ISSN 1022-8594
JUJS

Vol. 41, No.1, pp.11-30

An Alternative Approach for Characterizing Cigarettes Smoking among Bangladeshi Adults

Mohammad Alamgir Kabir* **and Md. Asraful Alam**
Department of Statistics, Jahangirnagar University

Abstract

Smoking cigarettes is a preventable public health problem. This study examines the cigarettes smoking patterns among Bangladeshi adults by CART an alternative approach for its suitability than others such as binary logistic regression, multinomial logistic regression, chi-squared automatic interaction detector, quick unbiased statistical test. A nationally representative dataset of 1037 respondents extracted from Global Adult Tobacco Survey, Bangladesh. CART was used to characterize the cigarette smoking patterns among adults aged 15 years and above. The algorithm builds a tree model to classify “average number of cigarettes smoked per day” using some attributes as predictors. CART was found easy to understand compared to other traditional techniques. Logistic regression model requires the parametric assumption (PA) of the dependent variable. However, this PA often restricts when data have a mixture of categorical and continuous variables. For 2CAT and 3CAT models the classification accuracy (%) of CART is the highest compared to other techniques such as CHAID, QUEST, BLR, and MLR. So, the alternative approach CART is the best in terms of all aspects for characterizing cigarette smoking among adults in Bangladesh.

Keywords: cigarettes smoking; classification and regression tree; Bangladesh; GATS.

Introduction

Tobacco consumption (TC) is a preventable public health problem. In 2011, tobacco related illnesses killed almost 6 million people, with nearly 80% of these deaths occurred in low- and middle-income countries. By the year 2030, 8 million people are projected to die annually due to tobacco related diseases [1, 2]. TC and associated consequences are decreasing rapidly or levelling off in developed countries and some middle-income countries, but the incidence is still high in developing countries including Bangladesh [3-5]. The prevalence of TC among aged 15 years and above in Bangladesh is still high with using diverse tobacco products. Males consumed more

* E-mail of correspondence: alamgir@juniv.edu, alamgirfa_juniv@yahoo.com

tobacco than females (48.5% vs. 25.4%). However, smoking tobacco products, namely, cigarettes and *bidis* are popular among males [1, 2]. Several demographic, socioeconomic, environmental, and programmatic factors were found to be associated with TC among adults in Bangladesh [6-10].

Many tobacco related studies have employed logistic regression in their analysis and they mostly analyzed categorical variables with dichotomous outcomes [6, 8-18]. Linear regression cannot deal with dependent variables that are categorical in nature and the alternatives are a number of regression techniques, including logistic regression [19, 20]. Frequently "logistic regression" refers to the technique for problems in which the dependent variable is dichotomous (the category of dependent variable is limited to two categories). Logistic regression can also be used in more than two category dependent variables and referred to as multinomial or ordinal logistic regressions [21-24].

In comparison to logistic regressions, classification and regression tree (CART), a data mining technique have not been widely applied for tobacco related research. There are a few studies for characterizing smoking patterns among adults on developed countries that employed this class of methods [25-27]. However, CART has enormous statistical properties over traditional methods, such as binary and multinomial logistic regressions and other data mining techniques, namely, chi-square automatic interaction detector (CHAID), and quick, unbiased, efficient statistical tree (QUEST) [19, 25, 28-31]. Some of the advantages of CART are: it is purely non-parametric and is independent of distribution assumptions; it can handle both continuous and categorical data; it can be applied to skewed or multi-modal data without requiring the independent variables to be normally distributed; it can handle missing data; it is relatively automatic ‘machine learning’; it requires less input for analysis; it possesses visualization characters and its results are simple to interpret even for non-statisticians.

Therefore, this study examines the smoking patterns among adults by CART method and to compare findings with other traditional techniques using nationally representative data extracted from Global Adult Tobacco Survey.

Materials and Methods

The data and sampling

The data, the detailed methodology of data collection, sampling procedure, questionnaires and relevant information were reported in Global Adult Tobacco Survey (GATS) Bangladesh report-2009 [32]. Briefly, based on the sampling frame from Bangladesh Bureau of Statistics (BBS), the implementing agency of Bangladesh population census in 2001, the GATS was a three-stage stratified cluster sample of households. In the first stage, 400 primary sampling units (PSUs) (200 from rural and another 200 from urban areas) were selected with probability proportional to size. In the second stage, a random selection of one secondary sampling unit (SSU) per selected PSU was done. The SSUs were based upon the enumeration areas (EAs) from Bangladesh Agricultural Census, 2008. Each EA's consisted of 200 households in rural areas and 300 households in urban areas. In the third stage, households were selected systematically within the listed households from a selected SSU (an average of 28 households to produce equal male and female households based on design specifications). One respondent was randomly selected for interview from each selected eligible household to participate in the survey. About 10,751 (96.0%) households and 9,629 (86.0%) individuals were successfully completed the interview. The sample design for Bangladesh provides cross-sectional estimates for the country as a whole as well as by urban, rural and gender.The current study utilized 1,037 cigarettes smokers separately for characterizing smoking patterns among adults.

