Malaysian Journal of Mathematical Sciences 12(3): 401–419 (2018)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Application of Classification and Regression Trees Algorithm to Classify Children Ever Born: BDHS 2011

Saadati, M.[1], Bagheri, A. *[2], and Rana, S.[3]

[1,2]*National Population Studies & Comprehensive Management Institute, Shahid Beheshti Street, Tehran, Iran*
[3]*Department of Mathematical and Physical Sciences, East West University, Aftabnagar, Bangladesh*

E-mail: abagheri_000@yahoo.com
* *Corresponding author*

Received: 4 July 2017
Accepted: 30 August 2018

ABSTRACT

Study the nature of fertility determinants as influential development indicators is a necessity in densely populated countries like as Bangladesh. In addition, investigating these factors without considering the convenient statistical methods may result in misleading conclusions. The main purpose of this article is to classify one of the most important principal of fertility, children ever born, by applying Classification and Regression Trees (CART) algorithm. To achieve this goal, children ever born of ever-married women age 12-49 years old from Bangladesh Demographic and Health Survey 2011 data has been classified by Classification and Regression Trees algorithm according to a number of candidate demographic and socio-economic predictors. Marriage duration, couple's educational level, division, and religion were determined to be the most influential predictors by extracted classification model. The efficiency of CART algorithm has been proved by accuracy of the model.

Keywords: Fertility, Children Ever Born, Decision Trees, Regression Trees, Bangladesh Demographic and Health Survey.

1. Introduction

A large majority of women in Bangladesh before they reach at the age of twenty bear children. As a result, population grows rapidly. Therefore, reduction of the population growth and identifying the important factors which significantly associated with fertility is needed (Asaduzzaman et al., 2009).

Some surveys such as Bangladesh Demographic and Health Survey (BDHS) have been doing since 1993 to gather data for evaluating the implemented family planning programs (GOB, 1994). To manage increasing of the population, the variables that control growth rate should be studied. So it is crucial to model Children Ever Born (CEB) which represents potentiality of population growth. Traditional statistical methods such as Ordinary Least Squares (OLS) are not manageable to use, or inconvenience in addressing count variables. In this situation, the OLS estimations are not efficient, and standard errors estimates are not consistent, because some of basic assumptions such as normality are not met. There are also some possibilities that the OLS estimations could cause some count predictions to be negative. These problems have motivated researchers to search for alternatives methods.

A subset of discrete response regression models are count response models which are nonnegative integer responses with right skewed of the distribution. In recent years, Poisson Regression (PR) models or a Negative Binomial Regression (NBR) model have been applied to model count response variable considering some covariates (Bagheri 2018; Bagheri et al. 2017; Saadati 2015; Hilbe 2011). King (1989) and Winkelmann and Zimmermann (1994) proposed some generalized event count models based on the PR, NBR, and the binomial distributions as Generalized Poisson regression (GPR) model. A number of works have suggested various models to deal with extra Poisson variation in data (Femoye 1993).

Ordered categorical data (ordinal data) are ordinary data in scientific fields. To model these data, Agresti (2002) introduced Cumulative link models as the most popular class of ordinal regression models. The regression framework in these models has been provided similar to linear models while the response is considered as a categorical variable. Further, the proportional odds model is introduced by McCullagh (1980) which is a cumulative link model with a logit link. Proportional-odds Cumulative Logit (CL) model, the most popular model

for ordinal data, considers a threshold and uses cumulative probabilities up to it. Thus, the whole range of ordinal categories will be binary at that threshold.

The method of handling missing values in most of the mentioned methods is to eliminate cases having missing values; this is not efficient and increases the possibility of introducing bias in the study (Song and Lu 2015).

Classification in machine learning and statistics is the problem of recognizing a new observation belongs to which set of categories, according to a training set of data (data which contain observations whose category membership is known). In these models, classified data are used for building a model that can be applied to deduce the class of unclassified data.

There are so many popular and efficient classification methods such as clustering, k-means, discriminate analysis, and decision trees that are first introduced in 1960's. Decision trees are one of the most applicable data mining methods that are used to extract valuable information from large datasets and to present it in easy to interpret manner. Decision trees have been widely used in several disciplines due to this fact that both qualitative and quantitative variables can be considered as target or independent variables. By applying the tree model derived from historical data, predicting the result for future records is easy (Hastie et al. 2009).

