Scholars Journal of Applied Medical Sciences (SJAMS)
Sch. J. App. Med. Sci., 2017; 5(11A):4311-4329

(An International Publisher for Academic and Scientific Resources)
www.saspublishers.com
ISSN 2320-6691 (Online)
ISSN 2347-954X (Print)
DOI:10.36347/sjams.2017.v05i11.006

Exploration of Risk Factors Associated with Adolescent Alcohol Consumption using Cutting Edge Recursive Partitioning Techniques

Kuldeep Kumar[1], V.K Tiwari*[2], Sherin Raj[3], Niharika Kapadia[4]
[1]Professor, School of Business, Bond University, Gold Coast, Queensland, Australia
[2]Professor & Head, Dept. of Planning & Evalutaion, National Institute of Health and Family Welfare, New Delhi
[3]ARO, National Institute of Health and Family Welfare, New Delhi
[4]Pursuing MBA, School of Business, Bond University, Gold Coast, Queensland, Australia

Original Research Article

***Corresponding author**
Prof. V.K.Tiwari

Article History
Received: 26.10.2017
Accepted: 03.11.2017
Published: 30.11.2017

Abstract: The purpose of this article is to explore and identify risk factors influencing drug use in school going adolescents aged 10 to 19 in a hilly state in the North-Eastern part of India. This article will explore the data collected from the National Institute of Health and Family Welfare, New Delhi, by using cutting edge Recursive Partitioning techniques such as Discriminant Analysis, Decision Tree Method, Artificial Neural Network and the Stochastic Gradient Boosting to build a predictive model. Out of 3069 randomly selected participants who undertook the Adolescent Reproductive and Sexual health (ARSH) questionnaire a subset have been used to form this data set. Utilization of Artificial Neural Network, Stochastic Gradient Boosting and the Random Forest models produce higher accuracy and classification in contrast to other measures. These models will be useful in the prediction of associated risk factors that contribute to adolescent alcohol consumption.
Keywords: Adolescents, Alcohol risk factors, Artificial Neural Networks, Decision Trees, Random Forest, Stochastic Gradient Boosting

INTRODUCTION

Alcohol consumption among adolescents is becoming increasingly prevalent, and is causing serious life threatening complications on a global scale [1]. Studies have shown that underage drinking can significantly affect physiological and psychological development. In addition to these developmental effects, adolescents are more likely to engage in other detrimental behaviours such as illicit drug use, risky sexual behaviours, and victimisation [1].

These behaviours are more likely to manifest in those children and adolescents that consume alcohol at an earlier age. Studies that assess the risk factors that may significantly contribute to adolescent alcohol use, is providing useful frameworks for intervention programs [1].

Until recently, most studies on alcohol consumption have largely been conducted in developed-western countries. Global research, however, is revealing that developing countries require more emphasis, India being of increasing concern, as the prevalence of alcohol consumption in this country has increased by 55% over the past two decades. Interventions are largely focused on deterring adolescent use by addressing the associated risk factors of alcohol consumption. Despite success in determining these factors in adults, complexity still remains in identifying risk factors in adolescence [1].

Studies predominately approach the identification of associated risks factors for alcohol consumption based on two stages – factors that influence initiation, and facilitate ongoing use. Gopiram and Kishmore [2] focused on a study of users, and non-users, and elucidated that an individual's sense of curiosity, state of wellbeing, and their social network, are strong drivers that initiate alcohol consumption [2]. These results are reinforced in a study by Saddichha, Sinha, and Khess [3] that conducted research in patients recovering from alcohol addiction at a rehabilitation facility [3]. It was revealed that peer pressures, role models, and the nurturing environment contributed to the initiation phase of addiction. In terms of the continued addiction to alcohol, patients reported that their social network and other psychosocial contexts such as work, and traumatic past events, contributed to their ongoing use. The aforementioned studies provide insights into the emergent factors that influence adolescent alcohol use. A plethora of research demonstrates that the nurturing environment, and a family history of alcohol consumption are significant predictors of alcohol use in adolescence. Other psychosocial predictors include: peer substance abuse, the rate of change in societal structures, exposure to certain technologies, and parental methods employed [3].

A review of the literature demonstrates the ARSH dataset is best explored by the following categories: psychosocial and peer factors, demographic, socio-economic class, media exposure, and the use of alcohol, tobacco, and illicit drugs. As a scan of the literature reveals these factors as likely to contribute to alcohol consumption, there is an emerging concern to identify which of these variables contribute to adolescent alcohol use. These associated risks factors will be explored through the analysis of particular sub-sections of the ARSH data set.

There is now emphasis on creating predictive models that focus on these risk factors and these are explored in the data collection from the National Institute of Health and Family Welfare, New Delhi (NIHFW). This paper examines the variables that influence alcohol consumption in adolescence. This study includes the following research objectives:

- To examine and identify the main variables leading to alcohol consumption in adolescents.
- To create a model through percussive techniques that uses risks factors to measure the likelihood of alcohol consumption in adolescents.

MATERIAL AND METHODS

Data collection was performed by the National Institute of Health and Family Welfare, New Delhi. The data set was generated by Tiwari *et al*. [4] using a questionnaire as part of a study on Adolescent Reproductive and Sexual Health (ARSH) in Mizoram, August 2012 [4]. Data was collected from 3069 randomly selected participants aged from 10 to 19 years from private, missionary and government schools across two locations (Aizawl and Champhai district), both serviced by ARSH Programs [4]. For the purpose of this study, various non-disruptive variations were made, reducing the data set to 3041 participants. The survey consisted of 121 questions and only 67 were found to be relevant and applicable for the analysis of report. The variables used in this report can be categorised into social, demographic and behavioural factors affecting adolescent alcohol consumption and can be seen below:

- Demographic: Sex, Age, Marital Status, Grade, Subject Stream, Type of Education, Primary language of Education, Part-Time Employment, Part- Time Earnings, Household Income and Type of Family.
- Substance Use & Frequency: Tobacco, Drugs (illicit and medicinal), and Alcohol Frequency
- Social Activity: Attending Party/Picnic, Substances Available, Leisure Activities, Pornography Usage
- Reasons for Substance Use
- Social/Peer Substance Use and Frequency
- Following predictive modelling techniques are applied to the above mentioned data set and their predictive power was obtained.

