Classification system model for project sustainability

S Hendra [a,1,*], H R Ngemba [a], R Azhar [a], R Laila [a], N P Domingo [b], R Nur [c]

[a] Information Technology Department Universitas Tadulako Palu, Indonesia
[b] Asian Language and Cultures Department, University of California, Los Angeles, California
[c] Public Health Department, Universitas Tadulako, Palu, Indonesia
[1] syaiful.hendra.garuda@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received September 3, 2022
Revised November 15, 2022
Accepted December 6, 2022
Available online December 31, 2022

Keywords
Decission Tree
Data Mining

ABSTRACT

One of the problems faced by the state-owned electricity company (PT. PLN) in Indonesia is the difficulty of monitoring the progress of an ongoing project so that it requires a technology that can help project managers in monitoring project implementation. The data in this study consisted of 117 Win project data and 89 Lose project data with a total of 206 data. The system development used extreme programming with algorithmic testing, namely the configuration matrix. The result of this research showed that the model could produce an accuracy of 92.68% with an error percentage of 7.32%, which means that the model produced good accuracy in implementing the C4.5 algorithm in recognizing patterns of project development. The first implication of the proposed approach is that it can establish project work monitoring services. The second implication is that project managers can improve company performance.

1. Introduction

PT. Perusahaan Listrik Negara (PLN/State-Owned Electricity Company) is a state-owned company. PLN is a company engaged in the electricity sector, starting from operating power plants to transmitting to people throughout Indonesia. As a state-owned company, service quality must be one of the priorities given by companies to the people in Indonesia [1], [2]. Besides the main functions and tasks of the Generating and Network Project Implementing Units, they also have duties and responsibilities ranging from land acquisition, supervision of the construction period to the completion of mass contracts for substation construction projects, transmission lines and generators throughout Indonesia as well as establishing relationships with stakeholders and the community [3]. Based on roadmap for developing new and renewable energy (EBT) by PLN, there are 512 projects in 2018; 774 projects in 2019; 1,040 projects in 2020; 1,438 projects in 2021; 996 projects in 2022; 871 projects in 2023; 1,299 projects in 2024; 7,323 projects in 2025; 20 projects in 2026; 639 projects in 2027 with a total of 14,912 projects [4].

Due to the large number of projects to be implemented, PT.PLN (Persero) is required to tighten the selection of contractors, and supervise the implementation of electricity projects in the future [5],[6]. This aims to avoid delays in project completion, as happened in the implementation of the power plant development acceleration program. Supervision is necessary to ensure that the implementation of organizational or institutional activities runs according to planning and in accordance with applicable regulations [7]–[9]. The supervisory process is carried out every week to monitor the progress of project work and see any obstacles that occur in the field. In the supervisory process, one of the problems faced by the manager is that it is very difficult to know the progress of the project because there is no media that facilitates reporting and monitoring of project progress to

http://doi.org/10.31763/aet.v1i3.689

find out the next decision to be made. Therefore, the manager also does not know Win (continued) or Lose (replaced) information. Project supervision is always carried out every week to find out the progress of the project, but there are frequently miscalculations of project work progress. If this problem is allowed to continue, project work cannot be completed on time according to the target and project follow-up will be hampered. This is because the manager does not know the problems in the field quickly and can experience large cost losses. Based onthe existing problems, technology support is needed to facilitate monitoring of existing projects because with technology, work will be more efficient and effective (Cascio and Montealegre, 2016; Linton and Solomon, 2017).

Several studies use data mining technology to help companies to improve company performance, as was done by [14]–[16]. This study aims to help state-owned companies, especially project managers, in monitoring the development of project work to determine the results of the project's Win/Lose target. Thus, it can help project managers to solve project monitoring problems which will be useful for them in making decisions related to project performance. In this study, the C4.5 algorithm will be used to identify patterns of project development classifications. The dataset used in this work will be described in Section 2. Section 3 discusses the application of project progress classifications and algorithm testing. Finally, this paper is concluded in Section 4.

