Implementation of decision tree using C5.0 algorithm in preference and electability survey results on regional head election in Aceh

M Marzuki, M Iqbal, A Nivada, H Sofyan, T Usman, N Nazaruddin, M Munawar and R Rasudin

1Department of Statistics, Universitas Syiah Kuala, Banda Aceh, Indonesia.
2Department of Political Science, Universitas Syiah Kuala, Banda Aceh, Indonesia.
3Department of Mathematics, Universitas Syiah Kuala, Banda Aceh, Indonesia.
4Department of Informatics, Universitas Syiah Kuala, Banda Aceh, Indonesia.

E-mail: marzuki@unsyiah.ac.id

Abstract. The decision tree is one of the methods of classification in data mining. There are many algorithms used to construct the tree model; one of them is C5.0 algorithm. The tree model with C5.0 algorithm was carried out based on the survey result dataset of the preference and electability of the regional head selection pre-campaign year 2018 in one of the districts in Aceh. The datasets consisted of 5 predictor variables, i.e. sub-districts, age, main occupations, highest education, and attracting factors from regional head candidate candidates. Variable categories of decisions ranged from candidates A, B, C, and D. The distribution of datasets was divided into training data and testing data using the k-fold cross-validation method. The optimum tree model formation was based on the accuracy value of model and coefficient of Kappa. The result showed that the best tree model was constructed using testing data on S = 10. The accuracy of the model and the Kappa coefficient were 0.8427 and 0.7208, respectively. There were three rules generated with five nodes. The main predictor variable contributing to the optimum model was the attracting factor of candidates and sub-districts.

1. Introduction

Classification is a data mining method that predicts label variables based on criteria variables [1]. Decision trees, artificial neural networks, Bayesian classifiers, Bayesian networks are some of the methods for data classification [2]. The decision tree is the most popular method and is easier to use in terms of application and interpretation of classification results [3][4].

The results of the classification of decision tree are presented in the form of tree diagrams [5], by grouping each object based on certain characteristics. These characteristics are derived from the attributes of the decision maker, in the form of categorical data, numerical data, or both [6].

Decision tree has developed in the preparation of its algorithm. Many algorithms that have been developed at this time, such as the ID3 algorithm (iterative dichotomiser 3), C4.5 algorithm, and C5.0
algorithm. The basis of all decision tree algorithms is the ID3 algorithm, whereas the C4.5 algorithm is the development of the ID3 algorithm, and the C5.0 algorithm is the development of the C4.5 algorithm [7].

The C 5.0 algorithm is already widely used today. One of the reasons for using the algorithm is to cope with the large data and the analysis process is faster [8]. This algorithm is also able to remove some attributes that do not have a major influence on classification classes [9].

This algorithm can be applied to any type of data [7]. One of the datasets that can be used is data survey preference and electability of regional head selection in one of the regencies/cities in Aceh Province. This study aims to find a model of decision trees and rules. By using this method, it is expected to determine the characteristics of each pair of candidates for regional head. These characteristics are in the form of voter demographic information and the strengths of each candidate pair [3][9][10].

2. Methods
This research uses survey results of preference and electability survey of regional head elections in one regency/city in Aceh Province in year 2018. There are 4 candidates for regional head pairs (A, B, C, and D) who are divided in the classification. The compiler characteristics of the candidate pairs of regional heads are the sub-district of the voter (X₁), the age of the voter (X₂), the voter's last education (X₃), the main occupation of the voter (X₄), and the interesting factors of the candidate pair of regional head (X₅).

Voters come from 8 districts. These sub-districts are denoted as K₁ to K₈. The age of the voters is categorized into 6 categories, namely 17-20 years (U₁), 21-30 years (U₂), 31-40 years (U₃), 41-50 years (U₄), 51-60 years (U₅), and above 60 years (U₆). The last voter educational background is categorized into 8 categories, namely Elementary school, Junior high school, Senior high/Vocational school, Diploma 1/Diploma 2, Diploma 3/Diploma 4, Bachelor degree, Master degree, Doctoral degree, and not go school. The main occupation of voters is categorized into 14 categories, namely farmers, fishermen, teungku/ustadz, civil servants, teachers/lecturers, informal sector or street vendors, small/medium entrepreneurs, large entrepreneur, BUMN/private employees, housewives, students, retired, casual work, and haven't found a job. Whereas there are 7 categories for interesting factors from the candidate pair of regional namely having a good track record (have no disability) (F₁), being religious (F₂), close to the people (F₃), committed to the people's economy (F₄), experienced in government (F₅), have vision, mission, and good work programs (F₆), and others (F₇).

The sampling in this study used stratified random sampling method, with the village being used as a stratum. Whilst for determining the number of samples of each village, a comparable allocation method is used, using the equation \( n_i = \frac{N_i \times n}{N} \) where \( n_i \) is the sample number of each village to-\( i \), \( N_i \) is the number of permanent voter lists (DPT) in the village to-\( i \), and \( N \) is the number of (DPT) for all villages, and \( n \) is the number of samples. There were 1000 respondents in determining the number of initial samples, and after the data cleaning, the number of samples that remained and could be processed was 890 respondents.

