A DECISION TREE CLASSIFIER FOR PREDICTING VOTER TURNOUT IN MALAYSIAN GENERAL ELECTION

Kuah Chong Hua
Jastini Mohd Jamil
*Izwan Nizal Mohd Shaharanee

DOI: https://doi.org/10.37178/ca-c.21.5.065

Kuah Chong Hua, *Department of Decision Science, School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia.*
chonghua@merdeka.org

Jastini Mohd Jamil, *Department of Decision Science, School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia.*
jastini@uum.edu.my

****Izwan Nizal Mohd Shaharanee*** *, Department of Decision Science, School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia.*
Corresponding author's email: nizal@uum.edu.my

Abstract.
Malaysia practices democracy in shaping the country's future. Hence, each citizen is entitled to have a vote in her election. Since independence, Malaysia had undergone 14 general elections. The 84.8% voters' turnout rate in 2013 general election is the highest turnout rate that has been recorded in the Malaysian election history. However, when voters' turnout rate is being compared with voting age population and the number of eligible voters, the actual participation rate is considered low. Thus, the main objective of this study is to predict Malaysian voters' turnout in the 2008 and 2013 general elections using classification tree algorithms. The datasets used in this study are the Asian Barometer Survey datasets. Datasets of 2014 and 2010 were used to examine the factors that determine voters' turnout in 2013 and 2008, respectively. Three selection decision tree algorithms used in this study are CHAID, CART, and C5.0. It is found that between these three methods, CHAID perform the best in predicting Malaysian voters' turnout during the general election. However, other feasible approaches such as Support Vector Machine (SVM), Random Forest and Boosting C5.0 can also be used and evaluated to predict voters' turnout.

Keywords: Voters' Turnout, Data Mining, Decision Tree Classification, CHAID, CART, C5.0

Introduction

Malaysia has been practicing democracy and allows all eligible voters to choose their desired leader in the country's general elections to shape the country's future. Malaysia has held 14 General Elections (GEs) after being granted its independence on 31 August 1957. Generally, Malaysia's GE is held every five years. Malaysian citizens aged above 21 years old is eligible to vote on the condition that they register with the Election Commission (EC). However, Malaysia does not enforce compulsory voting. In year 2013, it was estimated that 17.9 million individuals fulfilled the requirements to vote. However, only 13.2 million individuals registered with the EC as shown in Table 1 below. Since then, there was a significant increase in the percentage of registered voters, but it is still low when the figure is compared with the voting age population (VAP) for each year.

Fowler [1] observed that election outcomes and public policies can be changed when the number of voters increases during the election. Thus, some policies may only be beneficial to a particular group of people if the political participation or voter's turnout is imbalance. As a result, government's performance will drop due to the policies that are made were not tailored well enough to the specific cluster of people. Therefore, voters' turnout in any election plays an important role in deciding which political parties or leaders who will lead the country. It is crucial for both the ruling and the opposition political parties to identify who will vote because the numbers of voters will decide their victory.

Related Works

Voters' Turnout Model

There are three contextual political participation models that are used to explain the voting patterns in Malaysia. These models are the sociological model, political mobilisation model and psychological involvement model. The sociological model describes the voters, their socio-demographic background such as age, their gender and education background. Political mobilisation discloses the context of communication between political candidates and the citizens. On the other hand, psychological involvement model defines the interest about politics among voters.

Sociological Model

The sociological model suggests that socio-demographic variables such as age, gender and education are important factors in explaining voters' turnout. Numerous studies have been conducted to connect socio-demographic factors and voting behavior in the past decades. For instance, Mohd Hed and Grasso [2] revealed that the participation of young voters in any political engagement are less as compared to older voters. Although education, ethnic groups and gender do not affect political activism in Malaysia significantly, Blais [3] and Tenn [4] suggested that education is a factor that contributed to the rate of voters' turnout in the United States (US). In addition, Blais [3] pointed out that other variables such as government employees, marital status, religiosity and income were also correlated with political participation in the US. In a study conducted by Norris [5], demographic factors such as age, gender, income level, education level, club membership, religiosity and cultural attitude were significantly related to voters' turnout. Therefore, focusing on voters' societal context plays an important role in understanding the voters' turnout.

Political Mobilization

The concept of political mobilisation refers to activities that intend to motivate masses of organised or unorganised participants to express themselves and to undertake a particular political action to accomplish political aims. In short, it is practice of manipulating the existing distribution of power [6]. Political mobilisation model was developed by Rosenstone and Hansen [7]. They point out that campaigns and interpersonal conversations about politics had causal influence on voting. Political parties try to mobilise voters by using various campaigning methods from offline and traditional methods to online and modern campaigning ways. Traditional campaigning requires vast resources in order to obtain the support and trust from the voters [8]. On the other hand, modern campaigning methods use available information and communication techniques to influence voters. Welsh [9] claimed that modern campaigning that utilised social media platforms such as Facebook and Whatsapp encouraged more people to be engaged in politics and this indirectly brought changes to the political scenario in Malaysia.