Direct Logistic Regression

Logistic Regression is a commonly used technique to study the relationship of set variables to determine their predictive power and contribution in determining particular outcomes.

Discriminant Analysis (DA)

The aims of DA are to develop a discriminate function that groups one or more continuous or binary independent variables as a measure of predicting the dependent variable.

Artificial Neural Networks (ANN)

The Artificial Neural Networking (ANN's) has been the most widely used method of data mining application due to the ease of use, technological power and flexibility. ANN's models such MLP have a specific architectural map consisting of three primary layers: input, hidden, and output. The hidden layer is described as the middle component and is termed 'the activation function' as it operates to form complex linear relationships between the input and output layers [5].

Decision Trees

The Decision Tree (DT) also known as a classification tree is a conventional statistical analysis technique which maps observations (predictor/independent variables) about an outcome or an item (target/dependent variable). Observations are represented as branches and target variables as leaves. This analyses tool allows for easy and effective algorithm interpretation [6]. The DT is built on three important components: (1) The selection of the splits, (2) The decisions when to declare a node terminal or to continue splitting it (3) The assignment of each terminal node to a class [7]. Decision trees have may properties and capable of handling variable selection, variable interaction detection, non-linear relationship detection, missing value and outlier handling etc.

Random Forest

The Random Forest (RF) is an extension of the DT method. It uses a multitude of decision trees which resembles a 'forest-like' map that classifies an object.

Random forest algorithm consists of drawing a bootstrap sample and then fitting a large CART tree to this bootstrap sample which is unpruned. At each split in the tree we consider only limited number of randomly selected variables. These steps are repeated 200-500 times and finally we average the predictions to predict a new record. Random forests have superior predictive performance over CART trees and have lower variance as compared to a single CART tree. All the properties of DT are inherited in random forest. However, they are not as interpretable as a single CART tree. The performance of RF depends on number of trees and random number of variables chosen at each split. One method to interpret Random Forest is through variable importance which is done by computing variable importance score in each CART tree in the forest and then taking the average of the values for each variable.

Stochastic Gradient Boosting (Using TreeNet)

The Stochastic Gradient Boosting method using TreeNet is a powerful data mining approach based on the DT process. The algorithm synthesises thousands of small decision trees that are built in a sequential error-correcting process to formulate an accurate model for regression and classification. Benefits of this model include: Automatic predictor selection, Resistance to outliers, Resistance to over fitting via a slow update process and compensatory mechanisms for data omissions [8].

RESULTS

Logistic Regression

Logistic Regression has been performed to determine the significant risk factors that lead to alcohol youth consumption. Of the independent variables 67 were analysed as shown in Appendix 1.1. Interpretation of the Omnibus Tests of Model Coefficients was considered first to assess the performance and “goodness of fit” of the model by addressing that the explained variance in the data is significantly greater than the unexplained variance. The Hosmer and Lemeshow test reinforced the performance of the model with a significance level greater than 0.05 (Appendix 1.2 and 1.3). In addition, the Cox & Snell and Nagelkerke pseudo R square statistics showed that between 74.3% and 100% of the variability is explained by this set of variables (Appendix 1.4). Inclusion of these tests provides adequate evaluation for model fitness and performance.

Table 1 below illustrates how well the model is able to forecast the correct category for each case. It seems for original observations model can correctly classify 92.5% observations. However, when we do the cross validation it classifies only 81.4% observations correctly.

Table-1: Logistic Regression Classification Table

Classification

		Alcohol	Predicted Group Membership		Total
			Yes	No	
Original	Count	Yes	924	214	1138
		No	12	1891	1903
	%	Yes	81.2	19.8	100.0
		No	0.7	99.3	100.0
Cross-validated[b]	Count	Yes	832	306	1138
		No	258	1645	1903
	%	Yes	73.1	26.9	100.0
		No	13.6	86.4	100.0

a. 92.5% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. .81.4% of cross-validated grouped cases correctly classified.

Values shown in Appendix 1.1 which are less than 0.05 have been identified as significant. The significant variables consist of: age in months and friends consuming alcohol. The strongest predictor of adolescent alcohol consumption was friends consuming alcohol, with an odds ratio of 0.742. This result confirmed literature findings and indicated that adolescents who consumed alcohol were 0.742 times more likely to if they had friends consuming alcohol. The derived logistic regression equation for forecasting adolescent alcohol consumption is modelled as the following:

Z= 213.329 - .299 (Friends Consuming Alcohol) - 0.697 (Age in Months)

The above regression model indicates that if the probability (z) is more than 0.5 we can be 95% confident that the risk factors are associated alcohol consumption in adolescents. If this probability is less than this threshold we can be 95% confident that the variables are not associated with alcohol consumption

DISCRIMINANT ANALYSIS

The purpose of discriminant analysis is to predict risk factors that contribute to adolescent alcohol consumption. This method enables us to determine which independent variables are significantly influencing alcohol consumption and those independent variables which are not. The F ratios shown below in the table of Tests of Equality of Group Means (Appendix 2.1), shows fifty variables that significantly vary between the two groups at a 10% level of significance. Of these, drinking in general, use of tobacco products and the frequency of drinking alcohol were the most important independent variables to discriminate the functions.