The tools of data collection

GATS in Bangladesh used two types of questionnaires: the household questionnaire and the individual questionnaire. The questionnaires were based on GATS core and optional questions. The Ministry of Health & Family Welfare of Bangladesh with the consultation of local agencies, National Institute of Preventive and Social Medicine, National Institute of Population Research and Training, and Bangladesh Bureau of Statistics and international collaborators such as WHO South East Asia Regional Office and Centers for Disease Control and Prevention conducted the survey. The survey used electronic system that facilitates the complex skip pattern used in the GATS questionnaire, as well as some in-built validity checks on questions during the data collection. A repeated quality control mechanism was used to test the quality of questionnaire programming. The main steps involved in quality control checks were: version checking for household and individual questionnaires, checking date and time, skipping patterns and validation checks. The data were suitably weighted for well representation of the country.

Dependent and independent variables

The dependent variable for the CART model is “average number of cigarettes smoked per day” and a total of two models were analysed namely, 2 category model (2CAT) and 3 category model (3CAT).

In Bangladesh, the range of usage was 1 to 60 cigarettes per day [32]. Those who never smoked, or smoked but not every day in past 30 days before the survey were excluded because the main objective of this study is to understand the behaviour and characteristics of daily smokers using data mining technique. To compare the efficiency of CART models with other data mining algorithms, namely, CHAID, QUEST, Binary Logistic Regression (BLR), and Multinomial Logistic Regression (MLR), the dependent variable was also grouped into two-category (2CAT) and three-category (3CAT) models. The grouping for the 2-category models was guided by the median of the number of cigarettes smoked per day. The value is 7 cigarettes per day. For the 3-category models, the tertiles of the number of cigarettes smoked per day were used. The tertiles are 5 and 10 cigarettes per day. In addition, the cut-off points were also supported by the existing literature [25, 26] and the GATS data structure. Following the theory and literature on smoking behaviour, and taking into consideration the nature of data and the proposed CART technique, independent variables including gender, place of residence, highest level of education, wealth index, age when first started smoking cigarettes, smoking caused serious illnesses, advertisement of cigarettes at point of sale, and health warning labels in cigarettes packets were used to characterize the smoking behaviours among adults. The details of the variables and their coding for analysis are presented in **Table 1**.

Table 1: Variables from GATS and their Coding for Analysis

Variables Name	Specific question asked or how variable was derived	Coding for analysis
Average number of cigarettes smoke per day: **RESPONSE VARIABLE**		
Cigarettes; **B06a**	On average, how many of the cigarettes do you currently smoke each day? Options: “0” cigarette; a range of “1-60” cigarettes; smokes product but not everyday	

Variables Name	Specific question asked or how variable was derived	Coding for analysis
The set of independent variables selected for current study		
Gender, **A1**	Record gender from observation	1=male, 2=female
Residence	What is the place of residence	1=urban, 2=rural

Education (**HEDU**)
A4
	The highest level of education? 1=no formal education; 2=less than primary school completed; 3=primary school completed; 4=less than secondary school completed; 5=secondary school completed; 6=high school completed; 7=college or university completed; 8=post graduate degree completed; 77=don’t know; 99=refused	1=no formal education (option-1);
		2=less than or equal to primary (options 2 & 3);
		3=more than primary (options 4, 5, 6, 7& 8);

Wealth Index (**WI**) **A6**	The household or any person in the household has: a. electricity; b. flush toilet; c. fixed telephone; d. cell telephone; e. television; f. radio; g. refrigerator; h. car; i. moped/scooter/motorcycle; j. washing machine; k. bicycle; l. sewing machine; m. *almirah*/wardrobe; n. table; o. bed or cot; p. chair or bench; q. watch	1=poor (option 1 &2) 2=middle (option 3) 3=rich (option 4 & 5)

	or clock. Options based on factor analysis: 1=1st quintile; 2=2nd quintile; 3=3rd quintile; 4=4th quintile; 5=5th quintile	
Age when first started smoking cigarettes (**SAGE**) **B4**	How old where you when you first started smoking tobacco daily? (continuous form)	Dataset contains 12 to 36 years old adults
Smoking causes serious illness (**SSI**) **H1**	Based on what you know or believe, does smoking tobacco cause serious illness? Options. 1=yes; 2=no; 7=don't know; 9=refused	1=yes (option 1) 2=no (option 2 and 7)
Advertisement at point on sale (**APS**) **G4a1 & G4a2**	In the last 30 days, have you noticed any advertisements or signs promoting the cigarettes in stores where the products are sold? Options: 1=yes; 2=no; 7=don't know; 9=refused	1=yes (option 1) 2=no (option 2 and 7)
Health warning labels in cigarette packets (**HWP**) **G2 & GG2**	In the last 30 days, did you notice any health warnings on cigarette packages? Options. 1=yes; 2=no; 3=did not see any cigarette packages; 9=refused.	1=yes (option 1) 2=no (option 2)

Classification and Regression Tree (CART)

CART is a predictive model that classifies the data into leaf and node divisions viewed as a tree. Each branch of the tree represents a variable for classification and the leaves of the tree branch out according to some splitting algorithms. The CART produces rules that are mutually exclusive and collectively exhaustive and categorizes data on each branch point without losing any of the data. The total number of observations in a parent node is equal to the sum of the number of observations contained in its two children nodes.