Psychological Involvement

Voters' turnout is strongly related to the voters' political interest [10]. In other words, the more interest a voter in politics, the more likely the voter will vote in an election. Verba, Schlozman [11] found that educational and parental influence affected a person's political interest. In addition, the internet and social media are crucial variables in influencing voters' turnout. The Internet is the fastest way for political parties to engage with young voters. This, in turn, will contribute to a higher voter turnout among young voters [12]. A study conducted by Rauf, Hamid [13] showed that there was a positive relationship between the ability to access internet and participation in politics. Wang [14] suggested that the role of Internet as a medium in politics as a place of information seeking and opinion expressing was present in his study. Thus, the internet plays an important role to enhance as well as providing voters with the necessary information related to elections, political parties involved and candidates' background. Furthermore, it also stimulates the turnout rate and participation of young voters in an election. Therefore, the changes and improvements of communication and technology are influencing the electoral behavior of voters as well as the Malaysian political landscape.

Decision Tree Classification Algorithms

CART

The Classification and Regression Tree (CART) was developed by Breiman, Friedman [15]. CART can be applied in both categorical and continuous data. CART performs binary splits and using Gini or Entropy splitting rules to achieve an optimal purity node. It is a predictive model, which explains how an outcome variable's values can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable.

The following are the steps taken to implement CART algorithm:

(1) Start at the root node.

(2) Find the split set that minimises the sum of Gini indexes and use it to split the node into two child nodes to achieve an optimal purity node.

$$\text{Gini (D)}= 1-\sum_{i=1}^{n} (p_i)^2 \quad ; \qquad (3.1)$$

where p_i is the relative frequency of class i in D.

D is the dataset

(3) If a stopping criterion is reached, then exit.

(4) Prune the tree based on cost-complexity pruning.

(5) A test will be performed to examine the accuracy of the Model, if the model evaluation criteria is not satisfying, repeat step 2 – 4.

CHAID

Chi-square automatic interaction detection (CHAID) is a decision tree technique, based on adjusted significance testing (Bonferroni testing). The technique was developed in South Africa and was published in 1980 by Gordon V. Kass, who had completed a PhD thesis on this topic (Kass [16]. CHAID can only be conducted in categorical data. CHAID determines the size tree based on the Chi- Square test for independence.

The following are the steps taken in modelling using the CHAID algorithm:

(1) Start at the root node.

(2) Find the split set based on *Likelihood ratio Chi-Squared Statistics* and use it to split the node into two child nodes to achieve an optimal purity node.

Chi Square, $\chi^2 = \sum_{j=1}^{n} \frac{(e-o^2)}{e}$; (3.2)

where e is the expected value, o is the observe value
p_i is the relative frequency of class i in D
D is the dataset

(3) If a stopping criterion is reached, then exit.
(4) Prune the tree based on chi-square test for independence.
(5) A test will be performed to examine the accuracy of the Model, if model evaluation criteria is not satisfying, repeat step 2 – 4.

C5.0

Quinlan [17] combine and modify early version of C4.5 and ID3 and invented C5.0. The C5.0 offers new features such as the winnowing, boosting, generate smaller tree and unequal costs for different types of errors [18]. The C5.0 algorithm has become the industry standard for producing decision trees, because it does well for most types of problems directly out of the box. Compared to more advanced and sophisticated machine learning models (e.g. Neural Networks and Support Vector Machines), the decision trees under the C5.0 algorithm generally perform nearly as well but are much easier to understand and deploy. C5.0 uses information gain theory as the purity criterion to split the dataset and applied pessimistic pruning for the pruning process.

The following are the steps taken in implementing the C5.0 algorithm:
(1) Start at the root node.
(2) Find the split set based on *Entropy* measure and use it to split the node into two child nodes to achieve an optimal purity node.

Entropy (D)= - $\sum_{i=0}^{n} -p_i \, log_2 \; p_i$; (3.3)

where p_i is the relative frequency of class i in D
D is the dataset

(3) If a stopping criterion is reached, then exit.
(4) Prune the tree based on pessimistic pruning.

A test will be performed to examine the accuracy of the Model, if model evaluation criteria is not satisfying, repeat step 2 – 4.

Research Procedures

Data

The Asian Barometer Survey (ABS) collected public opinion on issues such as political values, and governance in 14 East Asia countries. Therefore, the dataset used in this study are secondary data that came from the third and fourth waves of ABS. The third wave and fourth wave of ABS were carried out in October 2010 and 2014, respectively. These surveys were usually carried out 17 or 18 months after any Malaysian general elections. Datasets of 2014 and 2010 were used to examine the factors that determines voters' turnout in 2013 and 2008, respectively.

Table 2 below, summarises the dimensions, attributes, measure level of the dataset that are selected for this study. Participation in the general election is the target variable in this study while the other 19 attributes are treated as predictor variables.

Table 2: Dimensions, Attributes and Measure Level of Dataset

	Dimensions	Attributes	Measurement Level
1	Participation in Elections	– Did you vote in the … General Election?	– Nominal
2	Socio-economic background	– Gender – Ethnic group – Age group – Marital status – Highest education level – Religion – Employment status – Settlement	– Nominal – Nominal – Ordinal – Nominal – Ordinal – Nominal – Nominal – Nominal
3	Political Mobilization	– Attend a campaign meeting or rally? – Try to persuade others to vote for a certain candidate of party? – Try to help or work for a party or candidate running in the election? – **Did you or a member of your family receive BR1M or any other government hand-out before the election period?* – Receive any reward or anything in exchange for your vote or support from any party?	– Nominal – Nominal – Nominal – Nominal – Nominal
4	Psychological Involvement	– How interested would you say you are in politics? – How often do you follow news about politics and government? – When you get together with your family members of friends, how often do you discuss political matters? – How often do you use the Internet including social media networks to find information about politics and government?	– Nominal – Ordinal – Ordinal – Ordinal

* These questions only appear in ABS wave 3 datasets.