Similar to stepwise variable selection in regression analysis, for selecting the most relevant input variables, decision tree can be applied. After identifying relevant variables, variables which play major roles should be known. Furthermore, variable importance can be computed according to the reduction of model accuracy after removing it. In most conditions, if a variable have effect on more records, the importance of that variable will be greater.

Moreover, several advantages of decision trees exist such as, fast modeling process, using both numerical and categorical data, simplifying input and target variables by making major subcategories, splitting variables by extensively searching of all possibilities, interpreting resulted tree easily, using nonparametric approach, handling missing values and heavy skewed data, and being robust against outliers. Decision tree method has also some disadvantages. When particularly using a small data set, it can be subject to overfitting and under fitting which can limit the generalizability and robustness of the resulting models. Moreover, powerful correlation between different potential input variables may result in selecting variables that improve the model statistics but are not causally related to the interested outcome.

In many studies on fertility, the number of women's CEB is modeled considering socio-economic variables. The commonly used models are the standard PR or CL model. PR model is suitable when the number of CEB is non-negative count without extra value and varies in a long range. But in application, both overdispersion and under-dispersion can exist in the CEB values. As a result, no longer inference by PR model based on the estimated standard errors is valid. As noted in Winkelmann and Zimmermann (1994), the number of CEB often does not follow equal-dispersion assumption when its mode is 2. Therefore, the standard PR model which assumes equal-dispersion is not appropriate to model data about household fertility decision. CL model is applied when CEB is categorized as an ordinal variable with small range. When the number of categories is large, gaining accurate model depends on adding interaction terms in the model. In this way, the accuracy of the model will be increased but the interpretation also will become complicated. In this situation, applying a model that could overcome these problems is needed (Saadati and Bagheri, 2015; Bagheri and Saadati, 2014).

The main aim of this article is to classify CEB of ever-married women age 12-49 from BDHS 2011 by applying one of the most applicable classification algorithms, Classification & Regression Trees (CART). To do this, the article arranged in four sections. Section 2 describes methods, which introduces decision trees and specifically CART algorithm. Section 3 outlines the results and discussions of applying CART on BDHS 2011 data. Some remarkable conclusions are described in Section 4.

2. Methods

Song and Lu (2015) declared that a classification problem contains four major components including a response or outcome variable, independent variables, learning dataset, and test or future dataset. Decision trees include another two components of a prior probability for each response and a decision loss or cost matrix.

Decision tree method is a great statistical tool for classification, prediction, interpretation, and data management that has quite a lot of possible applications in different field of studies. It has three types of nodes as a root node, internal nodes, and leaf nodes.

Branches symbolize chance outcomes that derive from root and internal nodes. Classification trees are a type of decision tress which split the dataset into categories belonging to the ordinal (or categorical) response variable. The

idea in a standard classification tree is to divide the dataset according to the data homogeneity. To quantify the homogeneity in Classification trees accurate measures of impurity such as entropy or Gini index according to computing proportion of the data that belong to a class are used.

THAID (Automatic Interaction Detection), Chi-Squared Automatic Interaction Detection (CHAID), Iterative Dichotomiser 3 (ID.3), QUEST (Quick, Unbiased, Efficient, Statistical Tee), and CART (Breiman et al., 1984) are the most applicable classification tree algorithms (Timofeev, 2004). In the CART algorithm Gini index as a generalization of the binomial variance is used for data classification.

2.1 CART Analysis

CART is a form of binary recursive partitioning which means in a decision tree each group of cases indicated by a node can only be categorized into two groups. Thus, each original node called as a parent node can be split into two child nodes. Moreover, the term recursive represents that the binary partitioning process can be applied repeatedly. CART Algorithm consists of four following basic steps:

Tree Building: CART begins at the root node including the learning dataset. Then, it finds the best possible variable to divide the node into two child nodes. All possible splitting variables called splitters and all possible values of the variable for splitting are checked to find the best variable.

The algorithm searches to maximize the average purity of the two child nodes to choose the best splitter. The two most common splitting functions are the Gini and Twoing (Timofeev, 2004). Gini impurity and Gini gain are used to select splitting points, attribute variables, and values of chosen variables that are given by:

$$i(t) = 1 - \sum_{i=1}^{m} f(t,i)^2 = \sum_{i \neq j} f(t,i) f(t,j) \tag{1}$$

$$\Delta i(s,t) = i(t) - P_L.i(t_L) - P_R.i(t_R) \tag{2}$$

where $f(t,i)$ is the probability of getting i in node t, and the target variable takes values in $1, 2, 3, ..., m$. P_L and P_R are the proportion of cases in node

t divided to the left and right child node, respectively. The splitting process stops when there aren't any Gini gain or the preset stopping rule are satisfied.