As we mentioned earlier, CART has enormous statistical properties over other techniques. The detailed description about CART and its application were found elsewhere [19, 25, 28-31, 33-35]. Briefly, to develop a CART for classification, each predictor is chosen based on how well it fits the records with different predictions. The entropy metric is used to determine whether a split point for a given predictor is better than the others. The CART algorithm splits the independent variable into two separate hyper-rectangular areas according to performance measures. From the algorithmic point of view, CART has a forward stepwise technique that adds model terms and a backward technique for pruning, while selecting important variables that are useful in the model. The output of the models is a hierarchical structure that consists of a series of "if-then" rules to predict the outcome of the dependent variable.

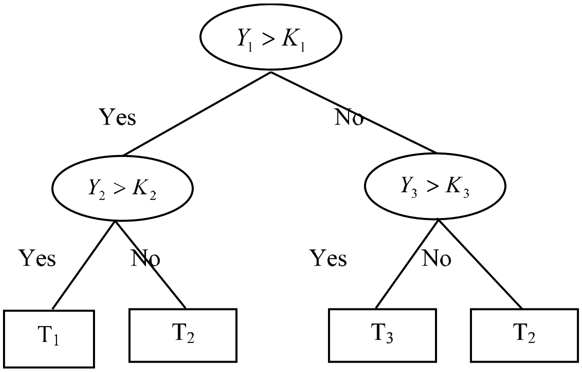

Figure 1: A Typical CART Model for Classification.

Notes: Ovals are the intermediate nodes and rectangles are terminal nodes, K_1, K_2 and K_3 are splitting values of the variables Y_1, Y_2 and Y_3, respectively.

For example, at each intermediate node (ovals in **Figure 1**) of the tree, a condition is tested on the variables (e.g., Y_1, Y_2 and Y_3). The split then takes place according to whether the condition is satisfied. The observations that satisfy the condition are grouped in the left branch while the remaining grouped in the right branch. Based on the splitting values

(K_1, K_2 and K_3) of the variables, every data point ends up in one of the nodes called terminal nodes (T_1, T_2 and T_3). The criteria for each terminal node by retreating up the tree to the top node can then be determined. For instance, the first terminal node (terminal node farthest to the left) retreats up to the left edge of the tree, yielding the following rule: "If $Y_1 > K_1$ and $Y_2 > K_2$, then the observation will be classified as T_1 (first terminal node)." Other terminal nodes in the tree can be interpreted similarly.

To build CART model, the number of cases in parent and child nodes are determined based on classification accuracy (overall & % in the specific classes) and other diagnostic results such as index chart, gain chart, and risk estimates. Smaller the cases in parent and child nodes, higher the classification accuracy of the model but it will enlarge the size of the tree, and make difficult for interpretations. Therefore, on an average the analysis procedure run about 25-30 times to determine the optimum cases in parent and child nodes and to satisfy other diagnostic tests. The predictive ability of the data mining model is also evaluated by their classification accuracy through a cross-validation technique. In the present study ten-fold cross validation was used to estimate true classification rate. To be specific, the algorithm divided the data set into 10 groups. In each of 10 iterations, nine groups were used for training (constructing) the model and the one remaining group was used for testing the constructed model. This process was repeated nine times more with the alteration of testing groups to obtain the classification rates of each testing set. The final classification results from 10 different testing groups were then averaged to obtain the cross-validation error rates of the decision tree model.

Results

Descriptive statistics

About 1,037 cigarette smokers were used to develop CART model. In 2CAT model, about 53% smoked 1-7 cigarettes per day and 47% smoked 8 or more. Whereas in 3CAT model, 40.8% smoked 1-5 cigarettes per day, 38.5 % smoked 6-10 cigarettes per day and rest 20.7 % smoked 11 or more cigarettes per day (**Table 2**).

Table 2: The Dependent Variables and their Grouping for CART Models

Cigarette model		
2CAT		
Category	n	%
1-7 Cigarettes per day	554	53.4
8+ Cigarettes per day	483	46.6
3CAT		
Category	n	%
1-5 Cigarettes per day	423	40.8
6-10 Cigarettes per day	399	38.5
11+ Cigarettes per day	215	20.7
Total cigarettes smokers: 1037		

The association measure

The Chi-square test for categorical and F-test for continuous independent variables were run separately. It was found that the categorical independent variables selected through the CART algorithm was significantly ($P<0.10$) associated with the response variable. While for the continuous independent variables' age when first started smoking was found to be significantly ($P<0.05$) associated with the response variable (the test results are not shown here).