2. Method

This research was conducted to assist project managers in supervising and controlling the project being worked on. The research was conducted in Central Sulawesi, Indonesia. The types of data needed in this study were divided into 2 (two) categories, namely the primary data that the researcher used in this study including project data, the flow of the project inspection and control process, and the problems faced by managers in the field. This research was conducted using the Extreme Programming software development methodology. Extreme Programming is an approach or software development model that aims to simplify various stages in the development process so that it becomes more adaptive and flexible[17]. Researchers conducted observations at PLN Central Sulawesi with 206 project data as samples in this study. The data consisted of 117 Win project data and 89 Lose Project data. The algorithm method used in this study wasthe C4.5 algorithm. In general, the C4.5 algorithm according to[18] is to build a decision tree as follows:

- Select attributes as root
- Create a branch for each value
- Divide cases into branches
- Repeat the process for each branch until all cases on the branch have the same class

The C4.5 algorithm is one algorithm that has been widely used[19], especially in the machine learning area which has several improvements, including:

- The C4.5 algorithm calculates the gain for each attribute and the attribute that has the highest value will be selected as a node. The use of this gain improves the weakness of the ID3 which uses information gain.
- Pruning can be done at the time of tree construction or when the tree building process is complete.
- Able to handle continuous attribute.
- Able to handle missing data.
- Able to generate rules from a tree.

Entropy is a parameter to measure the level of diversity (heterogeneity) of a data set [20]. The greater the entropy value, the greater the diversity of a data set. To calculate entropy, the equation(1) was used.

$$Entropy\ (S) = \sum_{i=1}^{n} -Pi\ x\ log_2\ (Pi) \tag{1}$$

Note:

S : Case Set

n : Number of partitions S

Pi : The propotion of Si to S

Meanwhile, the formula for entropy in each variable is:

$$Information\ Gain\ (S, A) = Entropy\ (S) - \Sigma_v Values\ (A) \frac{|S_v|}{|S|} I(S_v) \tag{2}$$

Note :

A : Variable

V : Possible values for variable A

$|S_v|$: Number of Samples for the value v

|S| : Number of Samples for all data samples

Entropy (Sv): Entropy for a sample that has a value of v

After calculating the Entropy value, the calculation of the Information Gain value can be seen in equation 2.

$$Entropy\ (S) \sum_{i=1}^{n} \frac{|S_v|}{|S|} \times Entropy(S_i) \tag{3}$$

Note :

S : Case Colection

A : Attribute Data

n : the number of partitions attribute A

$|S_i|$: number of cases on the ith partition

$|S|$: number of cases in S

The last stage wastesting. The system testing stage was carried out to find out what errors appeared when the application was running and find out whether the system being built was in accordance with user needs. The test method used at this stage was a configuration matrix, which contained information about the actual and predictive classes provided by the classification system [21]. Furthermore, based on descriptive data, the results of the interviews were analyzed and a table of cases was made which would serve as the base case in this study. Figure 1 shows the implementation of the model in this study

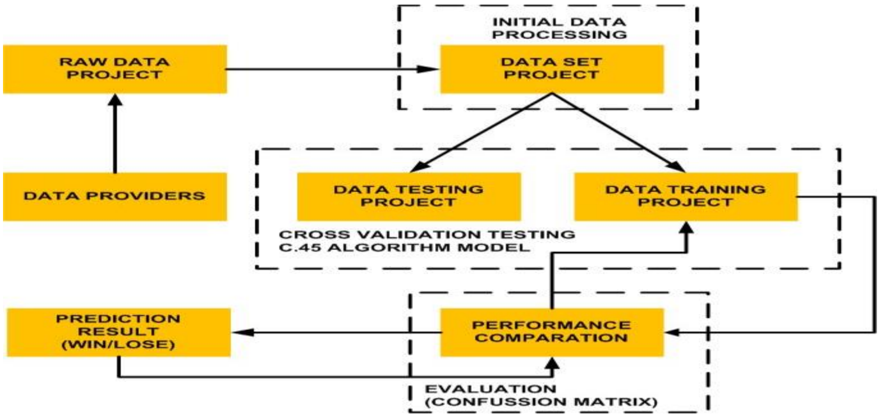

Fig. 1. The Model of Classification System For Project Sustainability

3. Results and Discussion

Comparison of the results of manual calculations and the results of system calculations was carried out to obtain class values whether it could be classified properly. Testing was done by comparing the contractor surveyor's original data on the development of work and system results. The description of the distribution of training data and test data is shown in Table 1.

Table 1. Data sharing

Information	Number of Data
Total Data	206
Total WIN Data	117
Total LOSE Data	89
Win Train Data (80%)	94
Lose Training Data (80%)	71
Win Test Data (20%)	23

The following are the steps for manual calculations based on the data used by the system. The first step taken was counting the number of cases for the classification of achievement, the number of cases for the classification of not achieving, the entropy of all cases and cases divided by attribute. After that, the gain for each attribute was calculated. The calculation results are shown in the Table. II. The total entropy line in Table II is calculated by equation 1 as follows:

Entropy (Total)

$$-\left(-\frac{71}{165} * log_2\left(\frac{71}{165}\right)\right) + \left(-\frac{94}{165} * log_2\left(\frac{94}{165}\right)\right) \qquad = 0.9859$$