The primary splitter for each node is the variable that splits the node perfectly and maximizes the purity of the resulting child nodes. A considerable advantage of this methodology comparing to multivariate regression modeling is in facing missing values in any of the predictor variables. If the primary splitting variable is missing for an individual observation, instead of discarding it a substitute splitting variable is required. A substitute splitter is a variable whose pattern is similar to the primary splitter. Thus, the best available information in the face of missing values and all observations with reasonable quality are used.

Stopping Tree Building: In building a decision tree, to avoid the model to become excessively complex, stopping rules must be applied. The process is stopped if there is only one observation in each of the child nodes; the distribution of all observations within each child node is identical to predictor variables which causes splitting impossible; or the researcher set depth option which is an external limit on the number of levels in the maximal tree. Stopping parameters must be chosen according to the goal of the analysis and the dataset characteristics. To avoid overfitting and underfitting , the target proportion of records in a leaf node could be selected to be between 0.25 and 1.00% of the full training data as a rule-of-thumb (Berry and Linoff,1999).

Tree Pruning: To build a decision tree model in a different way, at first it could be grown a large tree, and then by removing nodes that provide less additional information, prune it to optimal size (Hastie et al. 2001). To select the best possible sub-tree from various candidates, a common method is to consider the proportion of records with error prediction (the proportion in which the predicted occurrence of the target is incorrect).

The cost-complexity pruning method is used to generate a series of simpler and simpler trees which relies on a complexity parameter, denoted α. This parameter is gradually increased during the pruning process. If the resulting change in the predicted misclassification cost is less than α times the change in tree complexity, the child nodes in the terminal nodes are pruned away by beginning at the last level. Hence, to measure how much additional accuracy a category must add to the whole tree to warrant the added complexity, α could be used. By increasing α, more nodes of increasing importance are pruned away and simpler trees are resulted.

A predicted outcome class assigned to each node (even the root node). For each child node, node splitting process and the assigning of a predicted class to

each node is repeated and continued recursively till it is not possible to continue. During the tree-building process after pruning, it is impossible to detect which nodes will end up being terminal nodes. The predicted class allocated to each node depends on the decision loss (cost matrix); the assumed prior probability of each class within future datasets; and in each node, the fraction of subjects with each outcome in the learning dataset that end up. A node is assigned to class I if:

$$\frac{C(j|i)\pi(i)N_i(t)}{C(i|j)\pi(j)N_j(t)} > \frac{N_i}{N_j} \tag{3}$$

where $C(j|i)$ is cost of classifying i as $j, \pi(i)$ is prior probability of i, N_i is number of class i in dataset and $N_i(t)$ is number of class i in node. This method of node class assignment guarantees that the tree has a minimal expected average decision cost for future datasets. It is comparable to the learning dataset in which the probability of each outcome is equivalent to the assumed prior probabilities.

Optimal Tree Selection: The best tree which fit the learning dataset with higher accuracy is maximal tree which its performance on the original learning dataset (called the re-substitution cost), usually overestimates seriously the performance of the tree on an independent dataset (which obtained from a similar population). The aim of selecting the optimal tree is to reach the correct complexity parameter α in the way that not only the information in the learning dataset is fit but also not overfit. An independent data set generally is needed to find this value for α, but applying the cross validation method, this necessity can be avoided.

To validate a model building procedure, cross validation which avoids the requirement for a new validation dataset can be used. In CART cross validation, the entire tree building and pruning sequence is conducted N times and N sequences of trees are produced. To generate the tree performance estimate in predicting outcomes for a new independent dataset, as a function of terminal nodes number or complexity, trees within the sequences are matched up according to their terminal nodes number. The outcome of this process is a data-based estimate of the tree complexity that results in the best performance based on an independent dataset. When the tree is adequately complex to fit the learning dataset information but not so complex that noise in the data is fit, by applying this method a minimum cost happens (Song and Lu 2015).

3. Results

In this article, CART algorithm is applied to find influential factors on CEB of ever married woman age 12-49 using BDHS 2011 data. The BDHS 2011 was collected by two stages stratified sampling method. Urban and rural areas are strata of each division. The survey selected 18,000 residential households and 17842 ever married women age 12-49 years old were interviewed. Retrospective history of all woman births could be asked by a number of questions from each of them.