Referring to Appendix 2.2 the Eigenvalue of 69.997 is responsible for 100% of the explained variance and how well the discriminant function differentiates the group. In this case, the discriminant function is a good fit for the data. The Canonical Correlation 0.993, the square root (0.993^2 = 98.6%) means that 98.6% of the variance is explained by group differences (Appendix 2.2). The Wilks' Lambda score of 0.014 with a p value = 0.00 (64 degrees of freedom) indicates that 1.4% of the total variance is not explained between the two groups (Appendix 2.3).

Table-2: Discriminant Analysis Classification Table

Classification					
		Alcohol	Predicted Group Membership		Total
			Yes	No	
Original	Count	Yes	900	238	1138
		No	2	1901	1903
	%	Yes	79.1	20.9	100.0
		No	.1	99.9	100.0
Cross-validated[b]	Count	Yes	790	348	1138
		No	724	1179	1903
	%	Yes	69.4	30.6	100.0
		No	38.0	62.0	100.0
a. 92.1% of original grouped cases correctly classified.					
b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.					
c. 64.7% of cross-validated grouped cases correctly classified.					

The Standardized Canonical Discriminant Function table (Appendix 2.4) indicated that the two predictors are the following: friends taking drugs and alcohol in a social setting; and stress from study. These two factors contribute most in determining alcohol consumption in adolescents. The Structure Matrix (Appendix 2.5) has revealed that the frequency of alcohol and tobacco consumption are highly correlated with the discriminant function. The Functions at Group Centroids Table (Appendix 2.6) addresses how the two groups differ, the greater the difference between these values the less error there is in classification. The results reveal a high difference between groups making these classifications accurate.

The performance of the discriminant function is illustrated in the below Classification Results table 2. It indicates that 92.1% of original cases and 64.7% of cross-validated grouped cases were correctly classified.

Artificial Neural Networks

Artificial Neural Network analysis was performed on the data set using the Multilayer Perceptron to synthesize a predictive model. The Case Processing summary (Appendix 3.1) showed that 1361 cases were assigned to the training sample and 585 were allocated to the testing sample. The most important independent variables in dictating adolescent alcohol use as shown in the Independent Variable Importance table (Appendix 3.2) are frequency of alcohol consumption and tobacco use with gender being considered least important.

As shown in the Classification Table 3 below, 100% of those adolescents not consuming alcohol were classified correctly. In contrast 98.6% (544 of 552) of cases were classified correctly for those consuming alcohol. As this model classifies more than 95% of the cases correctly it is considered a good model.

The training model has a propensity to inflate the classification rate and therefore the testing sample is used provide clarity. The results show that 98.7% sensitivity by correctly classifying 220 out of 223 adolescent participants as alcohol consumers. Of the adolescents that did not consume alcohol 360 out of 362 were classified correctly with 99.4% sensitivity. As a result, based on the testing sample 99.1% of cases were classified correctly, indicating that this is a good model.

Table-3: Artificial Neural Network Classification Table

Classification				
Sample	Observed	Predicted		
		Yes	No	Percent Correct
Training	Yes	544	8	98.6%
	No	0	809	100.0%
	Overall Percent	40.0%	60.0%	99.4%
Testing	Yes	220	3	98.7%
	No	2	360	99.4%
	Overall Percent	37.9%	62.1%	99.1%
Dependent Variable: Alcohol				

Decision Trees

CART and CHAID were used as the growing methods to build the Decision Tree model. Sixty-seven independent variables were assigned for CART; however the pruning process refined the model to 5 significant independent variables (Figure 3) that influence alcohol consumption in descending order: frequency of alcohol, illicit drug use, legal medicinal drug use, frequency of tobacco use and peers taking drugs for fun. Below is a graphical representation (Figure 1) of the tree model which further supports current literature that adolescent alcohol use is a multifactorial issue that has several associated predictor variables.

The first decision node describes that if the frequency of alcohol use is less than 7, there is a 100% chance that the patient will not consume alcohol. If the frequency of alcohol use is greater than 7, there is a 97.5% probability of adolescents consuming alcohol and a 2.5% chance that participant will not engage in alcohol consumption. The remaining nodes represent the other significant variables in sequential order and describe the probability of alcohol consumption in adolescents.

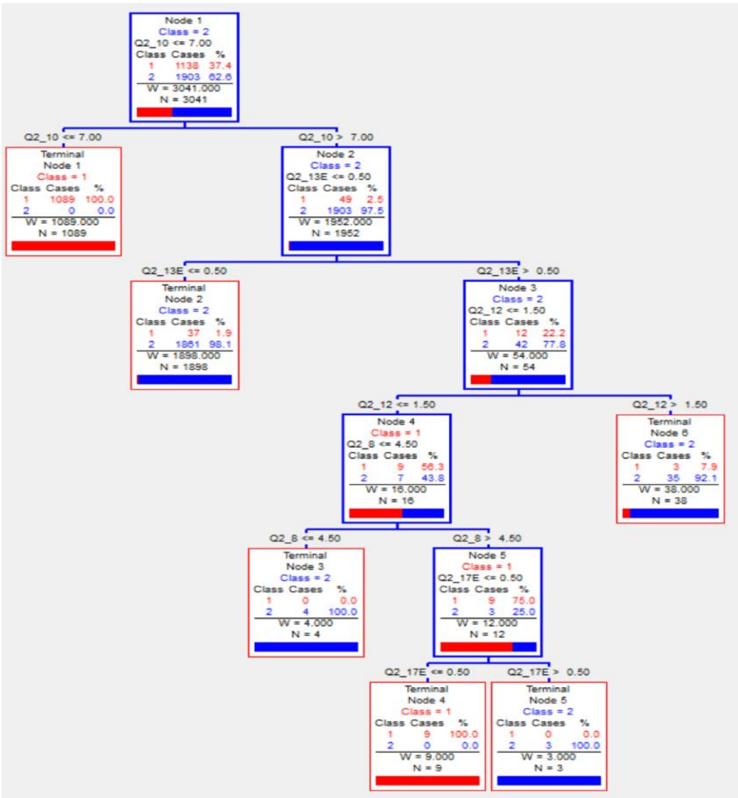


Fig-1: Decision Tree Using CART

The identified associated risk factors from the DT model largely reflect current literature findings on adolescent alcohol consumption. As shown in Figure 2 the model achieved a 72.52% specificity and 96.84% sensitivity with an overall classification of 81.62%. As a result, the DT model is a valuable application in predicting risk factors associated with adolescent alcohol consumption.