Entropy Project Type

Entropy (Total,Contruction)

$$= \left(-\frac{66}{159} * log_2\left(\frac{66}{159}\right)\right) + \left(-\frac{93}{165} * log_2\left(\frac{93}{165}\right)\right) \qquad = 0.9791$$

Entropy (Total,Maintanance)

$$= \left(-\frac{5}{6} * log_2\left(\frac{5}{6}\right)\right) + \left(-\frac{1}{6} * log_2\left(\frac{1}{6}\right)\right) \qquad = 0.65$$

Entropy Work/Day

Entropy (Total,Match)

$$= \left(-\frac{70}{70} * log_2\left(\frac{70}{70}\right)\right) + \left(-\frac{0}{70} * log_2\left(\frac{0}{70}\right)\right) \qquad = 0$$

Entropy (Total, Not Match)

$$= \left(-\frac{1}{95} * log_2\left(\frac{1}{95}\right)\right) + \left(-\frac{94}{95} * log_2\left(\frac{94}{95}\right)\right) \qquad = 0.0843$$

Entropy Duration Work

Entropy (Total > 1 Week)

$$= \left(-\frac{1}{95} * log_2\left(\frac{1}{95}\right)\right) + \left(-\frac{94}{95} * log_2\left(\frac{94}{95}\right)\right) \qquad = 0.0843$$

Entropy (Total < 1 Week)

$$= \left(-\frac{70}{70} * log_2\left(\frac{70}{70}\right)\right) + \left(-\frac{0}{70} * log_2\left(\frac{0}{70}\right)\right) \qquad = 0$$

Entropy Status

Entropy (Total,Reached)

$$= \left(-\frac{71}{71} * log_2\left(\frac{71}{71}\right)\right) + \left(-\frac{0}{70} * log_2\left(\frac{0}{70}\right)\right) \qquad = 0$$

Entropy (Total,Not Reached)

$$= \left(-\frac{0}{94} * log_2\left(\frac{0}{94}\right)\right) + \left(-\frac{94}{94} * log_2\left(\frac{94}{94}\right)\right) \qquad = 0$$

Entropy Follow Up

Entropy (Total, Continuin Work)

$$= \left(-\frac{68}{73} * log_2\left(\frac{68}{73}\right)\right) + \left(-\frac{5}{73} * log_2\left(\frac{5}{73}\right)\right) \qquad = 0.3603$$

Entropy (Total, Solving Problem) $= \left(-\frac{3}{92} * \, log_2\left(\frac{3}{92}\right)\right) + \left(-\frac{89}{92} * \, log_2\left(\frac{89}{92}\right)\right)$ = 0.2073

Entropy Percentage

Entropy (Total, > 50 %) $= \left(-\frac{70}{70} * \, log_2\left(\frac{70}{70}\right)\right) + \left(-\frac{0}{70} * \, log_2\left(\frac{0}{70}\right)\right)$ = 0.2073

Entropy (Total, < 50 %) $= \left(-\frac{1}{95} * \, log_2\left(\frac{1}{95}\right)\right) + \left(-\frac{94}{95} * \, log_2\left(\frac{94}{95}\right)\right)$ = 0.0843

Then, the gain calculation was done as follows.

Gain (Total,Attribute) $Gain(S, A) = Entropy\ (S) - \sum_{i-1}^{n} \frac{|S_i|}{|S|} * Entropy\ (Sn)$

Gain (Total,Project Type) $= 0.9859 - \left(\frac{159}{165} * 0\right) + \left(\frac{6}{165} * 0\right) = 0.0188$

Gain (Total,Work/Day) $= 0.9859 - \left(\frac{70}{165} * 0\right) + \left(\frac{95}{165} * 0.0843\right) = 0.9374$

Gain (Total,Work Duration) $= 0.9859 - \left(\frac{70}{165} * 0\right) + \left(\frac{95}{165} * 0.0843\right) = 0.9374$

Gain (Total,Status) $= 0.9859 - \left(\frac{71}{165} * 0\right) + \left(\frac{94}{165} * 0\right) = 0.9859$

Gain (Total,Follow Up) $= 0.9859 - \left(\frac{73}{165} * 0.3603\right) + \left(\frac{92}{165} * 0.2073\right) = 0.7110$

Gain (Total,Percentage) $= 0.9859 - \left(\frac{70}{165} * 0\right) + \left(\frac{95}{165} * 0.0843\right) = 0.9374$