3.1 Data Description

Among a number of socio-economic and demographic variables, only eleven explanatory variables are detected to be extremely related to the response variable that are defined as follows: CEB (response variable): a measure of each woman lifetime fertility experience up to the study time is her number of children ever born. It had 22 discrete value from 0 to 21 children. For our study purposes, four categories 0, 1, 2, 3 and more were considered.

- **Place of residence:** It is a place that women were living in the study time that could be even urban or rural areas.
- **Division:** For administrative purposes, the country consists of 7 divisions of Barisal, Chittagong, Dhaka, Khulna, Rajshahi, Rangpur, and Sylhet in 2011.
- **Religion:** It is the women's religion which could be one of the four religions of Islam, Hinduism, Buddhism, or Christianity.
- **Marriage Duration (MD):** The most important social and demographic indicator of women's exposure to the risk of pregnancy is marriage. Childbearing becomes socially acceptable in Bangladesh during a woman's marriage life. MD is a continuous variable in year that has been computed through finding the difference between women's age in the study and marriage time.
- **Women's job status and Husband's job status:** A job is an activity, often regular, and often performed in exchange for payment. These variables had two levels of employed and unemployed for both of women and their husbands.
- **Husband's age:** A continues variable in year that measures the age of women's husband in survey time.

- **Couple's educational level:** The original data set has two collinear variables as husband's and wife's education variables. To avoid multicollinearity problem, Haque et al. 2015) made the variable couple' education merging the two variables which has three categories 0 when at least one of husband or wife is below Secondary School Certificate (SSC), 1 for both completed SSC and 2 for both completed Higher Secondary Certificate (HSC) or above.

- **Contraceptive use:** It is a categorical variable that has four categories of no method, folkloric, traditional and modern methods.

- **Food security status:** The accessibility of food and a person's access to it refer to food security. Food security for a household can be defined as not living in hunger, fear or starvation for family members (Hunt 2009). Food security status has four categories as food secure when ever-married women did not face to any food insecurity conditions which had to be worry about it. Women were in class of mild food insecurity when they rarely or sometimes worry about not having enough food /or were unable to eat favored food. Ever-married women who sacrifice on eating rice and/or rarely or sometimes had to cut back on the quantity by reducing the size of the meal or number of meals were in category of moderate food insecurity. Women who were in category of severe food insecurity were those that never had square meals, and often had to skip the meals, and/or cut-back on food, and/or had to some other grain than rice, and/or asked for food from a relative or neighbor (Bangladesh Demographic and Health Survey 2011).

- **Wealth status:** The wealth status index is constructed by principal components analysis of household asset data. The wealth index is created in three steps which have been described in Bangladesh Demographic and Health Survey 2011. It is considered in this study as Poor status (categories of Lowest and Second quintiles), Middle status (category of middle quintile) and Rich status (categories of Fourth and Highest quintiles) (Uddin et al., 2011).

Table (1) presents CEB crossed by chosen predictor variables. All the eleven predictors were highly correlated to CEB as the chi-square test p-values show in this table. Moreover, except for couple's educational level, the most of women in other variables had CEB equals to 3 and more. CEB of women in at least one bellow SSC of couples' educational level were 3 and more while the most CEB of women in the other educational levels was 1 child; Women in higher educational levels had less number of CEB. There were two continues predictors in this study that descriptive statistics of them were presented in Table (2).

Husband's age and marriage duration had means equal to 39.55 and 15.05 years by standard deviations equal to 11.15 and 9.8 years, respectively.

Mode for CEB of ever married women in this study was equal to 3 children. In this data set, predictors such as husband's age, couples' educational level, and food security status had 6.8, 0.1 and 0.2 percentage of missing values, respectively.

3.2 Decision Tree

Figure (1) presents decision tree for CEB of 17842 ever married women in BDHS 2011 according to the nominated predictors by CART algorithm. The following rules can be extracted from this tree:

- Women whose $MD \leqslant 1.5$ years were childless without influencing any other predictors.
- Women whose $1.5 < MD \leqslant 3.5$ years and didn't use any contraceptive method were childless.
- CEB of women whose $3.5 < MD \leqslant 6.5$ years and didn't use any contraceptive method was 1 child.
- CEB of women whose $1.5 < MD \leqslant 6.5$ years and used contraceptive methods was 1 child.
- CEB of women whose $6.5 < MD \leqslant 10.5$ years and didn't use any contraceptive method or used traditional methods according to their divisions which were Chittagong and Sylhet or others were equal to 2 or 1, respectively.
- CEB of women whose $6.5 < MD \leqslant 10.5$ years and used folkloric and modern contraceptive methods was 2.
- CEB of women whose $10.5 < MD \leqslant 13.5$ or $13.5 < MD \leqslant 17.54$ years and their divisions were Barisal, Dhaka, Khulna, Rajshahi, and Rangpur and their couple's education was at least one bellow SSC were 2 or 3, respectively.
- CEB of women whose $10.5 < MD \leqslant 17.5$ years and their divisions were Barisal, Dhaka, Khulna, Rajshahi, and Rangpur and their couple's education was both completed SSC or both completed HSC was equal to 2.

- CEB of women whose $10.5 < MD \leqslant 17.5$ years and their divisions were Chittagong and Sylhet and their couple's education was at least one bellow SSC was 3.
- CEB of women whose $10.5 < MD \leqslant 17.5$ years and their divisions were Chittagong and Sylhet and their couple's education was either completed SSC or HSC according to their religion which was Islam or other religions were 3 or 2, respectively.
- CEB of women whose $MD > 17.5$ years and their couple's education was at least one bellow SSC was 3 children.
- CEB of women whose $17.5 < MD \leqslant 22.5$ years, their couple's education was both completed SSC or both completed HSC and their divisions were Dhaka, Khulna, and Rajshahi or Barisal, Rangpur, Chittagong and Sylhet were 2 or 3, respectively.
- CEB of women whose $MD > 22.5$ years, the couple's education were both completed SSC or both completed HSC was 3 children.

Misclassification matrix for classification model has been shown in Table (3) which indicates the accuracy of the classification model. The shaded cells in Table (3) signify correct classification or accuracy of the classification trees on Figures (1). The accuracy of the classification trees for this model can be calculated as Equation (4).The result state that the accuracy of the model is 67 percent which indicates that 67 percentages of women's CEB have been classified correctly. This value indicates that misclassification of the model is equal to 33 percent.

Table 1: Children Ever Born Crossed by Predictor Variables

Variables		**Children Ever Born**					**Chi-Square Test**	**p-value**
Name	**Value**	0	1	2	≥3	**Total**		
Place of residence	Urban	10.9	23.6	27.8	37.7	100	150.59	≤ 0.001
	Rural	9.8	19.1	24.0	47.1	100		
Division	Barisal	10.2	21.1	24.1	44.7	100	247.668	≤ 0.001
	Chittagong	9.3	19.8	22.0	48.9	100		
	Dhaka	11.7	21.7	24.2	42.4	100		
	Khulna	10.4	22.5	30.2	36.9	100		
	Rajshahi	10.3	21.3	29.6	38.8	100		
	Rangpur	8.8	21.8	26.4	43.0	100		
	Sylhet	10.4	15.6	19.9	54.1	100		
Religion	Islam	10.3	20.5	24.6	44.6	100	17.92	≤ 0.001
	Hinduism	9.4	22.6	31.1	37.0	100		
	Buddhism	11.1	25.0	19.4	44.4	100		
	Christianity	6.3	12.5	37.5	43.8	100		
Women's job status	Unemployed	10.3	20.5	24.7	44.6	100	35.048	≤ 0.001
	Employed	9.7	22.1	29.3	38.9	100		
Husband's job status	Unemployed	10.0	20.9	25.7	43.5	100	60.108	≤ 0.001
	Employed	14.7	14.7	17.0	53.5	100		
	Don't know	14.7	23.5	17.6	44.1	100		
Couple's education level	At least one bellow SSC	7.5	15.9	22.9	53.6	100	1641.630	≤ 0.001
	Both completed SSC	15.2	29.9	29.6	25.3	100		
	Both completed HSC	18.6	34.1	33.1	14.2	100		
Contraceptive use	No method	18.4	22.1	19.3	40.2	100	1237.363	≤ 0.001
	Folkloric method	1.6	7.8	17.2	73.4	100		
	Traditional method	5.1	13.7	23.3	57.8	100		
	Mordern method	4.0	20.7	30.9	44.4	100		
Food security status	Food secure	11.6	23.1	26.8	38.5	100	484.926	≤ 0.001
	Mild food insecurity	8.2	18.0	22.8	51.0	100		
	Moderate food insecurity	5.4	11.4	21.2	62.0	100		
	Severe food insecurity	7.9	11.2	22.7	58.2	100		
Wealth status	Poor	8.8	17.9	22.1	51.2	100	273.163	≤ 0.001
	Middle	10.4	20.1	24.7	44.8	100		
	Rich	11.2	23.2	28.1	37.5	100		