Table 3: Importance of Independent Variables to Predict Dependent Variables

Independent variables	2CAT (%)	3CAT (%)
Age when first started smoking	100	100
Place of residence	62.2	34.0
Gender	5.9	4.8
Smoking causes serious illness	14.4	7.1
Cigarettes ads. at point on sale	75.9	37.5
Health warning in cigs packs	40.9	4.9
Education	27.8	48.7
Wealth index	39.2	21.6

(%) indicates normalized importance of splitters

Variables importance score

The importance of variables was evaluated from the sum of the improvements in all nodes or percentage scores in which the variable appears as a splitter. Surrogates were also included in the calculation which means that a variable that never splits a node may still be assigned a large importance score. From a range of variables in the dataset, the CART software provides the “variable importance scores.” Variables that receives a 100% score (highest sum of improvements) indicates the most influential independent variable for predicting the dependent variable, followed by other variables based on their relative importance to the most important one. Any variables that do not make a significant contribution to the final model were excluded. The variable importance scores are reported in **Table 3** for all the 2 models. Age of smoking initiation appeared to be the most important variable in all the four models. Gender was the second most important variable in two models.

Model Summary

The model summary table provides some broad information on the models (**Table 4**). The specification provides information on whether CART is the chosen algorithm, dependent and independent variables, validations, maximum tree depth, minimum number of cases in parent nodes and minimum number of cases in child nodes. The results of the analysis display information on all the parent, child and terminal nodes, while showing the number of observations in each category of the dependent variable for every node, depth of the tree (number of levels below the root note), and independent variables included in the final model.

Table 4: Model Summary of CART

Specification and Result	2CAT	3CAT
Minimum Cases in Parent Node*	60	80
Minimum Cases in Child Node*	20	40
Total Number of Nodes	21	19
Number of Terminal Nodes	11	10
The tree depth is 5		

CART results and interpretation

In the 2CAT cigarette model, the total number of nodes is 21 of which 11 are terminal nodes (the nodes that did not split to further nodes). The overall classification accuracy is 62.1%, demonstrating that the constructed decision tree model correctly classifies more than 62% of the cases. In the root node (node 0), about 53% used 1-7 cigarettes daily and 47% used more than 8 cigarettes per day. This node was divided based on the best splitter. Among the independent variables, age when first started smoking is the most influential variable (best splitter) with normalized importance of 100%. The smokers who started smoking at age less than or equal to 14 years branched the left child node (node 1) and all the other cases branched to the right (**Figure 2**). Node 1 is the parent node and is partitioned by wealth index (with normalized importance of 39.2%) to form two child nodes (node 3 and node 4). There is no further division in node 3 and node 4. These two nodes are terminal nodes and the following rule is created: if the respondents started smoking at age less than or equal to 14 years old and were from a rich family, about 48% smoked 8 or more cigarettes per day. If they were from a poor or middle income family, about 78% smoked 8 or more cigarettes per day. Similarly, node 2 is the parent node and again divided by age when first started smoking with an improvement of 0.005, and formed two child nodes (node 5 and node 6). The respondents who started smoking at age less than or equal to 26 years were categorized in the left child node (node 5) and all other instances were in the right node (node 6). Since there is no further division in node 6, it is a terminal node. The rule is that if the respondents started smoking at age more than 26 years old, 74.1% used 1-7 cigarettes per day. For the cases that started smoking at age more than 14 years old but less than or equal to 26 years, the next influential variables (splitters) are education level with improvement of 0.004, cigarette advertisements at point of sale (with improvement of 0.003 and 0.007), wealth index (improvement of 0.002), place of residence (improvement of 0.003 and 0.002), and health warning on cigarette packets (improvement of 0.005). It is interesting to discover that among the respondents who started smoking between the age of 14 to 26 years, had no formal education, seen cigarette advertisements at point of sale and resided in the urban area, 56.2% smoked 8 or more cigarettes per day (node 17). However, 61.9%

smoked 1-7 cigarettes per day if they resided in the rural area (node 18). Smoking is thus more prevalent in urban areas. This pattern is also true among those who started smoking at the same age group, seen cigarette advertisements at point of sale, but with education (node 15 and 16). However, the proportion of heavy smokers was lower when compared to the group without formal education.

In the 3CAT cigarette model (**Figure 3**), the total number of nodes is 19 of which 10 are terminal nodes. The overall classification accuracy is only 46.6%. In the root node (node 0), about 40.8% smoked 1-5 cigarettes per day, 38.5% smoked 6-10 cigarettes per day and the remaining 20.7% smoked more than 11 cigarettes per day. Like the 2CAT model, age when first started smoking is the most important variable with normalized importance of 100%. The first splitter is place of residence with improvement of 0.007 which partitioned the root node into two child nodes (node 1 and node 2). Node 1 is now the parent node and is divided by the splitter of cigarette advertisements at point of sale (improvement of 0.005) to form two child nodes (node 3 and node 4). Node 4 is the terminal node. If the respondents were from urban area and had not seen cigarette advertisements at point of sale, 37.9% of them smoked 11 or more cigarettes per day. If the respondents had seen cigarette advertisements at point of sale (node 3), the next splitter is the highest education level (improvement of 0.005) which resulted in two nodes (node 7 and node 8). Since no further splits occurred in node 7 and 8, they are the terminal nodes.