Experimental Setting

The flow in this study is divided into five phases as shown in Figure 1 below. Phase 1 of this study is the pre-processing the dataset. The main reason of the existence of missing values is due to respondents refuse or unable to respond to the survey question. The missing values usually recorded as “No Response” (NR) or “Don’t Know” (DK) and will be filtered from the dataset. In Phase 2, the data is partitioned into training set and testing set with the ratio of 70% and 30% respectively. Testing set is used to validate the pattern generated from the training sample. The records in the training set are selected through a simple random sampling method. In Phase 3, the classification model of the datasets are being built. The classification model include CART, CHAID and C5.0. The control parameter of each model is summarized in Table 3, Table 4 and Table 5 respectively. In Phase 5, the model is evaluated through accuracy, sensitivity, specificity, positive prediction value (precision), negative prediction value, Area Under ROC curve (AUC). Lastly, Phase 5 is the discovery phase where voters’ turnout is being predicted.

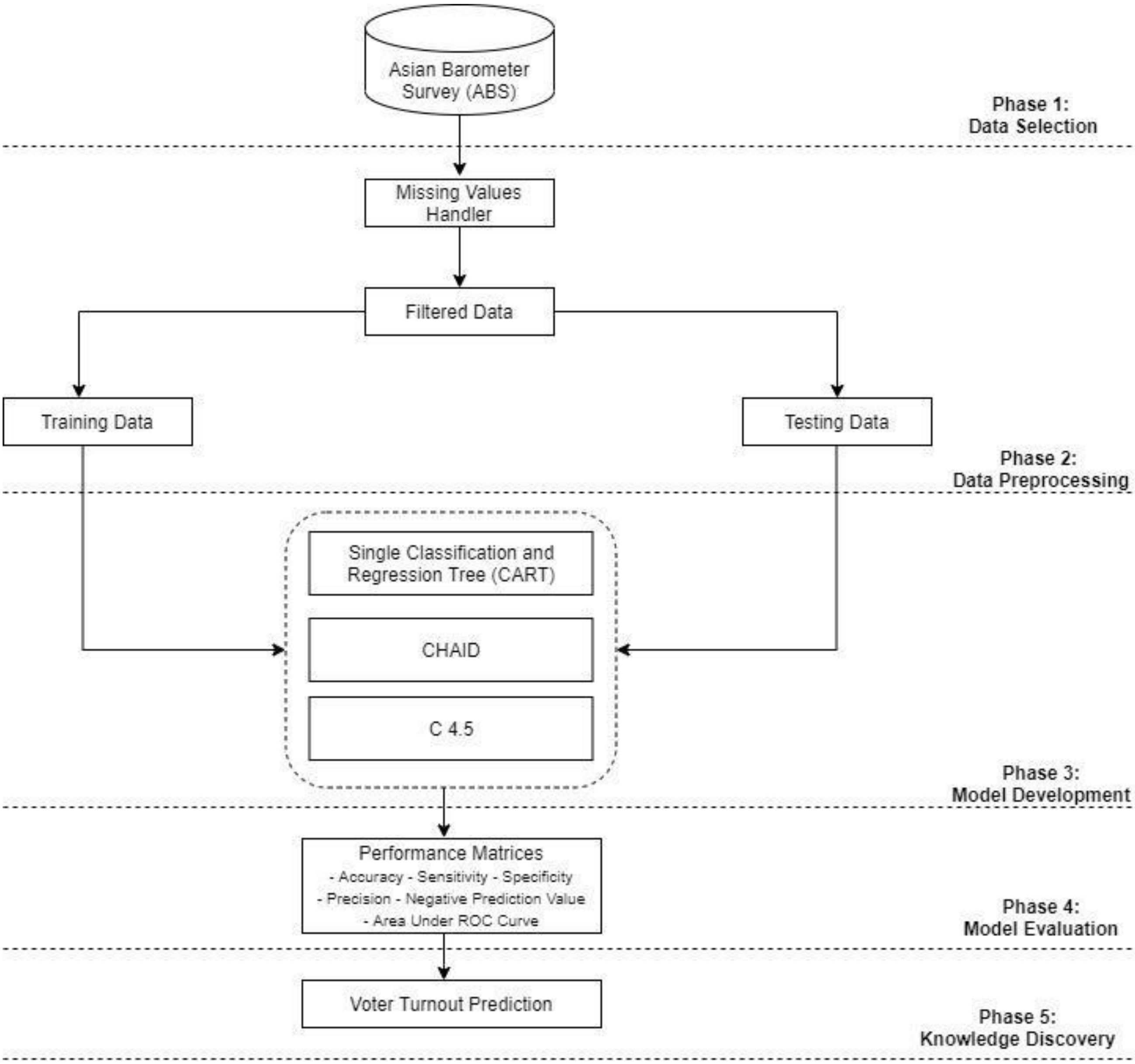

Figure 1. The Five Phases of the Experiment

Table 3: Control parameters for CART algorithm

Parameters	Value	Description
minsplit	20	The minimum number of observations that must exist in a node, in order for a split
minbucket	7	The minimum number of observations in any terminal node (which also equivalent to 1/3 of the minsplit)
maxcompete	4	The number of competitor splits retained in the output.
maxsurrogate	5	The number of surrogate splits retained in the output.
usesurrogate	2	Controls how surrogates are made use of in the model
xval	20	The number of cross-validations
surrogatestyle	0	Controls the selection of a best surrogate.
maxdepth	30	Limit the depth of a tree
cp	0.01	Control the size of the tree in order to select an optimal tree size.

