International Journal of Sciences:
Basic and Applied Research
(IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Application of the CART Model to Classify the Perception of Young Canadian Teenagers on the Effect of Marijuana on Health

Ruben Thoplan[a]*

[a]*Department of Economics and Statistics, Faculty of Social Studies and Humanities, University of Mauritius, Réduit, Mauritius*
[a]*Email: r.thoplan@uom.ac.mu*

Abstract

The use of the illicit drug, marijuana has increased over years among young teenagers in different parts of the world and its harm on the health is generally well-known. This paper attempts to study the perception of young adolescents of 13-15 years old residing in Canada towards the danger of marijuana on health. To do so, a classification and regression tree (CART) has been applied on the data from the 2012 National Anti-Drug Strategy (NADS) Youth Advertising Recall and Tracking Survey. The decision tree has been applied and pruned on a training data set (70%) and evaluated on the testing data set (30%). The results show that the main indicators which impact on the perception of a teenager towards the harm marijuana has on health are the perceptions towards psilocybin (another illicit drug), the province in which the teenager lives and whether he/she has been ever offered drugs. The overall error rate on the testing data set based on the confusion matrix is less than 20% and the area under the ROC curve is relatively high showing that the model is accurate in classifying the perception of young teenagers on the health marijuana has on health.

Keywords: marijuana; CART; pruned; psilocybin; ROC curve

* Corresponding author.
E-mail address: r.thoplan@uom.ac.mu.

1. Introduction

Marijuana, also known as cannabis, is an illicit drug which is known to have negative impacts on different spheres of life of a person. It has been shown by [1], that regular use of cannabis among adolescent is associated with other problems like the consumption of other prohibited drugs, crime, depression and suicidal behaviors. According to [2], the Canadian young adolescents top the lists of developed countries in the use of marijuana. The National Anti-Drug Strategy is an initiative of the government of Canada based on prevention, treatment and enforcement of law to combat the use of illicit drugs. As part of the prevention action plan, a mass media campaign was launched in 2012 to discourage young people from using illicit drugs, [3].

To evaluate the effect of the campaign in 2012, an advertising tracking and recall survey was conducted by the department of Justice of Canada among young adolescents of 13-15 years old. Based on the results from [4], it is reported that youth who saw the advertising in 2012 are more likely to be knowledgeable about drugs, their effects on physical health and about identifying marijuana to be harmful. In this paper, the perception of the young Canadian adolescents towards the harm that marijuana has on health is investigated. To do so, a classification and regression tree (CART) is applied to the 2012 National Anti-Drug Strategy (NADS) Youth Advertising Recall and Tracking Survey data obtained from [5].

It has been observed by [6] that as the rating of perception of risk associated to smoking marijuana gets lower, the rates of smoking marijuana increase. It is therefore important not to underestimate the perception of risk associated to marijuana among the young adolescents. Besides, sensitizing the young adolescent about the harm that marijuana has on the health is primordial. This paper attempts to understand the specificities of the young Canadian adolescents who perceive marijuana as dangerous to health and those who do not, through the CART method.

The paper is organized in four sections such that the second section introduces the CART methodology and shows its appropriateness to answer the objective of this study. In the third section, a CART model is applied to the data and the results are provided. The model is also evaluated to appreciate its accuracy in the classification task of young adolescents who distinguish marijuana to be dangerous to health and those who do not. The final section attempts some discussions and recommendations.

2. The CART Method

A decision tree consists of different set of nodes starting from the root node which splits into several branches leading to other nodes which further splits in other branches as required and terminate with a leaf node, [7]. A decision tree is very intuitive and easy to understand and can be used as classification models and regression models. The CART method, used for the purpose here, is a binary tree classification tool in data mining which can be used to analyze categorical or continuous outcomes. The analysis of categorical outcomes results in classification trees whereas the analysis of continuous outcomes results in regression trees. CART has had many practical applications like in Clinical trials [8], credit scoring [9], ecology [10] and other areas.

