Village Status Classification Based Tree Algorithm

Nadhira Nur Salima[1], Ahmad Ilham[2]
{nadhira@unimus.ac.id[1], ahmadilham@unimus.ac.id[2]}

Department Informatics, Universitas Muhammadiyah Semarang, Indonesia[1,2]

Abstract. Knowing the status of the village is very important to how developed the village is. The Central Statistics Agency (BPS) has carried out a classification process using the traditional scoring method, making it difficult to provide timely data. In this study, decision tree algorithm will be applied. The algorithm will test on the Podes2016 dataset focused on the Daerah Istimewa Yogyakarta (DIY) region, where the accuracy matrix is used to measure the algorithm's performance. The results showed that the proposed algorithm highest accuracy (87,40%) than the two comparison algorithms, such as the ID3 (83,81%) and C4.5 (80,62%). It can be concluded that the proposed algorithm can score very well and has good accuracy.

Keywords: Village status classification; podes2016; tree algorithm

1 Introduction

Indonesia is one of the largest archipelagic countries in the world. Administratively, the territory of Indonesia is divided into several regional levels, namely provinces, districts/cities, sub-districts, and urban villages, which are the smallest administrative areas (BPS, 2010). In addition, Indonesia is also famous for its natural resources, but these have not been optimally utilized to create a more prosperous life for the people. Inequality of development is still one of the problems faced. To overcome this, the government has drawn up a development plan contained in the Nawacita. One of the points is to build Indonesia from the periphery by strengthening regions and urban villages within the framework of a unitary state. The development carried out applies a decentralized system, namely development that spreads to all corners of Indonesia. To realize equitable development planning, it is necessary to have a link between the urban village and the city. This, in line with Tarigan's (2003) research, through the agropolitan concept, emphasizes that they can achieve village development well if the village is linked to urban development in the region. The existence of village funds is a tangible form of supporting development in village areas, especially to increase access to connectivity.

The village development program based on invite law no. 6 of 2014 aims to develop villages with more advanced life, both social, economic, and environmental resilience. This program went well enough to give birth to several village areas that changed their status to developed villages. This causes a shift in determining the characteristics of rural and urban status. Therefore, there is a need for uniformity in the use of concepts, definitions, and criteria for urban and rural areas in Indonesia.

The current classification of urban villages still refers to the Kemendes publication in 2016. Until to 2022, Kemendes has not provided the latest data regarding village data updates in Daeras Istimewa Yogyakarta in Indonesia where the status of villages may have changed.

ICONSEIR 2021, December 21, Medan, Indonesia

DOI 10.4108/eai.21-12-2021.2317488

Currently, many villages are undergoing changes in an advanced direction such as building public facilities, strengthening the economy, and others. Each village has different social, economic, condition and access characteristics which will continue to change over time. These criteria are used by Kemendes as an indicator to classify areas into rural or urban classifications. According to Tarigan (2003), regional development planning includes various aspects that take into account the interrelated roles of villages and cities. So that the status of a village is easily known by the government which can be the basis for development planning in rural areas.

Daerah Istimewa Yogyakarta (DIY) is one area that has a major contribution to developing natural and cultural tourism in Indonesia. However, there is still inequality of development in DIY. The Gini Ratio in March 2020 was 0.434, or an increase of 0.006 points compared to September 2019 of 0.428, which made DIY the highest Gini ratio in Indonesia [1]. Based on the above background, this research's main objective is to algorithm the classification of village status in the DIY region.

Several studies have been conducted for the classification of the rural status. As reported by [2], "Classifying Urban Villages and Rural Villages in Klungkung District Using the Mamdani Method.". This study aims to classify urban and rural status in Klungkung Regency using the Mamdani method. This study uses secondary data sourced from the Central Bureau of Statistics of Klungkung Regency in 2016. The study results show that the Mamdani method resulted in 52 villages classified as urban villages and seven villages as rural villages with an accuracy rate of 93%. In addition, there are differences in the total score and village status between the results using the Mamdani method and the original data.