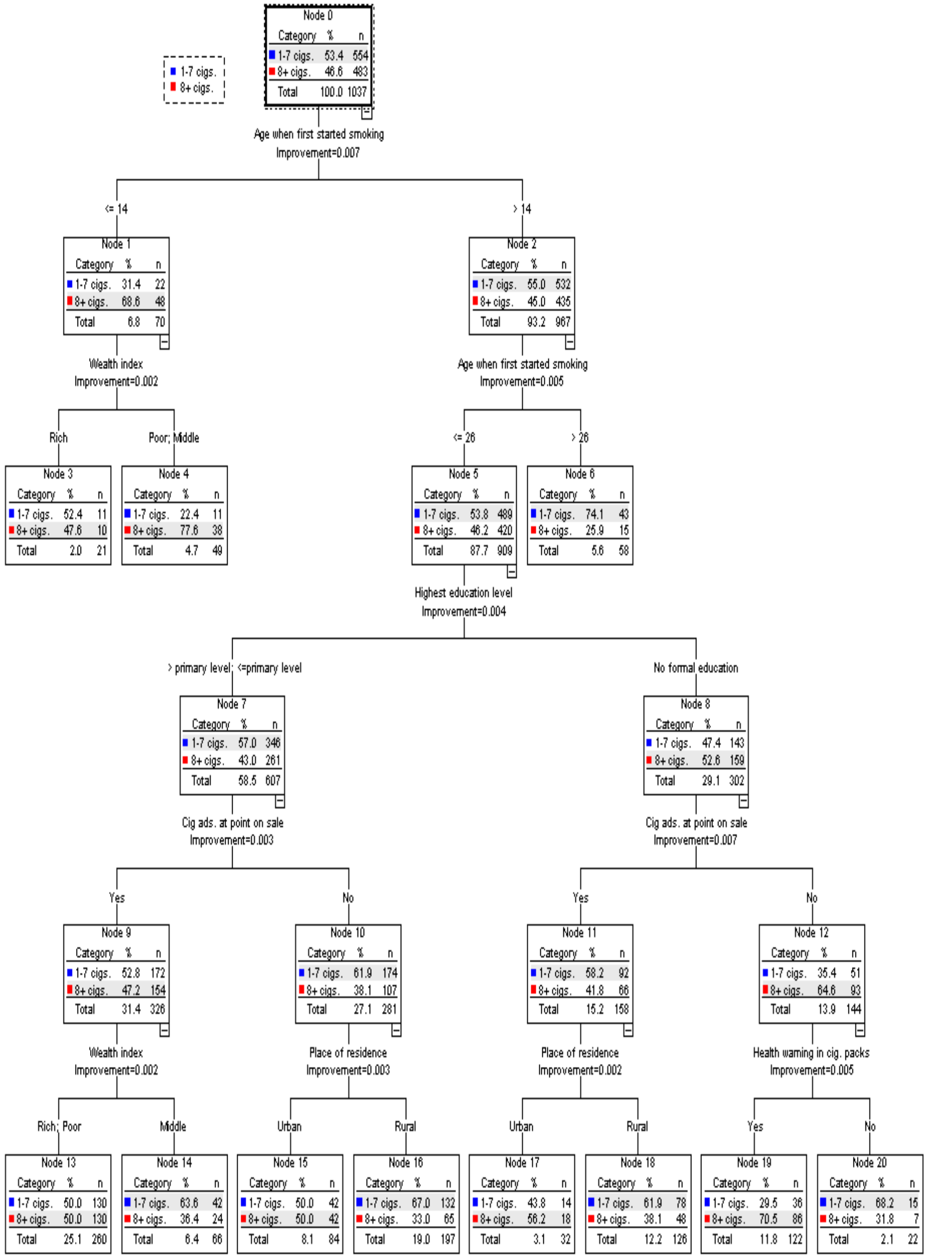

Figure 2: CART Model for Classifying Average Number of Cigarettes Smoked (2CAT) per Day