Actual Class	Total Class	Percent Correct	Predicted Classes	
			1 N = 1625	2 N = 1416
1	1,138	96.84%	**1,102**	36
2	1,903	72.52%	523	**1,380**
Total:	3,041			
Average:		84.68%		
Overall % Correct:		81.62%		
Specificity		72.52%		
Sensitivity/Recall		96.84%		
Precision		67.82%		
F1 statistic		79.77%		

Fig-2: Decision Tree Classification Table

Variable	Score
Q2_10	100.00
Q2_8	20.55
Q2_7	20.54
Q2_2A	18.42
Q2_14	15.76
Q2_13G	14.93
Q2_12	1.21
Q2_13E	0.74
Q1_4	0.30

Fig-3: Decision Tree Variable Importance

Random Forest

As the Random Forest model is an extension of the DT process it was built using CART as its growing method. All independent 67 variables were assigned for CART, however only 12 remained post pruning. The significant variables included of the following: Frequency of alcohol use, Frequency of tobacco use , Exposure to alcohol at parties, The use of illicit drugs, The use of tobacco products, Exposure to pornographic material, Unknown sources of viewing pornographic material, Friends consuming alcohol, CD/DVD/Video as the source of viewing pornographic material, Party and picnic with friends, Gender, Taking illicit drugs for fun.

The Variable Importance figure (Figure 4) below shows these significant variables in descending order. The model achieved 99.79% specificity and 96.13% sensitivity with an overall classification of 98.42% (Figure 5). This expansion from the DT method has identified 7 more significant variables without compromising accuracy. As the Random Forest model has the capabilities to accommodate large input data, it is a useful application for this large data set and is valuable in predicting risk factors associated with adolescent alcohol consumption.

Variable	Score
Q2_10	100.00
Q2_8	11.17
Q2_2A	9.67
Q2_13G	6.90
Q2_7	6.83
Q2_14	6.72
Q2_5	5.63
Q2_6D	4.90
Q2_11	3.98
Q2_6A	3.28
Q2_1	3.09
Q1_1	3.09

Fig-4: Variable Importance Random Forest

Actual Class	Total Class	Percent Correct	Predicted Classes	
			1 N = 1098	2 N = 1943
1	1,138	96.13%	**1,094**	44
2	1,903	99.79%	4	**1,899**
Total:	3,041			
Average:		97.96%		
Overall % Correct:		98.42%		
Specificity		99.79%		
Sensitivity/Recall		96.13%		
Precision		99.64%		
F1 statistic		97.85%		

Fig-5: Random Forest Classification Table

Stochastic Gradient Boosting (Using TreeNet)

As the Stochastic Gradient Boosting model using TreeNet is an advancement of the DT process, CART was still used as its growing method. All independent 67 variables were assigned, however only 10 remained post pruning (Figure 6). The following significant variables included as shown in the Variable Importance figure below:

- -Frequency of alcohol use
- -The use of legal medicinal drugs
- Age
- -Breakups with boy/girlfriend as the rational for friends taking drugs
- -The use of illicit drugs for fun
- -Household/parents monthly income
- -Viewing pornographic material
- -Government or private schooling education
- -Leisure time spent with friends
- -Viewing of pornographic material through internet/mobile

The model demonstrates 99.89% specificity and 96.31% sensitivity with an overall classification of 98.55% (Figure 7). This application is more accurate than the DT method and has identified 5 more significant variables that contribute to adolescent alcohol consumption. The accuracy of these results is due to the capacity to handle large data sets without over fitting.

Variable	Score
Q2_10	100.00
Q2_12	6.38
Q1_4	6.37
Q2_17A	6.08
Q2_13E	5.85
Q1_15	5.32
Q2_5	4.37
Q1_8	4.10
Q2_3D	3.75
Q2_6B	3.70

Fig-6: Variable Importance Stochastic Gradient Boosting

Actual Class	Total Class	Percent Correct	Predicted Classes	
			1 N = 1098	2 N = 1943
1	1,138	96.31%	**1,096**	42
2	1,903	99.89%	2	**1,901**
Total:	3,041			
Average:		98.10%		
Overall % Correct:		98.55%		
Specificity		99.89%		
Sensitivity/Recall		96.31%		
Precision		99.82%		
F1 statistic		98.03%		

Fig-7: Classification Table Stochastic Gradient Boosting

Appendices

Appendix 1 – Results and Interpretations for Logistic Regression Model

Appendix 1.1

Variables in the Equation						
	B	S.E.	Wald	df	Sig.	Exp(B)
Sex(1)	-32.207	2731.551	.000	1	.991	.000
Martial Status(1)	-8.400	16153.245	.000	1	1.000	.000
Area(1)	-15.960	1969.108	.000	1	.994	.000
Age in Months	-.697	81.798	.000	1	.002	.498
Religion			.000	4	1.000	
Religion(1)	25.291	289391.241	.000	1	1.000	96305888750.000
Religion(2)	9.471	4794.892	.000	1	.998	12976.473
Religion(3)	7.581	13248.780	.000	1	1.000	1960.522
Religion(4)	122.561	22933.865	.000	1	.996	.000
Standard of Studying			.001	4	1.000	
Standard of Studying(1)	-77.162	16447.079	.000	1	.996	.000
Standard of Studying(2)	-51.284	7896.884	.000	1	.995	.000
Standard of Studying(3)	-49.314	6232.879	.000	1	.994	.000
Standard of Studying(4)	47.850	1975.651	.001	1	.981	604097591700000000000.000