Table 4: Control parameters for CHAID algorithm

Parameter	Value	Description
alpha 2	0.05	Level of significance used for merging of predictor categories
alpha 4	0.05	Level of significance used for splitting of a node in the most significant predictor
minprob	0.001	Minimum frequency of observations in terminal nodes
minsplit	20	Number of observations in split response at which no further split is desired
minbucket	7	Minimum number of observations in terminal nodes
Stump	False	Only root node splits are performed

Table 5:
Control parameters for C5.0 algorithm

Parameter	Value	Description
Winnow	TRUE	A logical: should predictor winnowing (i.e feature selection) be used?
noGlobalPruning	TRUE	A logical to toggle whether the final, global pruning step is needed to simplify the tree.
CF	0.5	A number in (0, 1) for the confidence factor. lower factor levels will likely prune away the leaves which over specify the classification
Mincases	7	An integer for the smallest number of samples that must be put in at least two of the splits
fuzzyThreshold	TRUE	A logical toggle to evaluate possible advanced splits of the data.
earlyStopping	TRUE	A logical to toggle whether the internal method for stopping boosting should be used.

Evaluation Metrics

In this study, two evaluation metrics are being used to measure the performance of classification models. These evaluation metrics are the confusion matrix and area under curve (AUC) under receiver operating characteristic (ROC) (AUC under ROC).

Confusion Matrix

A confusion matrix is a table that provides information about the result of classification and actual values. Table 6 shows a sample of the confusion matrix. True negative (TN) and true positive (TP) represent the number of voters that have been correctly classified between the actual and predicted values. Whereas, false negative (FN) and false positive (FP) are the number of voters that have been falsely predicted or incorrectly classified. In addition to the confusion matrix, the performance metrics of a predictive model measures the accuracy rate, sensitivity rate, specificity, positive predictive values and negative predictive values. The formula of the performance metrics summarized in Table 7 below.

Table 6: The Confusion Matrix Table

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Table 7: The Performance Metrics

Performance Metric	Formula
accuracy rate	$\frac{TN+TP}{TN+FP+FN+TP}$
sensitivity rate	$\frac{TP}{FN+TP}$
specificity	$\frac{TN}{TN+FP}$
positive predictive value	$\frac{TP}{FP+TP}$
negative predictive value	$\frac{TN}{TN+FN}$

AUC under ROC

The Receiver Operating Characteristic (ROC) can be illustrated in two-dimensional graph where sensitivity rate is plotted against one minus specificity rate (1- specificity). The graph contains values from 0 to 1 in each of the axis. Each point on the ROC curves provide the true-positive rate and false-positive rate and visually it in attractive way to summarize the accuracy of predictions. Using the same graph, the area under the ROC curve is called an Area Under Curve (AUC). The closer the value to 1, the better the performance of a classifcation model.

Results

Figure 2(A) until Figure 2(F) illustrates Wave 3 and Wave 4 rules representation for CART, CHAID and C5.0 classification models. Voters' age (a variable) is the first rule of the breakdown of the tree for the three classification models. In CART and C5.0, voters age is split into two groups, below 30 and above 30. On the other hand, CHAID split age variable into 21-30, 31-50 and above 50 years old. It is interesting to note that voters below the age of 30 years old who helped or worked for a candidate or party in the election are most likely to vote in Wave 3 datasets. In Wave 4 datasets, youngsters who attended a rally or political campaign had a high probability of voting in an election.

The summary of model performance for Wave 3 and Wave 4 datasets are represented in Table 8 and Table 9 respectively. Among the three algorithms, CHAID performs the best since it provides 84.03% and 85.57% of accuracy on Wave 3 and Wave 4 datasets, respectively. Although CART model shows high degree of sensitivity (97.31%), however, the specificity value is very low (33.85%) in Wave 3 datasets. On the other hand, in Wave 3 datasets, C5.0 has the highest rate of specificity (60%) but has the lowest rate of negative predictive value. However, in Wave 4 datasets, C5.0 produces the highest true positive value (97.15%) but has a very low specificity value (13.33%). In other words, C5.0 predicts 86.67% of respondents are most likely to turn out in an election. The specificity of CHAID performs better when compared to CART and C5.0. However, the sensitivity value of CHAID is slightly lower as compared to CART and C5.0. All models perform fairly good from the point of view of AUC under ROC curve in Wave 3 and Wave 4 datasets. The value of AUC ranges from 0.5 to 1.0. The best value of AUC under ROC curve is between 0.9 to 1.0 whereas 0.5 means no predictive value.