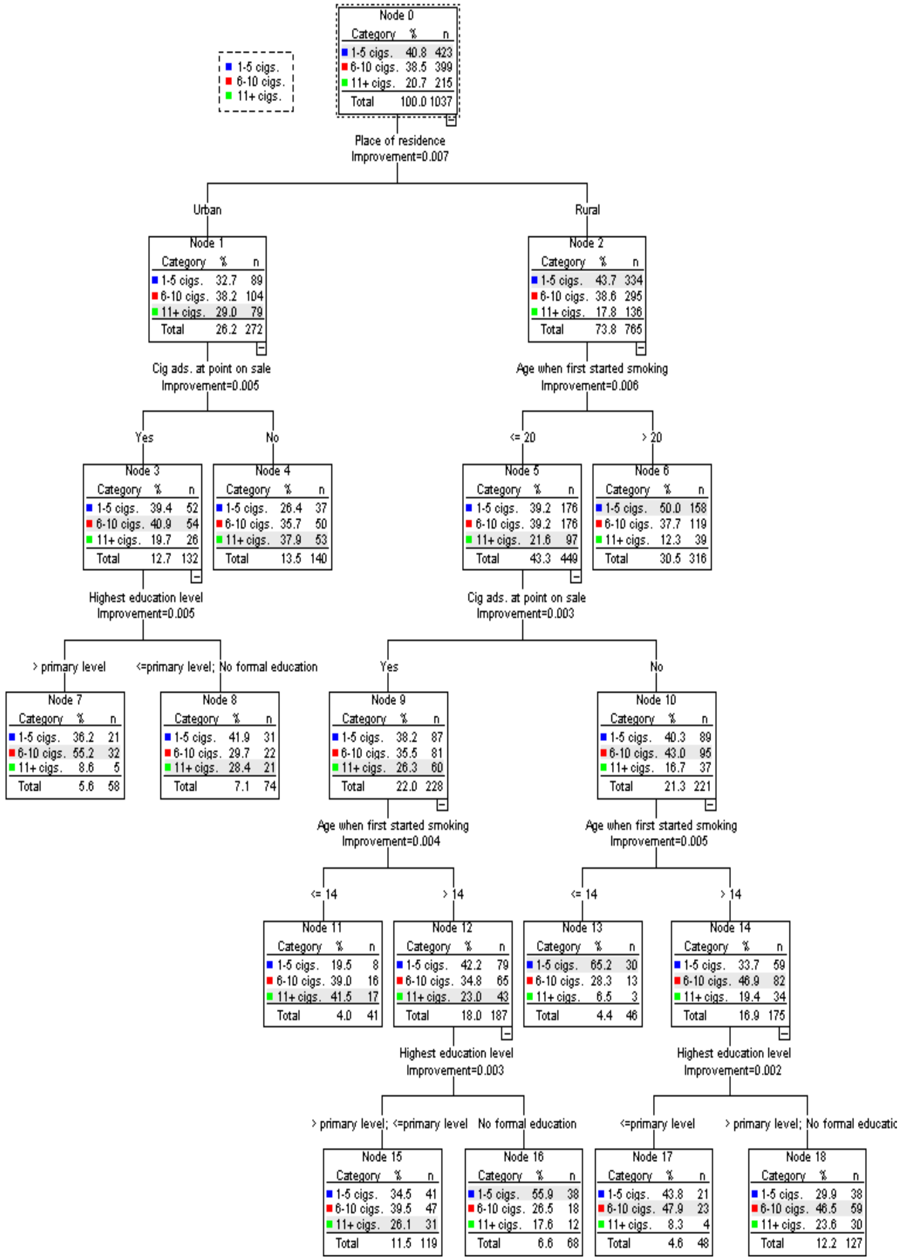

Figure 3: CART Model for Classifying Average Number of Cigarettes Smoke (3CAT) per Day

The following rule is developed: if the respondents were from urban area, seen cigarette advertisements at point of sale, and they have more than primary education, 55.2% of them smoked 6-10 cigarettes per day whereas, 28.4% smoked 11 or more cigarettes per day if they had less than or up to primary education. This shows that those with less education are more likely to be heavy smokers.

Discussions and conclusions

CART was used to characterize the cigarette smoking behaviour among adults aged 15 years and above in Bangladesh. The algorithm builds a tree model to classify “average number of cigarettes smoked per day” using some attributes as predictors. CART was found easy to understand compared to other data mining techniques [25, 28-31, 33-35], CART is appropriate because it defines groups that are consistent in their attributes but which vary in terms of the dependent variable and the results are presented graphically. In practice, many variables are not normally distributed and different groups may have markedly different degrees of variation or variance. Complex interactions or patterns may exist in the data and make modelling difficult when the number of interactions and variables become substantially large [31]. It was found that CART can solve such problem which may be difficult or impossible to solve using traditional multivariate techniques. CART is useful for handling highly skewed or multi-modal numerical data, as well as categorical predictors with either ordinal or non-ordinal structure. CART can competently handle data with a combination of categorical and continuous variables. In contrast with logistic regression [25, 28], CART is inherently non-parametric and no assumptions are made regarding the underlying distribution of the predictor variables. The CART algorithm can efficiently handle missing data through surrogates. In contrast with other multivariate techniques [25, 28], the method is relatively automatic ‘machine learning’ and less input is needed for analysis. CART shows visualization character, is descriptive in nature and simple for non-statisticians to interpret. Several possible predictors were inserted into the CART system to build the model. The procedure automatically excluded the variables that did not make any significant contribution to the final model for classifying the response variable. This study provided CART application and developed a comparison between CART and other traditional techniques especially logistic regressions.