3. Results and Discussion
Election of regional heads in one regency/city in Aceh Province whose data was taken for this study was followed by 4 pairs of candidates for regional heads. The candidate pair of regional head has certain characteristics based on voter demographics and interesting factors of the candidate pair of regional head. The mode for each attribute of the characteristic compiler of the prospective regional head pairs are shown in Table 1.
The dominant respondents came from K2 sub-District (19.10%) with ages between 31-40 years (36.29%), with high school/ vocational educational background (45.17%) and dominated by farmers (33.26%). The majority of respondents chose regional head candidates because they were close to the people (often interact).

Data partitioning is done as a first step to dividing data into training data and testing data. This partition uses the k-fold cross validation method, where the number of groups (k) in this study is 10 [9]. Each data set is analyzed using decision tree C5.0. The results of this analysis are compared for each placement of testing data in data sets 1 through 10 (S1 to S10). Comparisons are made using the accuracy value and Kappa coefficient. The results of the analysis are shown in Table 2.

**Table 1.** The attribute mode of the characteristic compiler candidate for regional head (n = 890).

| No. | Attribute | Modes                  | Percentage |
|-----|-----------|------------------------|------------|
| 1   | X₁        | K2                     | 19.10%     |
| 2   | X₂        | 31-40 years old        | 36.29%     |
| 3   | X₃        | Senior high/ vocational school | 45.17% |
| 4   | X₄        | Farmer                 | 33.26%     |
| 5   | X₅        | Close to the people (often interact) | 21.24% |

**Table 2.** Comparison of C5.0 decision tree results based on the layout of the data testing.

| No. | Layout of data testing | Accuracy | Kappa coefficient |
|-----|------------------------|----------|-------------------|
| 1   | S1                     | 0.7416   | 0.5435            |
| 2   | S2                     | 0.7614   | 0.5786            |
| 3   | S3                     | 0.7727   | 0.5916            |
| 4   | S4                     | 0.7191   | 0.5022            |
| 5   | S5                     | 0.7640   | 0.5823            |
| 6   | S6                     | 0.7416   | 0.5465            |
| 7   | S7                     | 0.7000   | 0.4612            |
| 8   | S8                     | 0.7111   | 0.4857            |
| 9   | S9                     | 0.6629   | 0.4005            |
| 10  | S10                    | **0.8427** | **0.7208** |

The results of comparison of decision tree C5.0 based on the data testing layout obtained the highest accuracy value if the testing data in the 10th data set (S10), which is 0.8427. Similarly, for the Kappa coefficient, which is 0.7208. Further analysis was done using data testing in the 10th set and other data sets were used as training data. The Decision tree that is generated when data testing in the 10th set is presented in figure 1.
Figure 1. Decision tree using data testing in the 10th set.

This decision tree model produces 3 rules. The three rules contain the dominance of class A and B as the class decision maker. The resulting rules are shown in Table 3.

Table 3. Rule of decision tree results for data testing in the 10th set.

| No. | Rule | n  | Gain |
|-----|------|----|------|
| 1   | IF $X_5$ is F2, F5, THEN A | 212 | 0.879 |
| 2   | IF $X_5$ is F1, F3, F4, F6, F7, AND $X_4$ is K1, K4, K5, K6, K7, K8, THEN B | 461 | 0.719 |
| 3   | IF $X_5$ is F1, F3, F4, F6, F7, AND $X_4$ is K2, K3, THEN A | 128 | 0.684 |

4. Conclusion
The results of this study obtained the main predictor attribute in the optimum model is an interesting factor of the candidates and sub-districts origin of the voters. The best decision tree model is the model generated from testing data in the 10th data set. This is due to the accuracy of the model and the highest kappa coefficient, which is 0.8427 and 0.7208, resulting in the 10th data set. The number of rules produced is 3 rules with 5 nodes.
References

[1] Abidin T F, Rizal S, Iqbalsyah T M and Wahyudi R 2020 Decision tree classifier for university single rate tuition fee system Int J Business Intelligence and Data Mining 17 2

[2] Nisbet R, Miner G and Yale K 2018 Handbook of Statistical Analysis and Data Mining Applications Second Edition (US: Academic Press)

[3] Kocakoç I D and Keser I 2019 Exploring Decision Rules for Election Results by Classification Trees KnE Social Science 107-115

[4] Ying K, Ameri A, Trivedi A, Ravindra D, Patel D and Mozumdar M 2015 Decision tree-based machine learning algorithm for in-node vehicle classification IEEE Green Energy and Systems Conference 71-76

[5] Revathy R and Lawrence R 2017 Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data International Journal of Innovative Research in Computer and Communication Engineering 5(1) 50-58

[6] Han J, Kamber M and Pei J 2012 Data Mining: Concepts and Techniques 3rd ed. (US: Elsevier Inc)

[7] Pang S and Gong J 2009 C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks Systems Engineering - Theory and Practice 29(12) 94-104

[8] Munawar, Marzuki, Hafnani and Iqbal M 2019 Multiple Correspondence Analysis of Factors Affecting People's Choice for the Next Mayor Journal of Theoretical Statistics and Its Applications 19(2)

[9] Teyyare E and Avci M 2016 Local voting behaviour in local elections: the case of 2014 local elections and zonguldak city Siyaset, Ekonomi ve Yönetim Araştırmaları Dergisi 4(1) 51-76

[10] Williams G 2011 Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (New York: Springer)