as harassment, assault, intimidation, persecution, abusive humiliation, violent acts, threatening the Head of the company or employees or their families

Information:
A: Mild
B: Minor
C: Moderate
D: Quite serious
E: Serious

Based on data from PT X, there are 15 types of violations as follows:

Table 2. Breach case data

No	Statement 1	Statement 2	Statement 3	Types of Violations
1	A	A	A	Mild
2	C	D	A	Mild
3	B	A	A	Serious
4	A	A	B	Mild
5	C	B	B	Serious
6	B	C	C	Serious
7	C	C	B	Mild
8	B	D	B	Mild
9	C	C	C	Serious

No	Statement 1	Statement 2	Statement 3	Types of Violations
10	D	C	C	Serious
11	D	C	D	Serious
12	D	C	D	Serious
13	E	C	D	Serious
14	E	D	D	Serious
15	E	E	E	Serious
16	A	B	D	Mild
17	C	D	E	Serious
18	B	E	A	Mild
19	B	E	D	Serious
20	A	D	C	Mild

The calculation of entropy and gain is a step to determine the result of the weight of each attribute [13] and the criteria for obtaining the highest value which is then used as the root of the decision tree [14].

$$Entropy\ (S) = \sum_{l-1}^{n} - pi * log_2\ pi \qquad (1)$$

Information:
S : Case Set
N : Number of partitions S
Pi : Proportion of Si to S

$$Gain\ (S,A) = Entropy(S) - \sum_{l-1}^{n} \frac{s1}{s} * Entropy(Si) \quad (2)$$

Information:
S : Case Set
A : Attributes
N : Number of partitions attribute A

The C4.5 algorithm is popularly used by many researchers to determine decisions by building decision trees. In simple terms, after determining the entropy and gain of the process, the next is to determine the attribute that is used as the root, after it is known which attribute is the root, the next process is to determine the branch for each root [15].

RESULT AND DISCUSSION

The first experiment was carried out using 15 data of cases of violations, further determined the values of entropy and gain. The highest gain value is then checked whether the entropy owned is worth 0 or is still worth. If the entropy is still valuable, the next derivative partition is carried out, namely the calculation of the entropy at the highest gain and the redetermination of the gain value and entropy.

Figure 2 is the result of the first partition, the highest gain value is in Statement 3 which is 0.5183. Each statement with an E value has an entropy value of 0, while for A, B, C, and D values it is not equal to 0. After obtaining the highest gain value, namely in Statement 3, it is necessary to recalculate the gain value and entropy and specifically for entropy which is not worth 0. Recalculation is called advanced partitioning.

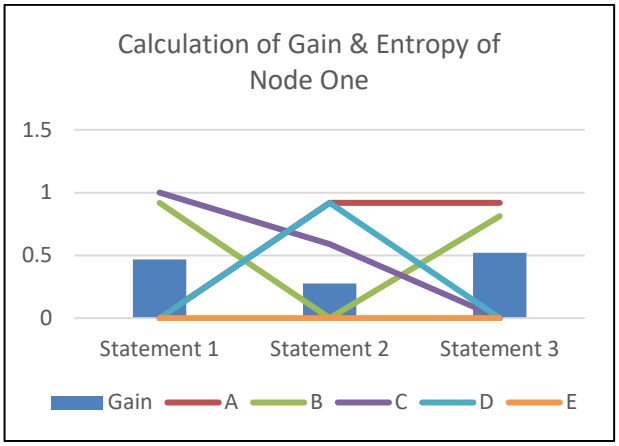

Figure 2. First Partition

The second partition is carried out in Statement 1 and Statement 2, the gain and entropy values are obtained, namely in Statement 1 the entropy value is worth 0 while the gain is worth 0.9183. In Statement 2, the entropy value of A is 1 while the gain value is 0.7850. The result of the second partition is in Table 2. Calculation Node 1.1 Statement 3A.

Table 2. Calculation of Node 1.1 Statement 3A

	Sum	Mild	Serious	Entropy
Statement 1				
A	1	1	0	0
B	1	0	1	0
C	1	1	0	0
D	0	0	3	0
E	0	0	3	0
Gain				0,9183
Statement 2				
A	2	1	1	1
B	0	0	0	0
C	0	0	0	0
D	1	1	0	0
E	0	0	0	0
Gain				0,7850

After the second partition, the next step is the calculation to get the result of node 1.2 based on the highest gain obtained on node 1.1. The derivative partition was carried out three times and obtained the highest gain of 0.8516 in statement 2 with an entropy value of 0. Next comes the decision tree like Figure 3.