Table 2: Descriptive Statistics of Continuous Predictor Variable

Variables	Mean	Median	Mode	Std. Deviation
Husband's age	39.55	39.00	30.00	11.15
Marriage duration	15.05	14.00	11.00	9.8.1

Table 3: Misclassification Matrix for CART

Observed Category	Predicted Category				Total
	0	1	2	≥3	
0	blue!10 1201	378	80	158	1817
1	441	blue!10 2088	763	398	3690
2	36	752	blue!10 2094	1635	4517
≥3	3	138	1089	blue!10 6588	7818
Total	36	68	213	88	17842

Table 4: Risk and Standard Error of CART for Training and Learning Data

	Risk	Standard Error
Learning Set	0.329	0.004
k-fold cross validity of training set	0.332	0.004

$$Accuracy = \frac{Numberofcorrectprediction}{Totalnumberofpredictions} = \frac{1201 + 2088 + 2094 + 6588}{17842} = 0.67 \tag{4}$$

Risk and standard error of classification tree in Figure (1) for training and learning data have been shown in Table (4). To fit CART algorithm to data set, data divided to two different groups of training and learning data and the model fits to these two groups. Indeed, training data for fitting the model and learning data for confirming the validity of the model are used. When the risk of these two data groups is close to each other, it confirms the validity of the fitted model. According to the results of Table (4), these values are almost equal which indicates the validity of classification model proposed by classification tree in Figure (1).