Another report is from [3]. They use Classical Quadratic Discriminant Analysis and Robust Quadratic Discriminant Analysis to classify rural status in Semarang Regency in the urban or rural village category. The data used is data collection of Potensi Desa (PODES) DIY Regency in 2011. Their research shows that 183 villages have rural status and 52 urban statuses with an accuracy rate of 87.23%. Meanwhile, the robust quadratic discriminant analysis resulted in a higher accuracy rate of 89.79%, where 167 villages had rural status and 68 had urban status.

This study aimed to compare the scoring algorithm conducted by BPS with the tree algorithm for the classification of rural status in DIY.

2 Methodology

2.1 Dataset

This study uses secondary data from data collection on village potential in the province of the Special Region of Yogyakarta (DIY) conducted by Kemendes in 2016. The data shows that 438 villages spread over five cities and 78 sub-districts, and village status (rural and urban). Table 1 shows the data variables used in this study.

Table 1. Description of the data variables is used.

Variables	Variables description	Measuring scale
Y	Rural status (labels)	Nominal
X1	Number of shopping group	Ratio
X2	Number of permanent markets	Ratio

Variables	Variables description	Measuring scale
X3	Number of junior high schools	Ratio
X4	Number of senior high school	Ratio
X5	Distance of the junior high schools to urban village office (kilometers (km))	Ratio
X6	Distance of the senior high schools to urban village office (km)	Ratio
X7	Distance of shopping group to urban village office (km)	Ratio
X8	Distance of permanent market to urban village office (km)	Ratio
X9	Percentage of household electricity users	Ratio
X10	Distance from hospital to the urban village	Ratio

2.2 Preprocessing

Data preprocessing is an early data mining technique to convert raw data into cleaner information that can be used for further modeling. The data used has missing values. Missing value often occurs when there is a problem in the collection process, such as an error in data entry. In this study, the data cleaning technique uses basic statistics to fill in the missing value with the mean value.

2.3 Algorithms tree based

Tree-based algorithms are regarded as one of the most effective and widely used supervised learning techniques. This algorithm provides high accuracy, stability, and interpretability for predictive algorithms. They can perform classification to map non-linear interactions very well. It is also easy to solve various problems (classification or regression). Notable decision tree algorithms include ID3 (Iterative Dichotomiser 3), C4.5 (successor of ID3), and CART (Classification And Regression Tree).

The ID3 algorithm is a math-based method to create a decision tree that can classify data objects that have classes. ID3 was first introduced by Quinlan (1979) in [4] . The rules generated by ID3 are hierarchical relations such as trees (having roots, vertices, branches, and leaves). Some researchers call the structure of ID3 a decision tree, but other researchers also call it a rule tree [5]. Information gain, commonly called gain info, is a separation criterion that uses entropy measurements.

The C4.5 algorithm is a derived classification model from a decision tree to produce a decision tree developed by Ross Quinlan [6]. C4.5 developed from previous ID3 algorithm. It can use the decision tree generated by C4.5 for classification, so C4.5 is often referred to as a statistical classifier. In 2011, the machine learning software Weka authors described the C4.5 algorithm as "an important decision tree program that is probably the most widely used machine learning workhorse in practice to date" [7]

The Classification and regression trees (CART) is a decision tree-based classification algorithm [8]. This algorithm is quite simple but very powerful. It aims to get an accurate data group as a classification characteristic that can describe the relationship between response variables and predictors. The resulting tree algorithm depends on the scale of the response variable; if the data response variable is continuous, then the resulting tree algorithm is regression trees (regression tree), while if the response variable has a categorical scale, then the resulting tree is classification trees.

2.4 Performance evaluation

In this study the confusion matrix is used as an performace evaluation of the model built. The confusion matrix summarizes the predicted results of the classification problem. The number of correct and incorrect predictions is summed up with calculated values and broken down by each class. The confusion matrix shows how your classification model gets confused when making predictions. It provides insight into the error made by the classifier and, more importantly, the type of error that is being made. Table 2 shows the confusion matrix.