Table 8: Summary of model performance for Wave 3 datasets

		CART	CHAID	C5.0
1	Accuracy [95% CI]	82.99% [78.14%, 87.14%]	**84.03%** [79.28%, 88.06%]	81.60% [76.63%, 85.90%]
2	Sensitivity (True Positive)	97.31%	95.07%	87.89%
3	Specificity (True Negative)	33.85%	46.15%	60.00%
4	Positive Predictive Value (Precision)	83.46%	85.83%	88.29%
5	Negative Predictive Value	78.57%	73.17%	59.09%
6	Area Under ROC Curve (AUC)	72.50%	80.00%	81.00%

Note: CI refers to confidence intervals

Table 9: Summary of model performance for Wave 4 datasets

		CART	CHAID	C5.0
1	Accuracy [95% CI]	84.88% [80.24%, 88.79%]	**85.57%** [81.00%, 89.40%]	84.19% [79.48%, 88.19%]
2	Sensitivity (True Positive)	95.62%	92.43%	97.15%
3	Specificity (True Negative)	17.50%	42.50%	13.33%
4	Positive Predictive Value (Precision)	87.91%	90.98%	85.97%
5	Negative Predictive Value	38.89%	47.22%	46.15%
6	Area Under ROC Curve (AUC)	73.30%	76.1%	67.85%

Note: CI refers to confidence intervals

Conclusion

The main objective of this study is to predict Malaysian voters' turnout in a general election using classification tree algorithms. Three classification models, CART, CHAID and C5.0, are being used to examine the factors that determine the factors voters' turnout in 2008 and 2013 general elections in Malaysia. Among these three models, CHAID have the highest accuracy in both datasets compare to CART and C5.0. The other criteria of CHAID such as true positive, true negative, positive predictive, negative predictive value and AUC under ROC curve are well accepted. Age of voters is the main predictor to differentiate voters' turnout among young and older voters. In 2008 GE, voters below the age of 30 years old are less likely to vote if they are not interested to join or help during the election. The impact of rally towards young voters are high in GE 2013. On the other hand, factors that influence voters from the middle age group to vote are factors such as the need to assist or help the candidates and party, interest in politics and the ability to join politics. In this study, it is found that sociological related variables such as religion, gender and employment status do not influence voters' turnout in all the algorithms used.

Limitations and Future Research

There are few limitations in this study. This study uses only three theoretical models in examining the relationship between voters' turnout and political participation models. Future research should include other theoretical models such as the cultural modernisation and rational choice theory. In terms of methodological aspect, there are many imbalance cases in both datasets. Thus, future study should consider other approaches such as resampling methods. In addition, other decision tree algorithms such as Support Vector Machine (SVM), Random Forest and Boosting C5.0 should also be used predict voters' turnout.

Acknowledgment

The authors would like to acknowledge the work that led to this paper, which was fully funded by the Research Generation University Grant, Universiti Utara Malaysia.

References

1. Fowler, A., *Electoral and policy consequences of voter turnout: Evidence from compulsory voting in Australia.* Quarterly Journal of Political Science, 2013. **8**(2): p. 159-182.
2. Mohd Hed, N. and M.T. Grasso, *Age group differences in political activism in Malaysia.* Journal of Youth Studies, 2020. **23**(6): p. 765-779.
3. Blais, A., *To vote or not to vote?: The merits and limits of rational choice theory*. 2000: University of Pittsburgh Pre.
4. Tenn, S., *An alternative measure of relative education to explain voter turnout.* The Journal of Politics, 2005. **67**(1): p. 271-282.
5. Norris, P., *Electoral engineering: Voting rules and political behavior*. 2004: Cambridge university press.
6. Nedelmann, B., *Individuals and parties—Changes in processes of political mobilization.* European Sociological Review, 1987. **3**(3): p. 181-202.
7. Rosenstone, S.J. and J.M. Hansen, *Mobilization, participation, and democracy in America*. 1993: Longman Publishing Group.
8. Farrell, D.M., *Campaign strategies and tactics.* Comparing democracies: Elections and voting in global perspective, 1996: p. 160-183.
9. Welsh, B., *"Saviour" Politics and Malaysia's 2018 Electoral Democratic Breakthrough: Rethinking Explanatory Narratives and Implications.* Journal of Current Southeast Asian Affairs, 2018. **37**(3): p. 85-108.
10. Blais, A., *Turnout in elections*, in *The Oxford handbook of political behavior*. 2007.
11. Verba, S., K.L. Schlozman, and H.E. Brady, *Voice and equality: Civic voluntarism in American politics*. 1995: Harvard University Press.
12. Hirzalla, F., L. Van Zoonen, and J. de Ridder, *Internet use and political participation: Reflections on the mobilization/normalization controversy*. The Information Society, 2011. **27**(1): p. 1-15.
13. Rauf, A.A.A., N.A. Hamid, and M.S. Ishak, *Modelling the effect of access to information, political interest and policy satisfaction on youth online political participation in Malaysia.* Jurnal Komunikasi: Malaysian Journal of Communication, 2016. **32**(2).
14. Wang, S.-I., *Political use of the Internet, political attitudes and political participation.* Asian Journal of Communication, 2007. **17**(4): p. 381-395.
15. Breiman, L., et al., *Classification and regression trees*. 1984: CRC press.
16. Kass, G.V., *An exploratory technique for investigating large quantities of categorical data.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 1980. **29**(2): p. 119-127.
17. Quinlan, J., *C5. 0 machine learning algorithm*. 1993.
18. Kuhn, M. and K. Johnson, *Applied predictive modeling.* Vol. 26. 2013: Springer.