Subject Stream			.000	2	1.000	
Subject Stream(1)	-93.948	7504.221	.000	1	.990	.000
Subject Stream(2)	-69.747	5924.717	.000	1	.991	.000
Type of School/College			.000	2	1.000	
Type of School/College(1)	-10.900	4640.885	.000	1	.998	.000
Type of School/College(2)	.682	4774.741	.000	1	1.000	1.978
Type of School/ College(1)	-10.224	6502.765	.000	1	.999	.000
Education Medium(1)	3.825	2162.510	.000	1	.999	45.849
Working Part Time(1)	13.062	8165.673	.000	1	.999	470842.758
Part-Time Earning	.014	2.485	.000	1	.996	1.014
Type of Family			.000	2	1.000	
Type of Family(1)	-9.868	4971.460	.000	1	.998	.000
Type of Family(2)	-40.325	4024.177	.000	1	.992	.000
Living with Parents			.000	2	1.000	
Living with Parents(1)	10.910	4920.196	.000	1	.998	54700.818
Living with Parents(2)	-11.993	5062.069	.000	1	.998	.000
Monthly Income	.000	.056	.000	1	.993	1.000
Party/ Picnic(1)	-4.540	1580.889	.000	1	.998	.011
Drink(1)	14.585	1998.656	.000	1	.994	2157635.236
Puffing (1)	-108.753	6397.786	.000	1	.986	.000
Drugs(1)	67.611	13111.673	.000	1	.996	230753738500000000000000000000.000
Other intoxication(1)	-1.883	1902.017	.000	1	.999	.152
Sport(1)	12.192	1519.714	.000	1	.994	197157.768
Listening Music(1)	51.912	2279.293	.001	1	.982	35078745320000000000000.000
Reading Novel, Megazine(1)	-33.198	3479.176	.000	1	.992	.000
Hanging out(1)	-2.573	4079.733	.000	1	.999	.076
Watching Movie(1)	-10.971	1682.780	.000	1	.995	.000
Any other (specify)			.001	2	1.000	
Any other (specify)(1)	111.601	40496.539	.000	1	.998	2.936E+48
Any other (specify)(2)	70.691	40457.003	.000	1	.999	501926122200000000000000000000.000
No Specific Activity(1)	23.116	6643.386	.000	1	.997	10944379260.000
Watch Pornographic Movies/ Video(1)	-61.887	76248.209	.000	1	.999	.000
Watching with Whom			.000	5	1.000	
Watching with Whom(1)	3.960	92098.295	.000	1	1.000	52.473
Watching with Whom(2)	72.243	94314.743	.000	1	.999	2370781616000000000000000000000.000
Watching with Whom(3)	38.057	92074.706	.000	1	1.000	33729678960000000.000
Watching with Whom(4)	-2.456	91893.514	.000	1	1.000	.086

Watching with Whom(5)	11.255	92251.435	.000	1	1.000	77253.661
CD/DVD/VIDEO			.000	2	1.000	
CD/DVD/VIDEO(1)	-213.555	41776.231	.000	1	.996	.000
CD/DVD/VIDEO(2)	-180.200	42063.531	.000	1	.997	.000
Internet/ Mobile			.000	1	.986	
Internet/ Mobile(1)	-35.078	2028.541	.000	1	.986	.000
TV			.000	2	1.000	
TV(1)	221.824	64082.954	.000	1	.997	2.171E+96
TV(2)	224.505	64870.873	.000	1	.997	3.171E+97
Magazine			.000	1	.998	
Magazine(1)	-16.388	6608.856	.000	1	.998	.000
Others			.000	1	.999	
Others(1)	9.463	6884.367	.000	1	.999	12869.515
NA			.000	1	.997	
NA(1)	-84.264	26628.562	.000	1	.997	.000
Taking Tobacco Products			.000	2	1.000	
Taking Tobacco Products(1)	50.312	25413.919	.000	1	.998	708154647100000000000.000
Taking Tobacco Products(2)	136.590	25618.029	.000	1	.996	2.090E+59
Frequency of Tobacco			.001	5	0.998	
Frequency of Tobacco(1)	106.646	4269.497	.001	1	.980	2.069E+46
Frequency of Tobacco(2)	182.865	6368.670	.001	1	.977	2.613E+79
Frequency of Tobacco(3)	60.266	3769.531	.000	1	.987	14894782610000000000000000.000
Frequency of Tobacco(4)	86.961	3954.993	.000	1	.982	5.844E+37
Frequency of Tobacco(5)	73.981	3753.240	.000	1	.984	1347211691000000200000000000000 00.000
Frequency of Alcohol			.006	6	0.982	
Frequency of Alcohol(1)	-266.196	9776.702	.001	1	.978	.000
Frequency of Alcohol(2)	-285.385	4156.003	.005	1	.945	.000
Frequency of Alcohol(3)	-305.663	6024.837	.003	1	.960	.000
Frequency of Alcohol(4)	-277.088	10608.758	.001	1	.979	.000
Frequency of Alcohol(5)	-254.520	6818.603	.001	1	.970	.000
Frequency of Alcohol(6)	-243.823	10995.396	.000	1	.982	.000
Drugs- SP Relipen etc(1)	-35.850	4519.813	.000	1	.994	.000
Drugs- Brown sugar, Cocain, heroin(1)	-33.580	3726.303	.000	1	.993	.000
Breaking up(1)	93.525	57462.226	.000	1	.999	4.144E+40
Stress of study(1)	63.975	5222.892	.000	1	.990	60805895200000000000000000.000
Friends (1)	8.940	4946.175	.000	1	.999	7632.450
Parents (1)	-80.961	9079.048	.000	1	.993	.000