In brief terms, the CART model is built by first finding the variable which best splits a data into two groups through an information gain measure, [7]. The data is then partitioned and the splitting criterion is applied again on the partitioned data, and the process continues recursively until very few observations are left or until no updating can be done anymore. The splitting criterion used in this paper is based on the entropy, an impurity measure which lies between 0 and 1. Entropy of 0 implies that all observations belong to the same class and a value of 1 implies that more information is required to classify observations in a specific class. In the case of CART, the tree which provides the greatest reduction in entropy would be considered as the best split. Adapting the notation used in [7], the entropy of a data can be formulated as follows:

$$\inf o(D) = -p\ln(p) - n\ln(n) \tag{1}$$

where p and n represent the proportion of the dataset in the “dangerous” and “ not dangerous” category respectively for the perception of young adolescents regarding the harm marijuana has on health.

Prior to performing any analysis, the data has been pre-processed such that the levels of the different categorical variables to be used as inputs have been merged to obtain dichotomous variables. The reason for this aggregation is to come up with decision rules that are easily interpretable from the CART. For example, responses like “somewhat agree” and “strongly agree” have been merged to a new category “agree” and “somewhat disagree” and “strongly disagree” have been merged to “disagree”. After all necessary pre-processing has been carried out; the model needs to be evaluated. To do so, the data is portioned into a training dataset of 70% randomly selected cases and a testing data set of 30% randomly selected cases. The model is applied on the training data set and evaluated on the testing set.

While running the model on the training data set, a cost-complexity pruning has to be carried out, [11] to avoid over-fitting. In simple terms, cost complexity represents a measure of the mean error reduced for each leaf node. To select the optimal sized tree, the one standard error (1-SE) rule can be used. Formally, the 1-SE rule is the smallest tree T_k which is obtained by choosing the first largest value of C_p such that

$$x_{error} \leq \min(x_{error}) + x_{std} \tag{2}$$

where C_p represents the complexity parameter, x_{error} represents the average cross validation error and x_{std} represents the estimate of the standard deviation of the cross validation error.

To evaluate the model on the training set, a confusion matrix and the Receiver Operating Characteristic (ROC) curve are used. From the confusion matrix, the overall error rate, false positive and false negative error rate can be obtained. The ROC curve plots the true positive rate against the false positive rate. If the CART model predicts “dangerous” in agreement with the actual outcome, this is referred to as true positive. However, if the model predicts “not dangerous” when the actual outcome is “dangerous”, then there is a false positive. The higher the area under the ROC curve, the more accurate is the classification. In this paper, all analyses are carried out using the rattle package available in the R programming language, [7, 12].

3. Results

In this section, the results obtained from rattle are presented. First, the decision tree is built fully on the training dataset and pruned to come up with an interpretable result. The cross validation relative errors are plotted for different complexity parameter as per figure 1. Using the 1-SE rule, the first largest value of C_p which satisfies equation (2) is 0.019. Using this complexity parameter, the decision tree is pruned accordingly as observed in figure 2.