Table 2. Confussion matrix.

	Predicted values		
Actual values	**Positive (P)**	**Negative (N)**	
Negative	True Positive (TP)	Fals Negative (FN)	**Sensitivity** $\frac{TP}{(TP+FN)}$
Positive	False Positive (FP)	True Negative (TP)	**Recall** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

$$\text{F1 Score} = 2\left(\frac{Recall \times Precission}{Recall + Precission}\right)$$

The definition of the term in the picture in Table 2, namely *Positive* (P) is positive observation; *Negative* (N) is observations are not positive or negative; *True Positive* (TP) is the observation is positive, and the prediction result should be positive; *False Negative* (FN) is positive observation, but negative prediction result; *True Negative* (TN) is The observation is negative and the prediction result should be negative; and *False Positive* (FP) is negative observation but positive prediction result.

Accuracy is represents the ratio of true (positive and negative) predictions to the overall data. Precision is the ratio of positive true predictions compared to the overall positive predicted outcome. Recall is the ratio of true positive predictions compared to the total number of true positive data. Specificity is the correctness of predicting negative compared to the overall negative data. F1 score is a weighted comparison of the average precision and recall.

3 Results and discussion

The average value for each group of urban and rural villages is obtained from the data used. It is said to be a rural village if it has a small number of educational facilities and the distance from the village to the school is quite far. Then, having health facilities in a quite far hospital can be traveled an average of 10.1 km, in contrast to urban villages, which only travel an average of 2.62 km to the village office. Furthermore, there are fewer shops and permanent

markets in rural villages for economic facilities, and they are far from the village office. A village with rural status has fewer user households than a village with urban status. Table 3, show

Table 3. The results of the average value of each variable.

Variables	Urban village	Rural village
X1	3,32	0,8
X2	0,64	0,52
X3	1,71	1,71
X4	0,97	0,97
X5	0,28	0,94
X6	0,9	5,1
X7	1	2,4
X8	1,33	2,57
X9	40,89	16,61
X10	2,62	10,1

The experiments are conducted using a computing platform based on 2.5 GHz Dual-Core Intel Core i5, 8 GB RAM, and macOS Catalina vers.10.15.7 64-bit operating system. The development environment is MS Visual Basic 6, PHP and MySql as database server.

First of all, from the overall data, we divided two are training (80% totaling 351 villages) and testing (20% totaling 87 villages). The training data aims to build a learning model, while testing aims to test the learning model. Table 4 shows the results of the classification using the CART algorithm, and Table 5 show the confusion matrix CART algorithm.

Table 4. Performance results of all the data using only CART algorithm.

Performance	Values	
	Training	Testing
Accuracy	0,874	0,908
Specificity	0,805	0,62
Precision	0,7801	0,6334
Recall	0,865	0,4153
F1-score	0,8662	0,8323

Table 5. Confusion matrix of all the data using only CART algorithm.

		Predicted		Total
		Urban village	Rural village	
Actual	Urban village	168	24	192
	Rural village	22	224	246
Total		190	248	438

As you can see in Table 5, which shows the results of overall data classification, there are 392 villages classified correctly according to the actual status. It can be interpreted that the CART's accuracy to the overall data is 89.50%. Then, the other validation measures are calculated.

Secondly, we compare CART with ID3 and C4.5 to determine which algorithm performs better. In a more detailed comparison, we present the comparisons in Table 6. The bold type

indicates the best value for each evaluation. As shown in Table 6, the first experiment (CART) outperformed the two comparison algorithms. Meanwhile, in the second experiment, the ID3 outperformed C4.5.

Table 6. Result of the comparison accuracy of three tree-based algorithms.

Algorithms	Accuracy
ID3	83,81%
C4.5	80,62%
CART	**87,40%**

In the last experiment, we compared the visualization of urban and rural village characteristics by classification results BPS and our work.