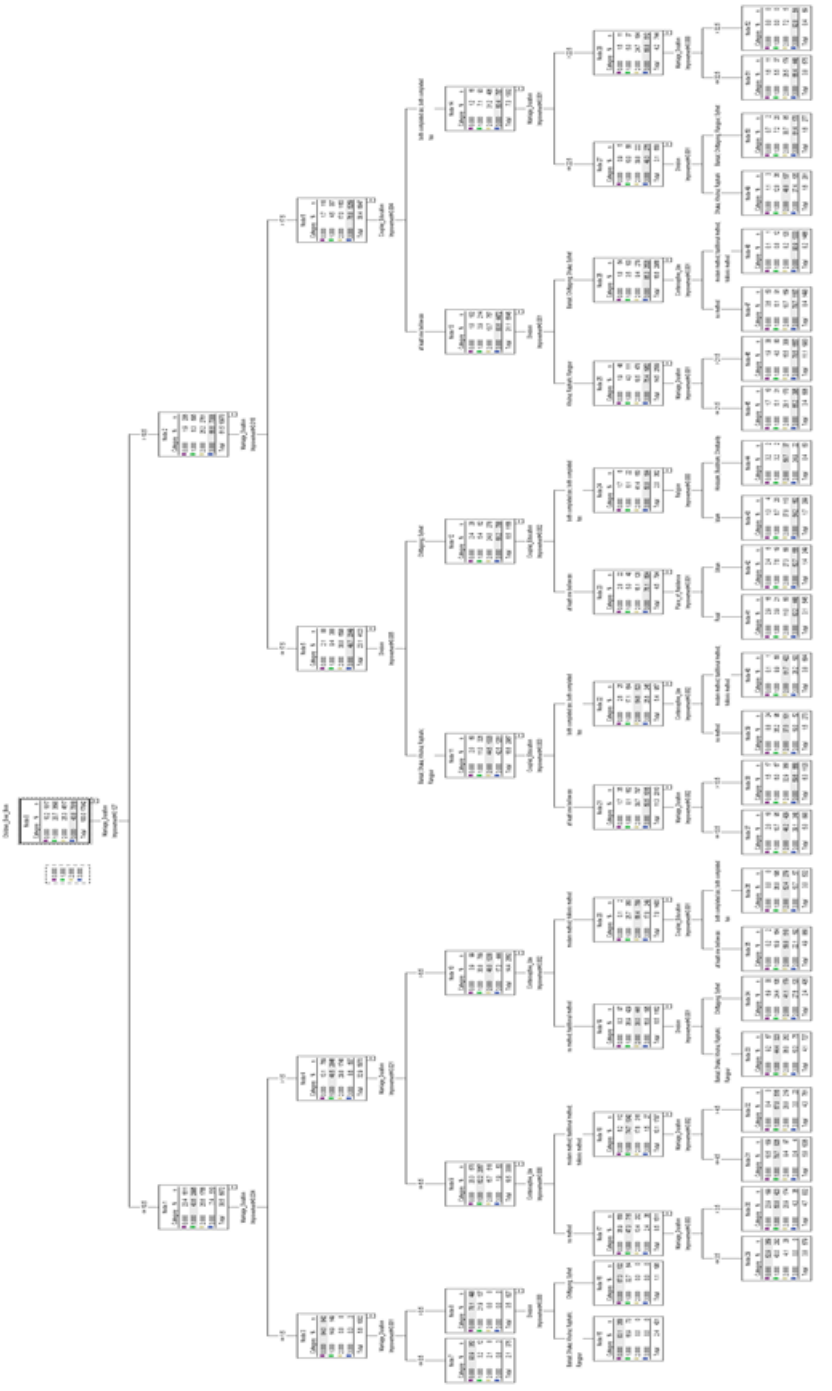

Figure 1: Decision Tree for CEB of Ever Married Women in Bangladesh by CART Algorithm

4. Conclusions

Since Bangladesh is going under a rapid growth of population, modeling children ever born variable is very important. There are many researchers that are working to the most efficient factors on this growth (Abedin and Rahman, 2012; Hasan and Sabiruzzaman, 2008; Haque et al. ,2015; Uddin et al., 2011; Farhana, 2013).

Islam e al. (2013) showed the effects of socio-demographic factors on CEB for domestic and non-domestic violence. Ahmmed and Nasser (2012) compared PR to Support Vector Machines (SVM), one of the well-known data mining methods. Kirkos et al. (2008) tried to study the three modern nonparametric techniques including SVM, neural network and decision tree and compared their generalization performances with classical ones through the task of classifying high risk-low risk child bearing pattern. Haque et al. (2012) attempted to compare multinomial logistic regression, ordinal logistic regression and decision tree (C4.5) on the basis of their capacities in predicting the total number of children ever born in 2007 BDHS data.

If CEB is considered as a count or ordinal variable, traditional methods like PR, GPR, NBR or CL can be used but they have some drawbacks. Thus to solve these problems, CART algorithm that is a dominant method with significant potential and analytical usefulness is introduced. The application of CART has been growing and is possible to increase in the future largely due to this fact that it is the best available solution for the considerable number of important problems.

According to CART decision tree in Figure (1), most of the predicted CEB was 3 and more according to the extracted tree which was still far from the government's goal of Bangladesh till 2016. To make family size small, the government should take essential steps to increase the education rate (76.9 percent couples are still at least bellow SSC), women's employment rate (86.7 percent of women are unemployed) and wealth status (36.1 percent of women are poor indexed).

Marriage duration (MD) was the most important predictors in this study which has been situated in the root. Longer MD resulted in more CEB. The length of marriage has a direct effect on fertility. It effects on lengthening the reproductive period and subsequently increasing fertility. Since older women have had a longer reproducing time comparing to younger women, it is realistic to anticipate a lower fertility for younger females comparing to older females. Many studies also declared this result (Bangladesh Demographic and Health

Survey 2011; Ahmmed and Nasser, 2012).

Couple's educational level also played a critical rule in this tree. CEB values changed on the educational level. Higher educational levels for couples resulted in less CEB. Haque et al. (2015) stated the same results.

Divisions also affected CEB of ever married women. According to the results of Figure (1), Sylhet followed by Chittagong divisions had the most CEB among the other divisions as 3 and more children (the same results can be concluded from Uddin et al. (2011), and Ahmmed and Nasser, 2012).

From Table 1, shows that 90 percent of people in Bangladesh are Muslim. According to the resulted decision tree, Ahmmed and Nasser (2012), Haque et al. (2015), and Uddin et al. (2011), CEB of Muslims are more than the other religions.

Acknowledgement

The authors are thankful to the Referees and Editor for their constructive comments and suggestions. This article is extracted from a survey under the title of âMining Demographic Data by Decision Treeâ which is supported by National Population Studies and Comprehensive Management Institute in Iran, in 2014 by the registered number of 20/15283.

References

Abedin, S. and Rahman, J. A. M. (2012). On the dynamics of high-risk fertility in Bangladesh. *International Journal of Human Science*. 9(2):1371-1378.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons, Inc.: New Jersey.