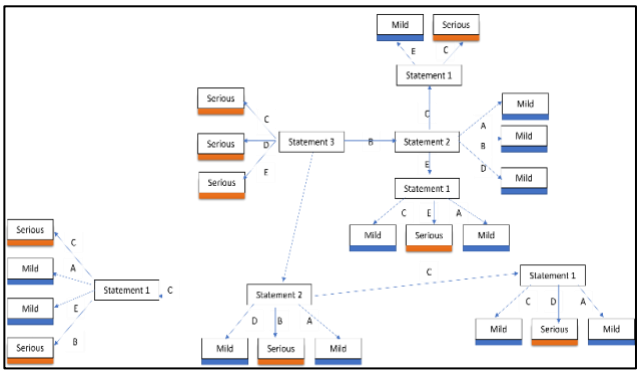

Figure 3. Decision Tree

Based on the decision tree, namely the final node that has been generated is then extracted into a rule or it can be concluded as follows:

1. if Statement 3 is of value A Statement 2 is of value A, then the type of violation is **Mild**
2. if Statement 3 is worth A Statement 2 is worth B, then the type of violation is **Serious**
3. if Statement 3 is worth A Statement 2 is worth C and Statement 1 is worth A then the type of violation **is Serious**
4. if Statement 3 is worth A Statement 2 is worth C and Statement 1 is worth B then the type of **Serious** violation
5. if Statement 3 is worth A Statement 2 is worth C and Statement 1 is worth C then the type of **Serious** violation
6. if Statement 3 is worth A Statement 2 is worth C and Statement 1 is worth E then the type of violation is **Mild**
7. if Statement 3 is worth A Statement 2 is worth D, then type of violation is **Mild**
8. if Statement 3 is worth A Statement 2 is worth E and Statement 1 is worth A then the type of violation is **Mild**
9. if Statement 3 is worth A Statement 2 is worth E and Statement 1 is worth C then the type of violation is **Mild**
10. if Statement 3 is worth A Statement 2 is worth E and Statement 1 is worth D then the type of **Serious** violation.
11. if Statement 3 is worth B Statement 2 is worth A, then the type of violation is **Mild**
12. if Statement 3 is worth B Statement 2 is worth B, then the type of violation is **Mild**
13. if Statement 3 is worth B Statement 2 is worth C and Statement 1 is worth C then the type of **Serious** violation.
14. if Statement 3 is worth B Statement 2 is worth C and Statement 1 is worth E then the type of violation is **Mild**
15. if Statement 3 is worth B Statement 2 is worth D, then the type of violation is **Mild**
16. if Statement 3 is worth B Statement 2 is worth E and Statement 1 is worth A then the type of violation is **Mild**
17. if Statement 3 is worth B Statement 2 is worth E and Statement 1 is worth C then the type of violation is **Mild**
18. if Statement 3 is worth B Statement 2 is worth E and Statement 1 is worth E then the type of **Serious** violation.
19. if statement 3 is of value C, then the type of **Serious** violation.
20. if statement 3 is of value D, then the type of **Serious** violation.
21. if statement 3 is of value, then the type of **Serious** violation.
22. if statement 3 is worth B Statement 2 is worth A, then the type of violation is **Mild**
23. if statement 3 is worth B Statement 2 is worth B, then the type of violation is **Mild**
24. if statement 3 is worth B Statement 2 is worth C, then the type of **Serious** violation.
25. if statement 3 is of value C, then the type of **Serious** violation.
26. if statement 3 is worth D, then the type of **Serious** violation.
27. if statement 3 is of value E, then the type of **Serious** violation.

Twenty-seven (27) rules are a class resulting from the classification of the types of violations committed can be **Serious** or **Mild.** The resulting rule can then be used as a pattern to determine the type of violation. The statement 3 attribute has a considerable influence in all partitioning performed. Based on the results of the C4.5 Algorithm that the statement 3 attribute is on the topmost node.

Validation of the rules that have been generated with the aim of evaluating the effectiveness of the C4.5 Algorithm in the classification of types of violations is found in figure 4. Validation is carried out twice, the first is validation with data that is used as a reference for algorithm calculations, namely with fifteen tests. The results obtained were thirteen tests successfully matched and two tests were not appropriate. The second validation is the conformity with the combination data against 125 combination case data. The results obtained were 106 appropriate case data and 19 cases were not appropriate.

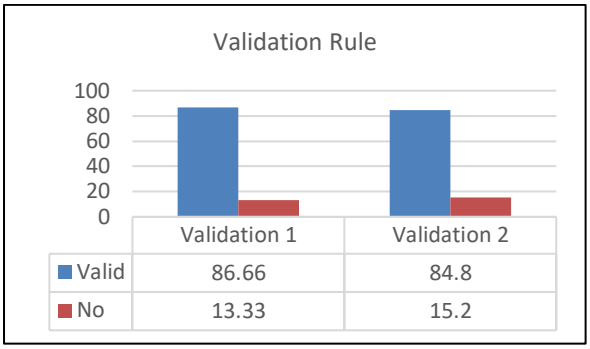

Figure 4 Rule Validation

CONCLUSION

Based on the results of research that has been carried out after determining the value of gain and entropy, three partitions were carried out, so a decision tree was obtained with 27 rules that can be used as a pattern to classify the type of violation. The results of this study can be concluded, namely the C4.5 Algorithm can be used to classify types of violations with an accuracy level above 80%. This was obtained from the first validation, which was 86.66% and the second validation was 84.8%. From the three attributes used, namely Statement 1, Statement 2, and Statement 3, it was found that Statement 3 had a major effect on each partition, this is evidenced by Statement 3 being on the top node. So that the C4.5 Algorithm not only has good performance on the classification of types of violations but can also produce an excellent decision tree.