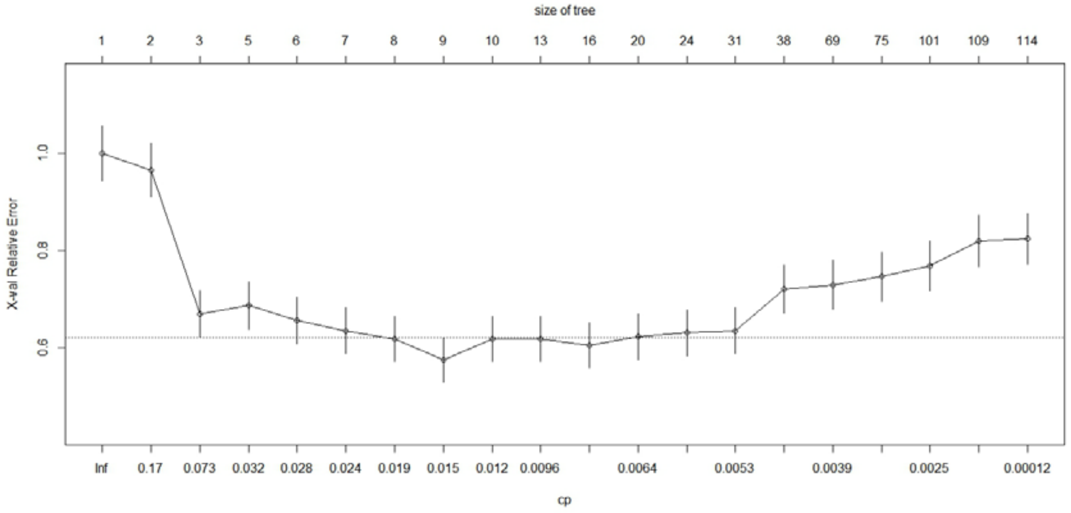

Figure 1: Plot of Cross Validation Relative Errors against Complexity Parameters

The results from figure 2 clearly shows that the perception young teenagers have towards the harm that marijuana causes on health is related to their perception regarding the frequency of use of marijuana and the harm on health which psilocybin causes. It is also related to the province in which they live and whether they have been offered drugs or not. The decision tree reads “yes” for every branch on the left of a node and “no” for every branch on the right of a node.

In this particular example, among adolescents who consider the harm of smoking marijuana once in a while to be serious, it is observed that there is a higher chance for an adolescent who does not live in Quebec or Manitoba and who perceives Psilocybin not to be dangerous to identify marijuana as being not dangerous to health. On the other hand, those who consider the harm of smoking marijuana once in a while to be not serious will identify marijuana as dangerous to health if they live in Quebec. Besides, an alarming result can be identified from a specific group of young adolescents who do not live in Quebec and who do not consider the harm of smoking marijuana once in a while to be serious. This group involves those who perceive Psilocybin to be dangerous to health and has been offered drugs in the past. The worrying result is that there is a higher risk for this category of young adolescents to recognize marijuana as being not dangerous to health.

To know whether the above statements made are valid, the decision tree needs to be evaluated on the testing data set. In a first instance, the confusion matrix is presented as per table 1. The overall error rate is 19.8% which is relative small for a classification task. Besides, the false positive and false negative error rates are both less than 15% which is quite small. This indicates briefly the accuracy of the model to predict a young adolescent who perceives marijuana to be dangerous to health and those who do not.