For Fun(1)	9.835	5067.407	.000	1	.998	18670.734
Friends taking Alcohol(1)	-.299	2551.483	.000	1	0.005	.742
Friends taking Drugs(1)	38.747	20035.875	.000	1	.998	67230787240000000.000
Breaking up			.000	2	1.000	
Breaking up(2)	91.847	57423.737	.000	1	.999	7.742E+39
Stress of Study			.000	1	.997	
Stress of Study(1)	-58.610	18565.938	.000	1	.997	.000
Friends taking			.000	1	.986	
Friends taking(1)	-43.278	2496.026	.000	1	.986	.000
Parents separated			.000	1	.999	
Parents separated(1)	-4.236	3621.632	.000	1	.999	.014
For Fun			.000	1	.998	
No Idea			.000	1	.997	
No Idea(1)	-19.769	4620.209	.000	1	.997	.000
Injectable			.000	2	1.000	
Injectable(1)	32.875	44257.530	.000	1	.999	189330218300000.000
Injectable(2)	.678	44532.385	.000	1	1.000	1.970
Puffs			.000	1	.999	
Puffs(1)	3.184	3229.334	.000	1	.999	24.142
Oral			.000	1	.987	
Oral(1)	65.631	3957.029	.000	1	.987	318417352100000000000000000.000
Not Known			.000	1	.989	
Not Known(1)	67.061	4675.101	.000	1	.989	133136953100000000000000000.000
Constant	213.329	52810.380	.000	1	.997	4.442E+92

a. Variable(s) entered on step 1: Sex, Martial Status, Area, Age in Months, Religion, Standard of Studying, Subject Stream, Type of School/College, Type of School/ College, Education Medium, Working Part Time, Part-Time Earning, Type of Family, Living with Parents, Monthly Income, Party/ Picnic, Drink, Puffing , Drugs, Other intoxication, Sport, Listening Music, Reading Novel, Megazine, Hanging out, Watching Movie, Any other (specify), No Specific Activity, Watch Pornographic Movies/ Video, Watching with Whom, CD/DVD/VIDEO, Internet/ Mobile, TV, Magazine, Others, NA, Taking Tobacco Products, Frequency of Tobacco, Frequency of Alcohol, Drugs- SP Relipen etc, Drugs- Brown sugar, Cocain, heroin, Breaking up, Stress of study, Friends , Parents , For Fun, Others, NA, Friends taking Alcohol, Friends taking Drugs, Breaking up, Stress of Study, Friends taking, Parents separated, For Fun, Others, No Idea, NA, Injectable, Puffs, Oral, Others, Not Known.

Appendix 1.2

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	2211.901	91	.000
	Block	2211.901	91	.000
	Model	2211.901	91	.000

Appendix 1.3

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	4	1.000

ppendix 1.4

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.743	1.000
a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.			

Appendix 2 – Results and Interpretations for Discriminant Analysis

Appendix 2.1

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
Sex	.944	95.800	1	1624	.000
Martial Status	1.000	.429	1	1624	.513
Area	.998	2.735	1	1624	.098
Age in Months	.983	27.949	1	1624	.000
Religion	.999	2.249	1	1624	.134
Standard of Studying	.997	5.404	1	1624	.020
Subject Stream	1.000	.000	1	1624	.991
Type of School/College	.980	33.587	1	1624	.000
Type of School/ College	1.000	.582	1	1624	.445
Education Medium	.991	15.320	1	1624	.000
Working Part Time	.997	5.554	1	1624	.019
Part-Time Earning	.999	.841	1	1624	.359
Type of Family	.999	1.569	1	1624	.211
Living with Parents	.999	1.885	1	1624	.170
Monthly Income	1.000	.075	1	1624	.785
Party/ Picnic	.919	142.780	1	1624	.000
Drink	.841	306.770	1	1624	.000
Puffing	.982	30.352	1	1624	.000
Drugs	.978	36.492	1	1624	.000
Other intoxication	.970	49.862	1	1624	.000
Sport	.981	30.783	1	1624	.000
Listening Music	.999	1.902	1	1624	.168
Reading Novel, Megazine	.995	8.888	1	1624	.003
Hanging out	.982	29.150	1	1624	.000
Watching Movie	.999	1.321	1	1624	.251
Any other (specify)	1.000	.734	1	1624	.392
No Specific Activity	1.000	.125	1	1624	.723
Watch Pornographic Movies/ Video	.920	140.970	1	1624	.000
Watching with Whom	.955	75.672	1	1624	.000
CD/DVD/VIDEO	.927	128.486	1	1624	.000
Internet/ Mobile	.926	129.749	1	1624	.000
TV	.921	139.140	1	1624	.000
Magazine	.921	139.913	1	1624	.000
Others	.922	137.020	1	1624	.000
NA	.920	141.798	1	1624	.000
Taking Tobacco Products	.835	320.871	1	1624	.000
Frequency of Tobacco	.867	248.723	1	1624	.000
Frequency of Alcohol	.071	21128.463	1	1624	.000