REFERENCES

[1] R. Benkercha and S. Moulahoum, "Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system," *Sol. Energy*, vol. 173, no. April, pp. 610–634, 2018, doi: 10.1016/j.solener.2018.07.089.

[2] N. Asiah and D. S. Rini, "Pengaruh Bystander Effect Dan Whistleblowing Terhadap Terjadinya Kecurangan Laporan Keuangan," *Nominal, Barom. Ris. Akunt. dan Manaj.*, vol. 6, no. 1, 2017, doi: 10.21831/nominal.v6i1.14336.

[3] S. Moral-García, C. J. Mantas, J. G. Castellano, and J. Abellán, "Using Credal C4.5 for Calibrated Label Ranking in Multi-Label Classification," *Int. J. Approx. Reason.*, vol. 147, pp. 60–77, 2022, doi: https://doi.org/10.1016/j.ijar.2022.05.005.

[4] T. H. Apandi, R. B. Maulana, R. Piarna, and D. Vernanda, "Menganalisis Kemungkinan Keterlambatan Pembayaran Spp Dengan Algoritma C4.5 (Studi Kasus Politeknik Tedc Bandung)," *J. Techno Nusa Mandiri*, vol. 16, no. 2, pp. 93–98, 2019, doi: 10.33480/techno.v16i2.659.

[5] E. Elisa, "Analisa dan Penerapan Algoritma C4 . 5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT . Arupadhatu Adisesanti," *JOIN*, vol. 2, no. 1, pp. 36–41, 2017.

[6] H. Yusti and K. Ameliza, "PENERAPAN ALGORITMA C 4.5 UNTUK PENETUAN KRITERIA ANGGOTA LAYAK PINJAM BERDASARKAN AD/ART KOPERASI," *J. Ilmu Komput. dan Bisnis*, vol. 9, no. Nov 2018, pp. 2044–2050, 2018, [Online]. Available: https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part.

[7] A. R. Panhalkar and D. D. Doye, "Optimization of Decision Trees using Modified African Buffalo Algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, 2021, doi: 10.1016/j.jksuci.2021.01.011.

[8] L. N. Rani, "Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit," *INOVTEK Polbeng - Seri Inform.*, vol. 1, no. 2, p. 126, 2016, doi: 10.35314/isi.v1i2.131.

[9] Y. Chen and Y. Zhou, "Machine learning based decision making for time varying systems: Parameter estimation and performance optimization," *Knowledge-Based Syst.*, vol. 190, pp. 1–10, 2020, doi: 10.1016/j.knosys.2020.105479.

[10] H.-B. Wang and Y.-J. Gao, "Research on C4.5 algorithm improvement strategy based on MapReduce," *Procedia Comput. Sci.*, vol. 183, pp. 160–165, 2021, doi: https://doi.org/10.1016/j.procs.2021.02.045.

[11] J. Yan, Z. Zhang, K. Lin, F. Yang, and X. Luo, "A hybrid scheme-based one-vs-all decision trees

for multi-class classification tasks," *Knowledge-Based Syst.*, vol. 198, p. 105922, 2020, doi: 10.1016/j.knosys.2020.105922.

[12] M. Calis *et al.*, "Algorithms for the management of frontal sinus fractures: A retrospective study," *J. Cranio-Maxillofacial Surg.*, vol. 50, no. 10, pp. 749–755, 2022, doi: https://doi.org/10.1016/j.jcms.2022.09.007.

[13] H. Huang, H. Wang, and M. Sun, "Incomplete data classification with view-based decision tree," *Appl. Soft Comput. J.*, vol. 94, p. 106437, 2020, doi: 10.1016/j.asoc.2020.106437.

[14] X. Meng, P. Zhang, Y. Xu, and H. Xie, "Construction of decision tree based on C4.5 algorithm for online voltage stability assessment," *Int. J. Electr. Power Energy Syst.*, vol. 118, no. July 2019, p. 105793, 2020, doi: 10.1016/j.ijepes.2019.105793.

[15] J. Shanthi, D. G. N. Rani, and S. Rajaram, "A C4.5 decision tree classifier based floorplanning algorithm for System-on-Chip design," *Microelectronics J.*, vol. 121, p. 105361, 2022, doi: https://doi.org/10.1016/j.mejo.2022.105361.