Table 1: Historical Statistics of General Elections (1955-2013)

Date of General Election	VAP[1] ('000)	VEP ('000)	Turnout ('000)	Increase of voters ('000)	Increase of voters (%)	Turnout (VAP) (%)	Turnout (VEP) (%)	Variance Turnout (%)
1955-07-27	-	1,280.9	1,060.6				82.8%	
1959-08-19	-	2,17.7	1,596.3	896.8	+70.0%	-	73.3%	-9.5%
1964-04-25	-	2,681.9	2,116.0	504.2	+23.2%	-	78.9%	+5.6%
1969-05-10	-	3,450.0	2,539.2	768.1	+28.6%	-	73.6%	-5.3%
1974-08-24	5,265.0	4,178.9	3,138.4	728.9	+21.1%	59.6%	75.1%	+1.5%
1978-07-08	6,067.2	5,059.7	3,810.0	880.8	+21.1%	62.8%	75.3%	+0.2%
1982-04-22	6,828.2	6,081.6	4,585.5	1,021.9	+20.2%	67.2%	75.4%	+0.1%
1986-08-02	7,893.9	6,791.4	5,052.8	709.8	+11.7%	64.0%	74.4%	-1.0%
1990-10-20	8,882.0	8,000.0	5,784.0	1,208.6	+17.8%	65.1%	72.3%	-2.1%
1995-04-24	10,175.0	9,012.4	6,155.5	1,012.4	+12.7%	60.5%	68.3%	-4.0%
1999-11-29	13,411.5	9,564.1	6,627.9	551.7	+6.1%	49.4%	69.3%	+1.0%
2004-03-21	13,802.5	9,756.1	7,209.8	192.0	+2.0%	52.2%	73.9%	+4.6%
2008-03-08	15,283.3	10,922.1	8,300.8	1,166.0	+12.0%	54.3%	76.0%	+2.1%
2013-05-05	17,883.7	13,268.0	11,256.6	2,345.9	+21.5%	62.9%	84.8%	+8.8%

Note:

VEP: adult citizens that are registered and eligible to vote; VAP: adult citizens that are within the prescribed age (21 years old and above) to vote regardless of being registered with the EC