Ahmmed, F. and Nasser M. (2012). Modeling and Predicting of Children Ever Born in Bangladesh. *International Conference on Statistical Data Mining for Bioinformatics Health Agriculture and Environment*, Department of Statistics, University of Rajshahi, Bangladesh.

Asaduzzaman, M.d., Rahaman H., Khan, Md. (2009). Identifying Potential Factors of Childbearing in Bangladesh. *Asian Social Science*. 5(3): 147-154.

Bangladesh Demographic and Health Survey (2011). *National Institute of Population Research and Training Dhaka*. Bangladesh, January 2013.

Bagheri, A. (2018). Studying the Influential Factors of Children Ever Born of Migrant Women to Tehran. *Scientific Journal of Ilam University of Medical Sciences*. 25(6):118-129.

Bagheri, A. Saadati, M. Razeghi, N BB. (2017). Identification of fertility preference determinations using Poisson regression. *Iranian Journal of Epidemiology*. 13(2):153-161.

Bagheri, A. Saadati, M (2014). Introduction and Application of CART Model to Classify Ideal Number of Children for 15-49 Year-Old Women, Semnan providence. *Journal of Population Association of Iran*. 17:77-111.

Berry, M.J.A., Linoff, G. (1999). *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley and Sons, Inc.: New York.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Inc.: California.

Famoye F. (1993). Restricted generalized Poisson regression model. *Communications in Statistics-Theory and Methods*. 22(5):1335-1354.

Farhana, S. (2013). Performance of Generalized Poisson Regression Model and Negative Binomial Regression Model in case of Over-dispersion Count Data. *International Journal of Emerging Technologies in Computational and Applied Sciences*. 4(6): 558-563.

Haque, M.d. A., Hossain Md. T. and Nasser M. (2015). Predicting the Number of Children Ever Born Using Logistic Regression Model. *Biometrics & Biostatistics International Journal*. 2(4): 4 pages.

Haque, M.d. A., Hossain M.d.T. and Nasser, M. (2012). Comparison Between Logistic Regression and Decision Tree in Predicting the Number of Children Ever Born to Matured Women in Bangladesh. *International Conference on Statistical Data Mining for Bioinformatics Health Agriculture and Environment*, Department of Statistics, University of Rajshahi, Bangladesh.

Hasan, M. and Sabiruzzaman (2008). Factors Affecting Fertility Behavior in Bangladesh: A Probabilistic Approach Research. *Journal of Applied Sci-*

ences. 3(1):70-76.

Hastie, T.J., Tibshirani, R.J. and Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Second Edition. Springer.

Hilbe, J. M. (2011). *Negative Binomial Regression.* Cambridge University Press: New York.

Hunt, P. (2009). World Food Crisis Worsens. *Irish Independent.* October 13, http://www.gorta.org/pdf/InTuition_13Oct09_Irish_Independent.pdf.

Islam, R., Alam, R. and Islam, R.B. (2013) Effects of socio-demographic factors on children ever born for domestic and non-domestic violence: Application of Path model. *Global Advanced Research Journal of Social Science.* 2(2): 38-46.

Kirkos, E., Spathis, C., and Manolopoulos, Y. (2008). Support vector machines, Decision Trees and Neural Networks for auditor selection. *Journal of computational Methods in Sciences and Engineering.* 8(3): 213-224.

King, G. (1989). Variance specification in event count models: From restrictive assumptions to a generalized estimator. *American Journal of Political Science.* 33: 762-784.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society.* Series B 42: 109-142.

Saadati M (2015). Factors Affecting Children Ever Born for 15-49 Year-Old Women in Semnan Using Poisson Regression. *Journal of Health System Research.* 11(3):627-637.

Saadati M, Bagheri A. (2015). Mining children ever born data; classification tree approach. *Indian Journal of Science and Technology.* 8(30):1-8.

Song, Y. and Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Biostatistics in psychiatry.* 27(2): 130-135.

Timofeev, R. (2004). *Classification and Regression Trees (CART)Theory and Applications.* Unpublished MSc. Thesis, Center of Applied Statistics and Economics Humboldt University, Berlin.

Uddin, Md. I., Bhuyan, K. C. and Islam, S. S. (2011). Determinants of desired family size and children ever born in Bangladesh. *The Journal of Family Welfare*. 57(2): 39-47.

Winkelmann R. and Zimmermann, K. F. (1994). Count data models for demographic data. *Mathematical population Studies*.4: 205-221.