Table 2. Calculation Node 1

	Cases Number	Win	Lose	Entropy	Gain
Total	165	71	94	0.9859	
		Project Type			
Construction	159	66	93	0.9791	0.0188
Maintenance	6	5	1	0.65	
		Work/Day			
Match	70	70	0	0	0.9374
Not Match	95	1	94	0.0843	
		Work Duration			
> 1 week	95	1	94	0.0843	0.9374
< 1 week	70	70	0	0	
		Status			
Reached	71	71	0	0	0.9859
Not Reached	94	0	94	0	
		Follow Up			
Continuin Work	73	68	5	0.3603	0.7110
Solving Problem	92	3	89	0.2073	
		Percentage			
> 50 %	70	70	0	0	0.9374
< 50 %	95	1	94	0.0843	

Based on Table 2, the results of the calculation of the 1st node show that the highest gain value is Status (0.9859), so the first node or root node is Status. Furthermore, by looking at the contents of the Status attribute, there was the value on each side, either the WIN or LOSE value was 0. Here, the decision tree making was complete. To find out the accuracy of the test results, confusion matrix was calculated so that it can be seen that this system wasable to classify how accurate the test data. The confusion matrix calculation can be seen in Table 3.

Table 3. Confusion Matrix

		Classification	
		WIN	LOSE
Observation Data	WIN	18	0
	LOSE	3	20

$$\text{Recall} \quad = \frac{18}{18+10} - \frac{18}{8} = 1 \times 100 = 100\%$$

$$\text{Precision} \quad = \frac{18}{18+10} - \frac{18}{8} = 1 \times 100 = 100\%$$

$$\text{Accuracy} \quad = \frac{18+20}{18+3+20+0} = \frac{38}{41} = 0.9268 \times 100\% = 92.68\%$$

$$\text{Error rate} \quad = \frac{0+3}{18+3+23+0} = \frac{3}{41} = 0.0732 \times 100\% = 7.32\%$$

The training data used in this study amounted to 165 data consisting of 71 data classified as Win and 94 data classified as Lose. Based on the training data, there were several parameters or information used to determine the classification process. These parameters included the project id, project name, project type, work/day, work duration, work status, follow-up and percentage. Meanwhile, in the multi-class classification form, the input data were classified into several classes. The form of multi-label classification was basically the same as multi-class where data were grouped into several classes, but in multi-label classification, data could be included in several classes at once. The last form of classification was hierarchical. Input data were grouped into several classes, however these classes couldbe regrouped into simpler classes hierarchically, for example in this study, the direction of movement was grouped into 12 directions which indeed could be simplified into 4 directions.

In performance measurement using confusion matrix, there are 4 (four) terms as a representation of the results of the classification process. The four terms are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Negative (TN) value is the number of negative data detected correctly, while False Positive (FP) is negative data but detected as positive data. Meanwhile, True Positive (TP) is positive data that is detected correctly. False Negative (FN) is the opposite of True Positive, so the data is positive, but is detected as negative data. The results of the confusion matrix calculation using the C4.5 method are shown in Table III with the number of packets detected as TrueNegative of 0 data, False-Positive of 20 data, False-Negative of 3 data and True-Positive of 18 data, yielding an accuracy value of 92.68%, precision 100%, recall 100% and error rate7.32%. The accuracy parameter is the percentage of the total data identified and assessed. This means that the value generated from the comparison of the data, or whether it is true WIN or LOSE identified, on the 92.68% data is classified as correct. The recall parameter is the deletion data that was successfully retrieved from the data relevant to the query. In binary classification, recall is known as sensitivity. The emergence of relevant data that is taken is to agree with the query, whichcan be seen by recall. Based onthe test results, it can be concluded that the system can produce an accuracy of 92.68% with an error percentage of 7.32%, which means that the system produces good accuracy for the implementation of the C4.5 algorithm in recognizing patterns of project development.

4. Conclusion

This research is very helpful for companies, especially project managers in knowing the development of project work by utilizing the C4.5 algorithm. To test the success of the C4.5 algorithm, we used the Confusion Matrix method with the test results obtained by a recall value of 100%, which indicates the success rate of the system in the deletion data that was successfully retrieved from data relevant to the query. The results of the recall value on the system test are very high, which means that the quality of the information displayed during the retrieval is very complete and relevant. Additionally,the precision value obtained is 100%, which means that the level of accuracy between the information requested by the user and the answer given by the system is 100%. The precision value obtained shows that the application is very useful and appropriate for end users in the monitoring process. Likewise, the result of testing an accuracy is92.68% with an error percentage of 7.32%, which means that the system produces a fairly good accuracy for the implementation of the C4.5 algorithm in recognizing patterns of project development. This study has several implications for the company.