1 http://www.idea.int/data-tools/question-countries-view/441/221/ctr

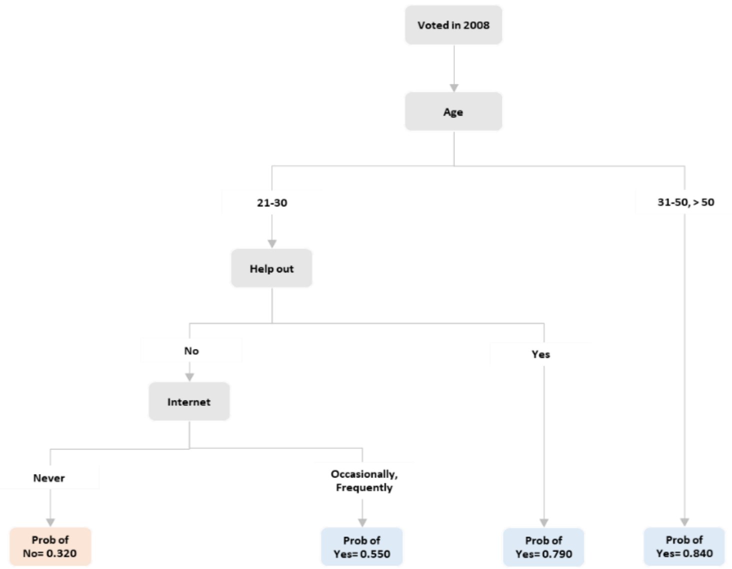

Figure 2(A). CART Rules Representation Wave 3 dataset

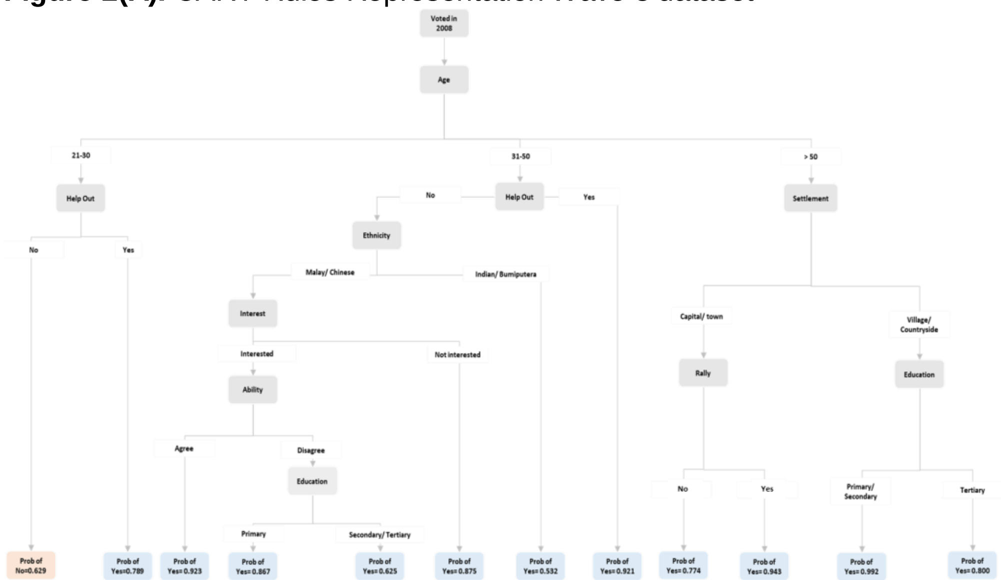

Figure 2(B). CART Rules Representation Wave 3 dataset

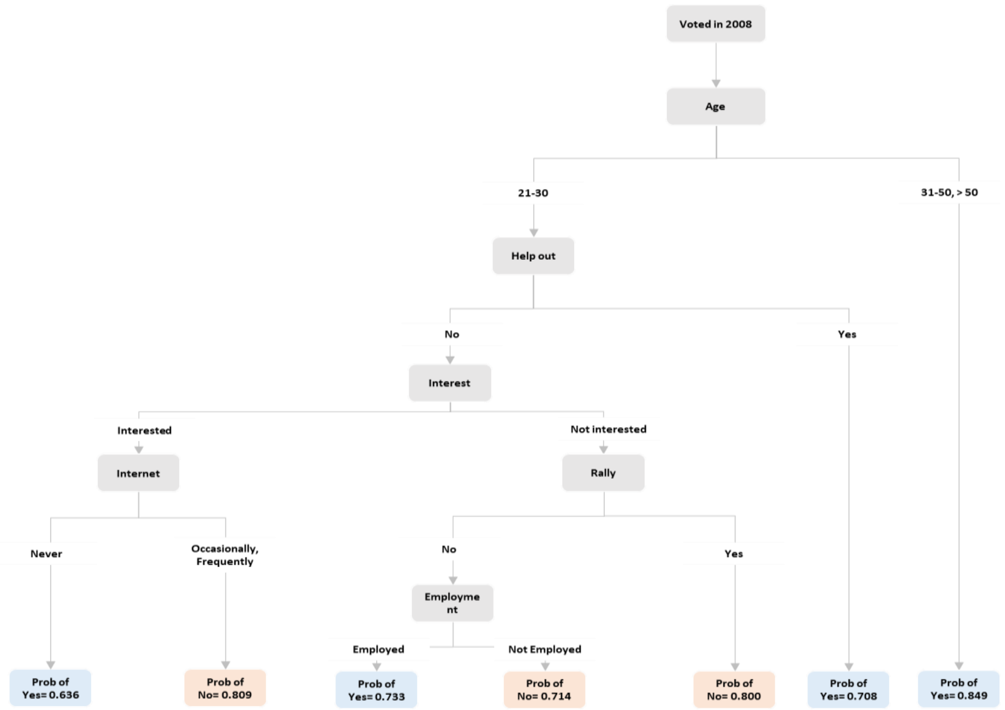