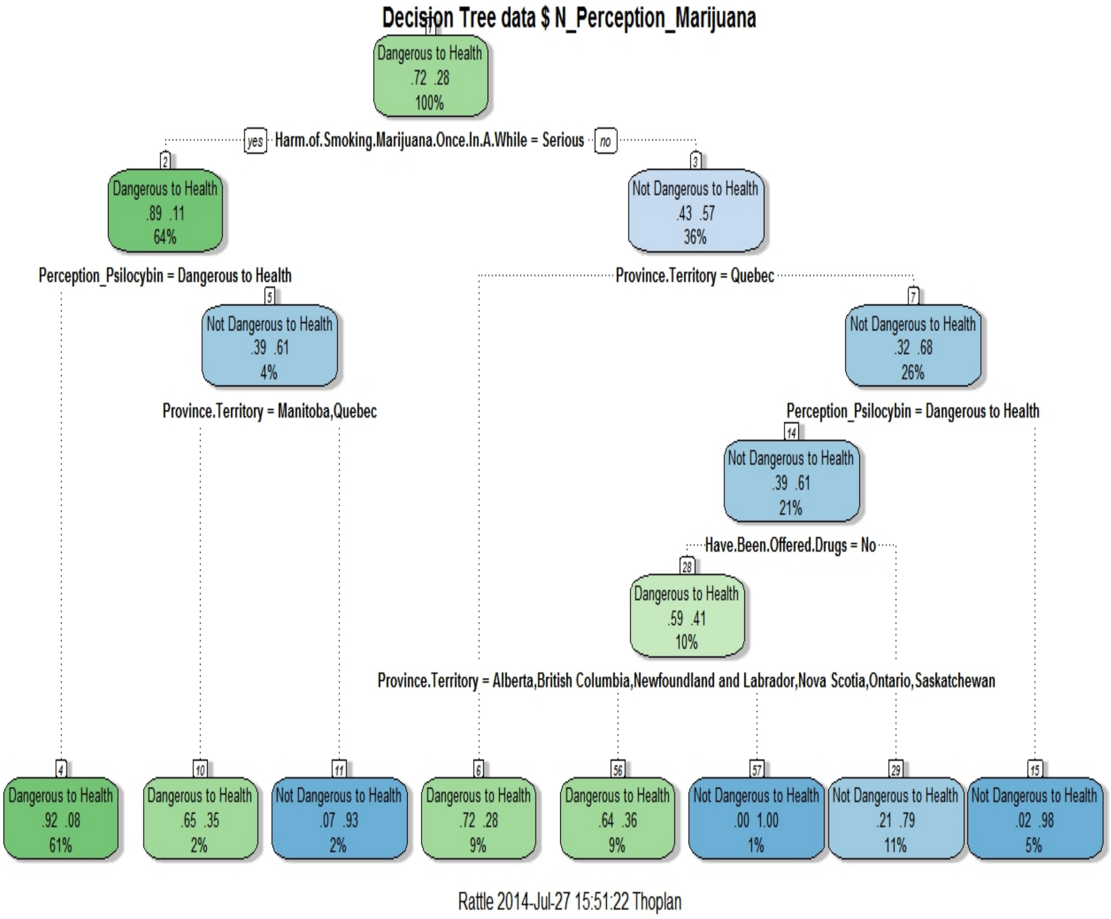

Figure 2: Classification and Regression Tree for Perception of Young Teenagers Regarding Harm marijuana has on Health

Table 1: Error Matrix on Testing Data Set (%)

Actual	**Predicted**	
	Dangerous to Health	**Not Dangerous to Health**
Dangerous to Health	63.9	6.3
Not Dangerous to Health	13.5	16.3

To confirm the accuracy of the classification model, the ROC curve is built and presented in figure 3. The further away the curve is from the diagonal line, the higher is the accuracy of the model. In this example, we have a relatively good accuracy as the area under the ROC curve is relatively large.

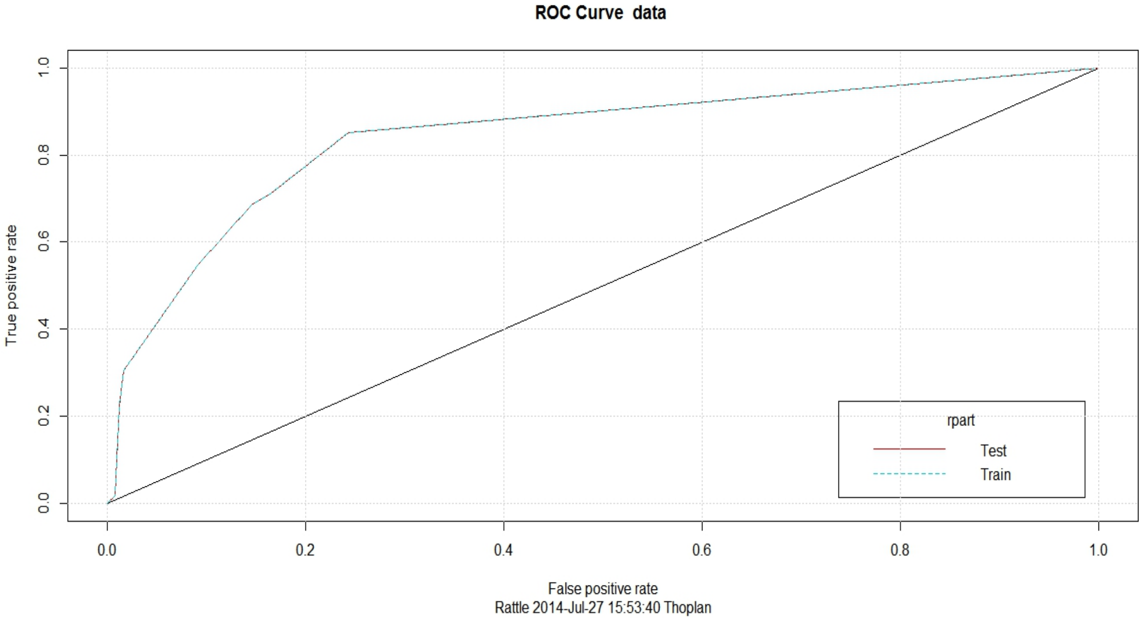

Figure 3: ROC Curve for Training and Testing Data Set

4. Discussions and Recommendations

In this paper, the perception of young Canadian adolescents has been scrutinized thoroughly using the CART method. As stated by [6], perception of risk is associated to the smoking of marijuana. In this effect, the perception of a young adolescent towards the harm that marijuana has on health cannot be neglected. It is primordial to understand the category of young adolescents who do not consider marijuana to be harmful to health and those who do. Some key factors which are related to the perception of a young adolescent in this effect have been identified. For example, the province and whether the adolescent has been offered drug in the past will impact on his/her perception towards the harm that marijuana has on health.