The first implication of the proposed approach is that it can establish project work monitoring services. The second implication is that project managers can improve company performance.

Acknowledgment

We thank the Faculty of Engineering, Universitas Tadulako for providing financial support for this research with Grant Number : 4007/UN28//KP/2020. We also thank the head of the Information Technology department for their support and the researchers also thank all manager State Electricity Enterprise in Palu City for helping us in this research.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] A. B. Setyowati, "Mitigating energy poverty: Mobilizing climate finance to manage the energy trilemma in Indonesia," *Sustain.*, vol. 12, no. 4, 2020, doi: 10.3390/su12041603.

[2] I. I. of E.-I. G. on P. S. R. I. Economics, "Electricity Governance Initiative : Case Of Indonesia," 2019.

[3] International Labour Office (ILO), The electronics industry in Indonesia and its integration into global supply chains, no. Workin Paper No. 330. 2019.

[4] IESR, "A Roadmap for Indonesia ' s Power Sector," 2019.

[5] S. Hardjomuljadi, "Factor analysis on causal of construction claims and disputes in Indonesia (with reference to the construction of hydroelectric power project in Indonesia)," *Int. J. Appl. Eng. Res.*, vol. 9, no. 22, pp. 12421–12446, 2014.

[6] PwC Indonesia, "Power in Indonesia," 2017. [Online]. Available: https://www.pwc.com/id/en/energy-utilities-mining/assets/power/power-guide-2017.pdf.

[7] K. Manghani, "Quality assurance: Importance of systems and standard operating procedures," *Perspect. Clin. Res.*, vol. 2, no. 1, p. 34, 2016, doi: 10.4103/2229-3485.76288.

[8] N. Azad *et al.*, "Leadership and management are one and the same," *Am. J. Pharm. Educ.*, vol. 81, no. 6, 2017, doi: 10.5688/ajpe816102.

[9] M. Aidah, H. Rasmita Ngemba, and S. Hendra, "A study of barriers to e-commerce adoption among small medium enterprises in Indonesia," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1306, no. 1, pp. 75–80, 2017, doi: 10.1145/3124116.3124124.

[10] K. M. Wilburn and H. R. Wilburn, "The Impact Of Technology On Business And Society," *Glob. J. Bus. Res.*, vol. 12, no. 1, pp. 23–39, 2018.

[11] S. Hendra, H. Rasmita, and F. A. Masse, "Aplication for Mapping and Supporting Small and Medium Industries," *Tadulako Sci. Technol. J.*, vol. 1, no. 1, pp. 26–34, 2019.

[12] J. D. Linton and G. T. Solomon, "Technology, Innovation, Entrepreneurship and The Small Business—Technology and Innovation in Small Business," *J. Small Bus. Manag.*, vol. 55, no. 2, pp. 196–199, 2017, doi: 10.1111/jsbm.12311.

[13] W. F. Cascio and R. Montealegre, "How Technology Is Changing Work and Organizations," *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 3, no. June, pp. 349–375, 2016, doi: 10.1146/annurev-orgpsych-041015-062352.

[14] P. Maroufkhani, R. Wagner, W. K. Wan Ismail, M. B. Baroto, and M. Nourani, "Big data analytics and firm performance: A systematic review," *Inf.*, vol. 10, no. 7, pp. 1–21, 2019, doi: 10.3390/INFO10070226.

[15] M. M. Hasan, J. Popp, and J. Oláh, "Current landscape and influence of big data on finance," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00291-z.

[16] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, 2017, doi: 10.1016/j.jbusres.2016.08.001.

[17] M. Al-Zewairi, M. Biltawi, W. Etaiwi, and A. Shaout, "Agile Software Development Methodologies: Survey of Surveys," *J. Comput. Commun.*, vol. 05, no. 05, pp. 74–97, 2017, doi: 10.4236/jcc.2017.55007.

[18] C. J. Mantas and J. Abellán, "Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data," *Expert Syst. Appl.*, vol. 41, no. 10, pp. 4625–4637, 2014, doi: 10.1016/j.eswa.2014.01.017.

[19] M. Sadikin and F. Alfiandi, "Comparative Study of Classification Method on Customer Candidate Data to Predict its Potential Risk," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, p. 4763, 2018, doi: 10.11591/ijece.v8i6.pp4763-4771.

[20] C. Neto, M. Brito, V. Lopes, H. Peixoto, A. Abelha, and J. Machado, "Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients," *Entropy*, vol. 21, no. 12, 2019, doi: 10.3390/e21121163.

[21] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.