Drugs- SP Relipen etc	.892	196.206	1	1624	.000
Drugs- Brown sugar, Cocain, heroin	.971	49.240	1	1624	.000
Breaking up	.983	27.965	1	1624	.000
Stress of study	.994	10.052	1	1624	.002
Friends	.963	62.697	1	1624	.000
Parents	.998	3.334	1	1624	.068
For Fun	.885	211.358	1	1624	.000
Others	.997	4.997	1	1624	.026
NA	.861	261.309	1	1624	.000
Friends taking Alcohol	.873	236.596	1	1624	.000
Friends taking Drugs	.988	19.518	1	1624	.000
Breaking up	.989	18.498	1	1624	.000
Stress of Study	.988	19.010	1	1624	.000
Friends taking	.989	18.866	1	1624	.000
Parents separated	.988	18.986	1	1624	.000
For Fun	.989	18.369	1	1624	.000
Others	.989	18.865	1	1624	.000
No Idea	.989	18.631	1	1624	.000
NA	.989	18.822	1	1624	.000
Injectable	.989	17.907	1	1624	.000
Puffs	.989	18.478	1	1624	.000
Oral	.989	18.714	1	1624	.000
Others	.988	19.222	1	1624	.000
Not Known	.988	20.080	1	1624	.000

Appendix 2.2

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	68.197[a]	100.0	100.0	.993
a. First 1 canonical discriminant functions were used in the analysis.				

Appendix 2.3

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.014	6745.230	64	.000

Appendix 2.4

Standardized Canonical Discriminant Function Coefficients	
	Function
	1
Sex	.005
Martial Status	.006
Area	-.023
Age in Months	-.050
Religion	-.001
Standard of Studying	.020
Subject Stream	.011
Type of School/College	.016
Type of School/ College	.014
Education Medium	-.013

Working Part Time	.028
Part-Time Earning	.236
Type of Family	.000
Living with Parents	-.007
Monthly Income	.006
Party/ Picnic	-.033
Drink	-.015
Puffing	.037
Drugs	-.014
Other intoxication	-.086
Sport	-.015
Listening Music	-.013
Reading Novel, Megazine	.020
Hanging out	-.025
Watching Movie	.028
Any other (specify)	-.001
No Specific Activity	-.035
Watch Pornographic Movies/ Video	.013
Watching with Whom	-.013
CD/DVD/VIDEO	-.138
Internet/ Mobile	.281
TV	.212
Magazine	.541
Others	-1.440
NA	.530
Taking Tobacco Products	-.005
Frequency of Tobacco	.002
Frequency of Alcohol	-.068
Drugs- SP Relipen etc	-.046
Drugs- Brown sugar, Cocain, heroin	-.012
Breaking up	.032
Stress of study	.038
Friends	-.046
Parents	.015
For Fun	.003
Others	-.055
NA	.021
Friends taking Alcohol	.009
Friends taking Drugs	.052
Breaking up	-.810
Stress of Study	.694
Friends taking	.465
Parents separated	-1.254
For Fun	.063
Injectable	.118
Puffs	.360
Oral	-.023
Others	.371
Predicted probability	.054
Predicted Value for Q2_9	.699
Predicted Pseudo-Probability for Q2_9 = 1	.023
Predicted Value for Q2_9	-.016
Predicted Pseudo-	-.336

Probability for Q2_9 = 1	
Predicted Probability for Q2_9=1	-.147

Appendix 2.5

Structure Matrix	
	Function
	1
Predicted Value for Q2_9	.904
Predicted Value[a]	.904
Predicted Probability for Q2_9=1	-.893
Predicted Probability for Q2_9=2[a]	.893
Predicted Pseudo-Probability for Q2_9 = 2[a]	.869
Predicted Pseudo-Probability for Q2_9 = 1	-.869
Predicted Pseudo-Probability for Q2_9 = 1	-.770
Predicted Pseudo-Probability for Q2_9 = 2[a]	.770
Predicted Value for Q2_9	.754
Predicted probability	.690
Frequency of Alcohol	.437
Taking Tobacco Products	.054
Drink	-.053
NA	.049
Frequency of Tobacco	.047
Friends taking Alcohol	.046
For Fun	-.044
Drugs- SP Relipen etc	.042
Party/ Picnic	.036
NA	.036
Watch Pornographic Movies/ Video	.036
Magazine	.036
TV	.035
Others	.035
Internet/ Mobile	.034
CD/DVD/VIDEO	.034
Sex	.029
Watching with Whom	.026
Friends	-.024
Other intoxication	-.021
Drugs- Brown sugar, Cocain, heroin	.021
Drugs	-.018
Not Known[a]	.017
Type of School/College	.017
Sport	-.017
Puffing	-.017
Hanging out	-.016
Breaking up	-.016
Age in Months	-.016

No Idea[a]	.014
NA[a]	.013
Friends taking Drugs	.013
Others	.013
Stress of Study	.013
Parents separated	.013
Friends taking	.013
Oral	.013
Breaking up	.013
Puffs	.013
For Fun	.013
Injectable	.013
Others[a]	.013
Education Medium	-.012
Stress of study	-.010
Reading Novel, Megazine	.009
Working Part Time	.007
Standard of Studying	-.007
Others	-.007
Parents	-.005
Area	.005
Religion	.005
Listening Music	-.004
Living with Parents	.004
Type of Family	.004
Watching Movie	.003
Part-Time Earning	-.003
Any other (specify)	-.003
Type of School/ College	-.002
Martial Status	.002
No Specific Activity	-.001
Monthly Income	.001
Subject Stream	.000
Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.	
a. This variable not used in the analysis.	