Figure 2(C). C5.0 rules representation Wave 3 dataset

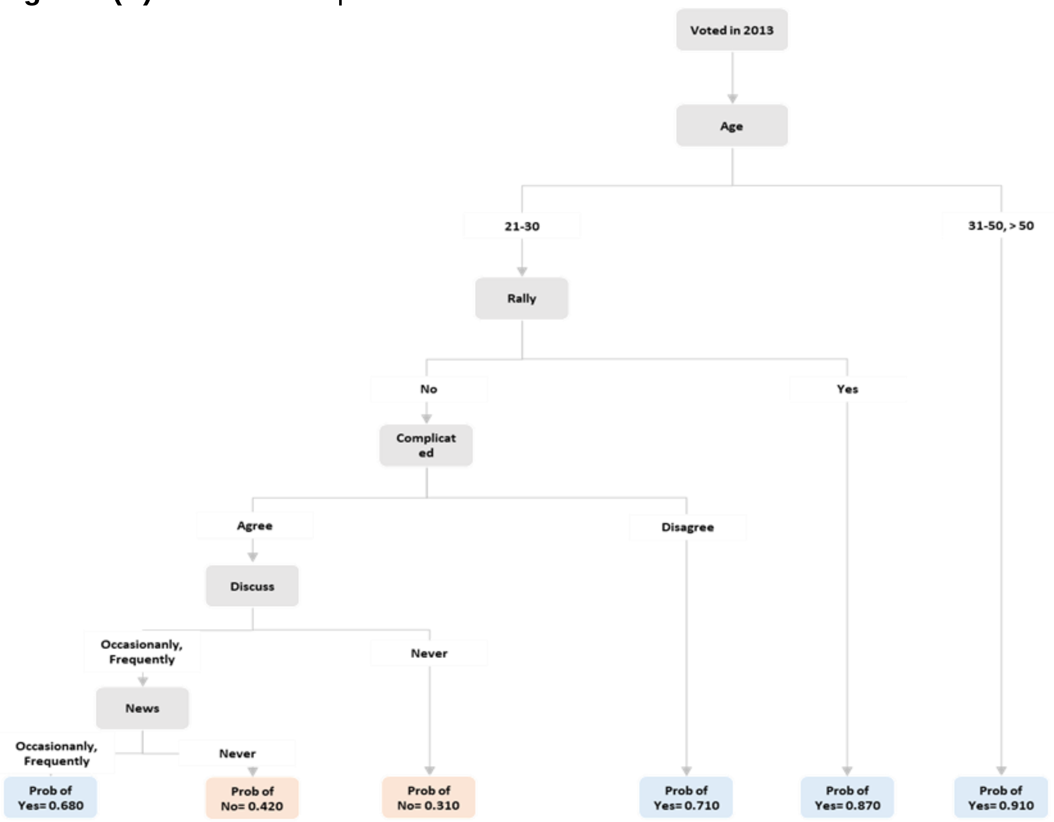

Figure 2(D). CART rules representation Wave 4 dataset

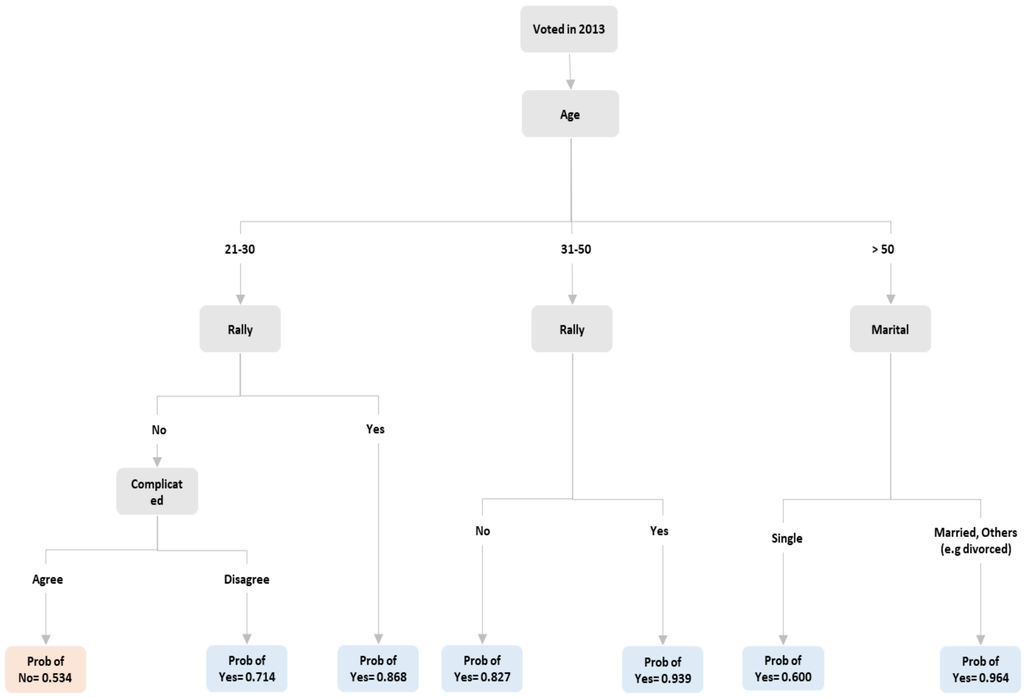

Figure 2(E). CHAID rules representation Wave 4 dataset

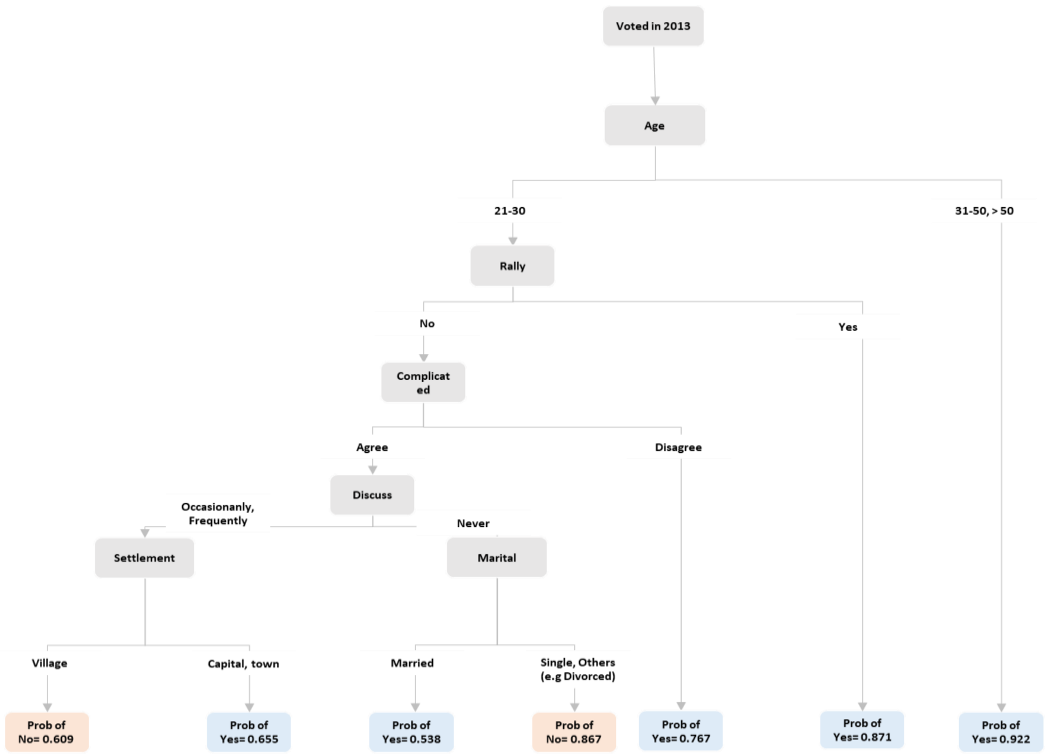

Figure 2(F). C5.0 rules representation Wave 4 dataset