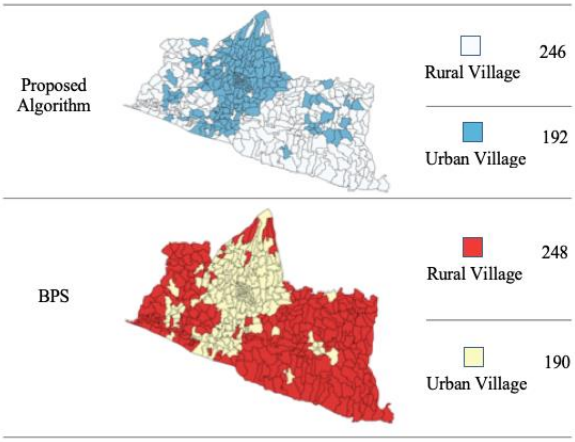

Fig. 1. Comparison of the classification results visualization between the proposed algorithm and BPS.

As shown in Figure 1, the results of the classification carried out by BPS are 190 urban villages spread over each district, such as 16 villages in Kulon Progo Regency, 54 villages in Bantul Regency, 8 villages in Gunungkidul Regency, 68 villages in Sleman Regency and 44 villages in the city of Yogyakarta. Meanwhile, 248 are classified as rural villages spread over five regencies. These include 72 villages in Kulon Progo Regency, 21 villages in Bantul Regency, 136 villages in Gunungkidul Regency, 18 villages in Sleman Regency, and only one village in Yogyakarta City.

In contrast to the proposed algorithm, the results show that 192 urban villages and 246 rural villages are spread across each district. For urban villages, there are 16 villages in Kulon Progo Regency, 50 villages in Bantul Regency, 17 villages in Gunungkidul Regency, 66 villages in Sleman Regency, and 43 urban villages in Yogyakarta City. As for rural villages are 72 villages in Kulon Progo Regency, 25 villages in Bantul Regency, 127 villages in Gunungkidul Regency, 20 villages in Sleman Regency, and only two villages in Yogyakarta City.

4 Conclusion

Based on the results of this study, there is a significant difference between the modeling carried out by BPS with the proposed tree-based capitalization results. From the performance approach, the proposed algorithm is better than BPS. Furthermore, the proposed algorithm is also compared with ID3 and C4.5, and the result is that the proposed algorithm is superior to the two previous algorithms.

Future research will be concerned with benchmarking the proposed method with other clustering techniques, such as DBSCAN, Fuzzy Cmeans, etc. and other meta-learning techniques, such as bagging and boosting also challenging to be studied in our future work.

Acknowledgments. We would like to express our gratitude to Computing Intelligent System Research Group (CISRG) for warm discussion about this research.

References

[1] B.-B. D. I. Yogyakarta, "Analisis Informasi Statistik Pembangunan Daerah Istimewa Yogyakarta 2016," *Yogyakarta: Badan Perencanaan Pembangunan Daerah-Badan Pusat Statistik Daerah Istimewa Yogyakarta*, 2016.

[2] N. K. Sumarwati, G. K. Gandhiadi, and T. B. Oka, "Mengklasifikasikan Desa Perkotaan dan Desa Perdesaan Di Kabupaten Klungkung Menggunakan Metode Mamdani," *E-Journal Matematika Vol. 7 (3)*, pp. 203–2010, 2018.

[3] A. S. Kurniasari, D. Safitri, and S. Sudarno, "Pemisahan Desa/kelurahan Di Kabupaten Semarang Menurut Status Daerah Menggunakan Analisis Diskriminan Kuadratik Klasik Dan Diskriminan Kuadratik Robust," *Jurnal Gaussian*, vol. 3, no. 1, pp. 1–10, 2014.

[4] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.

[5] S. Yang, J.-Z. Guo, and J.-W. Jin, "An improved Id3 algorithm for medical data classification," *Computers & Electrical Engineering*, vol. 65, pp. 474–487, Jan. 2018, doi: 10.1016/j.compeleceng.2017.08.005.

[6] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

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

[8] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "The CART decision tree for mining data streams," *Information Sciences*, vol. 266, pp. 1–15, May 2014, doi: 10.1016/j.ins.2013.12.060.