The classification accuracy (goodness-of-fit) of 10-fold cross validation among CART and other decision trees algorithms such as CHAID and QUEST were tested. It should be mention that the same training and testing data for calculating cross-validated errors were used. The result showed that the CART models (2CAT & 3CAT) of cigarette yielded higher classification accuracy than CHAID and QUEST models (**Table 5**). In addition, the classification accuracy of 10-fold cross validation of 2CAT CART model vs. Binary Logistic Regression (BLR) and 3CAT, CART model vs. Multinomial Logistic Regression (MLR) for average number of cigarettes smoked per day were tested separately. It is apparent that, 2CAT CART model yielded higher classification accuracy (**Table 5**) than BLR model (62.1 vs. 56.5). Though the classification accuracy is low due to more categories, but it also reveals that, 3CAT CART model yielded higher classification accuracy than MLR model (46.6 vs. 43.1).

Table 5: Classification Accuracy (%) of the Two Models to Predict Dependent Variables

Models[#]	2CAT	3CAT
CART	**62.1** (.422)	**46.6** (.614)
CHAID	58.8 (.457)	44.4 (.592)
QUEST	54.4 (.462)	37.1 (.640)
BLR	56.5	-
MLR	-	43.1

[#] CART-Classification and Regression Tree, CHAID-*Chi*-squared Automatic Interaction Detector, QUEST-Quick, Unbiased, Efficient Statistical Tree, BLR-Binary Logistic Regression, MLR-Multinomial Logistic Regression, the risk estimate of cross-validation are in parenthesis

In conclusion, it was clear that the logistic regressions (whether binary or multinomial) were less superior in terms of classification accuracy in the current data set, where there is a mixture of categorical and continuous independent variables. Both CART and logistic regression models can efficiently handle the nonlinear relationship between the independent and dependent variables. The main difference between the two techniques is that the logistic regression is a parametric approach that assumes the response variable follows the binomial or multinomial distribution, while CART does not require any distributional assumptions. The choice between the two techniques may depend upon the nature of data and no consensus exists about which of them best satisfies all conditions.

Limitations

Though the *bidi* smoking and smokeless tobacco products were also popular in Bangladesh among adults, but in CART analysis, these products were excluded due to data limitation.

References

[1] Eriksen, M., Mackay, J., & Ross, H., *The Tobacco Atlas*, 2012, Fourth Edition, American Cancer Society. ISBN-10:1-60443-093-1; ISBN-13:978-1-60443-093-6. Available at www.TobaccoAtlas.org

[2] World Health Organization, *WHO report on the global tobacco epidemic, 2009: implementing smoke-free environments*, 2009, Geneva, Switzerland: World Health Organization. Available at http://www.who.int/tobacco/mpower/en.

[3] Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S., Murray, C. J. L., *the Comparative Risk Assessment Collaborating Group*, et al., Selected major risk factors and global and regional burden of disease, 2002, *The Lancet*, 360, 1347–1360. doi:10.1016/S0140-6736(02)11403-6

[4] Mathers, C. D., & Loncar, D., *Projections of global mortality and burden of disease from 2002 to 2030*, PLoS Medicine, 2006, *3*, e442. doi:10.1371/journal.pmed.0030442

[5] Samet, J. M., & Wipfli, H. L., *Globe still in grip of addiction*, Nature, 2010, *463*, 1020-1021.

[6] Abdullah, A. S., Hitchman, S. C., Driezen, P., Nargis, N., Quah, A. C. K., & Fong, G. T., *Socioeconomic differences in exposure to tobacco smoke pollution (TSP) in Bangladeshi households with children: Findings from the International Tobacco Control (ITC) Bangladesh survey*, International Journal of Environmental Research and Public Health, 2011, 8(3), 842-860.

[7] Choudhury, K., Hanifi, S. M., Mahmood, S. S., & Bhuiya, A., *Socio-demographic characteristics of tobacco consumers in a rural area of Bangladesh*, Journal of Health Population and Nutrition, 2007, 25, 456-464.

[8] Kabir, M. A., Goh, K. L., & Khan, M. M. H., *Tobacco consumption and illegal drug use among Bangladeshi males: Association and determinants*, American Journal of Men's Health, 2012, 7, 128-137. doi: 10.1177/1557988312462737

[9] Khan, M. M. H., Khandoker, A., Kabir, M. A., Kabir, M., & Mori, M., *Tobacco consumption and its association with illicit drug use among men in Bangladesh*, Addiction, 2006, 101, 1178-1186. doi: 10.1111/j.1360-0443.2006.01514.x

[10] Palipudi, K. M., Gupta, P. C., Sinha, D. N., Andes, L. J., Asma, S., McAfee, T., et al., *Social determinants of health and tobacco use in thirteen low and middle income countries: Evidence from Global Adult Tobacco Survey*, PLoS One, 2012, 7(3), e33466. doi: 10.1371/journal.pone.0033466.

[11] Chen, Y. H., Yeh, C. Y., Chen, R. Y., Chien, L. C., Yu, P. T., Chao, K. Y., Han, B. C., *Moving toward people's needs for smoke-free restaurants: Before and after a national promotion program in Taiwan*, 2003–2005, Nicotine & Tobacco Research, 2009, 11(5), 503-513.

[12] Edens, E. L., Glowinski, A. L., Pergadia, M. L., Lessov-Schlaggar, C. N., & Bucholz, K. K., *Nicotine addiction in light smoking African American mothers*, Journal of Addiction Medicine, 2010, 4(1), 55-60.