As a whole, there is a high probability for the young teenagers who do not live in Quebec to perceive marijuana as being not dangerous to health. Indirectly, if we consider the link made by [6], there is a lower chance for the young teenagers of Quebec to consume marijuana compared to other provinces. In this respect, more investigation must be carried out to understand the reasons why the young of Quebec perceives marijuana to be dangerous compared to the young of other provinces. In so doing, strategies could be developed for policy implications using Quebec as benchmark in view of fitting drug addiction among young adolescents. Universal and selective preventions can be carried among mainly among young adolescents of non-Quebec provinces. Universal prevention can also be carried out among all non-Quebec states through a mass media campaign targeting the young adolescents. Selective prevention can be carried out in colleges among those who have already been offered drugs in the past to avoid them from trying marijuana in the future. This paper therefore provides a basis of Canadian policy makers to target their sensitization campaign accordingly.

As a further research, the random forests algorithm could be applied in view of improving the classification accuracy. The random forests algorithm is another classification technique in data mining which produces very accurate predictions and gives a measure of importance for each variable in the data set.

References

[1] D. M. Fergusson, L. J. Horwoord and N. Swain-Campbell, “Cannabis use and psychosocial adjustment in adolescence and young adulthood,” *Addiction,* vol. 97, no. 9, p. 1123–1135, 2002.

[2] Unicef Office of Research, “Child Well-being in Rich Countries: A comparative overview,” UNICEF Office of Research, Florence, 2013.

[3] Government of Canada, “NATIONAL ANTI-DRUG STRATEGY IMPLEMENTATION EVALUATION Final Report,” Evaluation Division, Office of Strategic Planning and Performance Measurement, Canada, 2010.

[4] Government of Canada, “Horizontal Initiatives - Departmental Performance Report 2011-12 – Supplementary Information Tables,” Department of Justice, 2012. [Online]. Available: http://www.justice.gc.ca/eng/rp-pr/cp-pm/dpr-rr/2011_2012/supp/hi-ih.html. [Accessed 29 07 2014].

[5] Government of Canada, “data.gc.ca,” Health Canada, 15 February 2012. [Online]. Available: http://data.gc.ca/data/en/dataset/ef6aa6eb-8c38-4a54-b8cf-469f810605ba. [Accessed 29 July 2014].

[6] NSDUH, “The NSDUH Report: Trends in Adolescent Substance Use and Perception of Risk from Substance Use,” Substance Abuse & Mental Health Services Administration, Center for Behavioral Health Statistics and Quality, Rockville, MD, 2013.

[7] G. Williams, Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Australian Capital Territory: Springer, 2011.

[8] K. R. Hess, M. C. Abbruzzese, R. Lenzi, M. N. Raber and J. L. Abbruzzese, “Classification and Regression Tree Analysis of 1000 Consecutive Patients with Unknown Primary Carcinoma,” *Clinical Cancer Research,* vol. 5, pp. 3403-3410, 1999.

[9] T.-S. Lee, C.-C. Chiu, Y.-C. Chou and C.-J. Lu, “Mining the customer credit using classification and regression tree and multivariate adaptive regression splines,” *Computational Statistics & Data Analysis*, vol. 50, no. 4, pp. 1113-1130, 2006.

[10] G. De'ath and K. E. Fabricius, “CLASSIFICATION AND REGRESSION TREES: A POWERFUL YET SIMPLE TECHNIQUE FOR ECOLOGICAL DATA ANALYSIS,” *Ecology,* vol. 81, no. 11, p. 3178–3192, 2000.

[11] L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen, Classification and Regression Trees, Chapman and Hall/CRC, 1984.

[12] T. M. Therneau and E. J. Atkinson, “An Introduction to Recursive Partitioning Using the RPART Routines,” Mayo Foundation, 2014.

[13] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.