[13] Hosseinpoor, A. R., Parker, L. A., Tursan d'Espaignet, E., & Chatterji, S., *Social determinants of smoking in low- and middle-income countries: Results from the World Health Survey*, PLoS One, 2011, 6(5), e20331. doi: 10.1371/ journal.pone.0020331.

[14] Padrão, P., Silva-Matos, C., Damasceno, A., & Lunet, N., *Association between tobacco consumption and alcohol, vegetable and fruit intake across urban and rural areas in Mozambique*, The Journal of Epidemiology and Community Healt*h*, 2011, 65, 445-453.

[15] Rachiotis, G., Siziya, S., Muula, A. S., Rudatsikira, E., Papastergiou, P., & Hadjichristodoulou, C., *Determinants of exposure to environmental tobacco smoke (ETS) among non-smoking adolescents (aged 11-17 years old) in Greece: Results from the 2004-2005 GYTS study*, International Journal of Environmental Research and Public Health, 2010, 7, 284-290.

[16] Rahman, M. M., Ahmad, S. A., Karim, M. J., & Chia, H. A., *Determinants of smoking behaviour among secondary school students in Bangladesh*, Journal of Community Health, 2011, 36, 831-838.

[17] Rudatsikira, E., Muula, S. A., Siziya, S., & Mataya, R. H., *Correlates of cigarette smoking among school-going adolescents in Thailand: Findings from the Thai Global Youth Tobacco Survey 2005*, International Archives of Medicine, 2008, 1, 8. doi: 10.1186/1755-7682-1-8

[18] Tarafdar, M. M. A., Nahar, S., Rahman, M. M., Hussain, S. M. A., & Zaki, M., *Prevalence and determinants of smoking among the college students in selected district of Bangladesh*, Bangladesh Medical Journal, 2009, 38(1), 7–12.

[19] Agresti, A., *An introduction to categorical data analysis*, 2007, Hoboken, NJ: John Wiley & Sons.

[20] Harrell, Jr. F. E., *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*, 2001, Springer-Verlag, New York.

[21] Bender, R., & Grouven, U., *Ordinal logistic regression in medical research*, Journal of Royal College of Physicians, 1997, 31(5), 546-551.

[22] Chan, Y. H., *Biostatistics 202: Logistic regression analysis*, Singapore Medical Journal, 2004, 45(4), 149-153.

[23] Chan, Y. H., *Biostatistics 305: Multinomial logistic regression*, Singapore Medical Journal, 2005, 46(6), 259-269.

[24] Kwak, C., & Clayton-Matthews, A., *Multinomial logistic regression*, Nursing Research, 2002, 51(6), 404-410.

[25] Moon, S. S., Kang, Y. K., Jitpitaklert, W., & Kim, S. B., *Decision tree models for characterizing smoking patterns of older adults*, Expert Systems with Applications, 2012, 39, 445–451.

[26] Schane, R. E., Ling, P. M., & Glantz, S.A., *Health effects of light and intermittent smoking: A review*, Circulation, 2010, 121(13), 1518-1522.

[27] Gervilla, E., Cajal, B., & Palmer, A., *Some relevant factors in the consumption and non-consumption of nicotine in adolescence*, Revista Iberoamericana de Psicologiay Salud, 2011, 2(1), 57-74.

[28] Giskes, K., Kunst, A. E., Benach, J., Borrell, C., Costa, G., Dahl, E., et al., *Trends in smoking behavior between 1985 and 2000 in nine European countries by education*, Journal of Epidemiology & Community Health, 2005, 59, 395–401.

[29] Ruben, D., & Canlas, Jr., *Data mining in healthcare: Current applications and issues*, 2009, Carnegie Mellon University, Australia.

[30] Soni, J., Ansari, U., Sharma, D., & Soni, S., *Predictive data mining for medical diagnosis: An overview of heart disease prediction*, International Journal of Computer Applications, 2011, 17 (8), 43-48.

[31] Srinivas, K., Rani, B. K., & Govrdhan, A., *Applications of data mining techniques in healthcare and prediction of heart attacks*, International Journal on Computer Science and Engineering, 2010, 2(2), 250-255.

[32] World Health Organization, *Global Adult Tobacco Survey (GATS) Bangladesh report-2009*, 2010, Dhaka, country office for Bangladesh: World Health Organization. Available at: http://www.searo.who.int/LinkFiles/

[33] Romei, A., & Turini, F., *Inductive database languages: Requirements and examples*, Knowledge and Information System, 2011, 26, 351–384. Doi: 10.1007/s10115-009-0281-4.

[34] Sarker, K. U., Moon, N. N., & Ahmed, S., *Classifying the practitioner's behavior in medical informatics by using data mining*, International Journal of Computer and Information Technology, 2011, 1 (2), 23-25.

[35] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J., & Hua, L., *Data mining in healthcare and biomedicine: A survey of the literature*, Journal of Medical System, 2012, 36, 2431–2448. Doi: 10.1007/s10916-011-9710-5.