Appendix 2.6

Functions at Group Centroids	
Alcohol	Function
	1
Yes	-9.771
No	6.971
Unstandardized canonical discriminant functions evaluated at group means	

Appendix 3 – Results and Interpretations for Artificial Neural Network

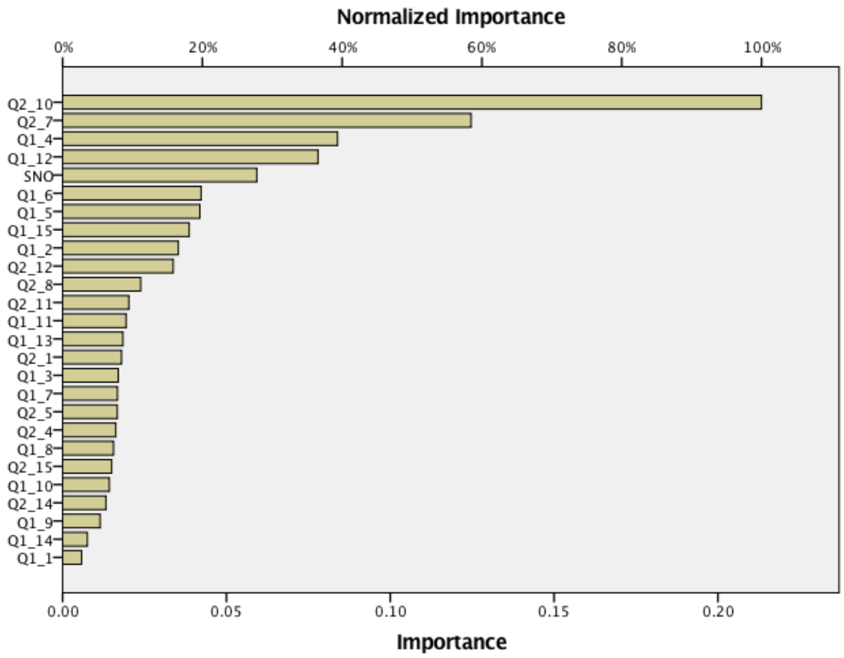
Appendix 3.1

Case Processing Summary			
		N	Percent
Sample	Training	1361	69.9%
	Testing	585	30.1%
Valid		1946	100.0%
Excluded		1095	
Total		3041	

Appendix 3.2

Independent Variable Importance		
	Importance	Normalized Importance
Sex	.006	2.7%
Martial Status	.035	16.5%
Area	.017	8.0%
Religion	.042	19.6%
Standard of Studying	.042	19.8%
Subject Stream	.017	7.8%
Type of School/College	.016	7.3%
Type of School/ College	.011	5.4%
Education Medium	.014	6.7%
Working Part Time	.019	9.1%
Type of Family	.018	8.6%
Living with Parents	.008	3.5%
Age in Months	.084	39.3%
Part-Time Earning	.078	36.5%
Monthly Income	.039	18.1%
SNO	.059	27.8%
Party/ Picnic	.018	8.4%
Watch Pornographic Movies/ Video	.016	7.6%
Watching with Whom	.017	7.8%
Taking Tobacco Products	.125	58.5%
Frequency of Tobacco	.024	11.2%
Frequency of Alcohol	.213	100.0%
Drugs- SP Relipen etc	.020	9.5%
Drugs- Brown sugar, Cocain, heroin	.034	15.8%
Friends taking Alcohol	.013	6.2%
Friends taking Drugs	.015	7.0%

Appendix 3.3



DISCUSSION

Risk factors associated with adolescent alcohol consumption are complex in nature. Despite this complexity using recursive techniques has revealed useful risk factors associated with adolescent alcohol use. This study composed of a dataset of 67 independent variables and by using various statistical modelling techniques it was revealed that 8 of these were significant risk factors associated with adolescent alcohol use. In comparison to traditional univariate and multivariate analytical models which is used in literature, the cutting recursive methods delivered superior modelling results.

Comparison of Classification Rates

This report applied 6 modelling techniques to a subset of the ARSH data set: Logistic Regression (LR), Discriminant Analysis (DA), Artificial Neural Networks (ANN), Decision Tree (DT), Random Forest (RF) and the Stochastic Gradient Boosting method.

Table-4: Classification Accuracy

Classification Accuracy		
Model	**Training**	**Testing**
Logistic Regression	92.5%.	92.5%%
Discriminant Analysis	92.10%	92.10%
Artificial Neural Network	99.40%	99.10%
Decision Tree Analysis	81.62%	81.62%
Random Forest	98.42%	98.42%
Stochastic Gradient Boosting	98.55%	98.55%

The above classification accuracy table (Table 4) shows that the ANNs gives highest accuracy with followed by SGB. However ANN has excluded quite a few observations and also depends on random seed. Therefore, accounting for these statistical errors it is concluded that Stochastic Gradient Boosting provided the best predicted accuracy of risk factors contributing to adolescent alcohol consumption. Nevertheless, each of these predictive models contains its own parameters and the classification accuracy depends on these. Each model is advantageous as each can be optimised with further statistical trials to develop ideal parameters.

Comparison of Significant Independent Variables

The aim of these models was to accurately derive associated risk factors that contribute to adolescent alcohol use. Accuracy was confounded due to the disparity between the nature of the ARSH dataset designed for adolescent reproductive sexual health, and the research for this paper – adolescent alcohol consumption. Logistic Regression and Discriminant

Analysis give statistically significant variables whereas non-parametric methods like ANN, Decision Tree, Random Forest and SGB just give variable importance analysis. From the analysis of these six different models we have identified eight significant variables which are common to at least one or more algorithms. For example Frequency of alcohol was found important by five models followed by frequency of tobacco use etc. These variables were consistent across both parametric and non-parametric methods discussed in the paper. The other variables consistent across different models were illicit drug use, legal medicinal drug use, peers taking drugs for fun etc. as shown in Table 5. It can be concluded that the important independent variables that emerged are consistent with literature.

Table-5: Comparison of Significant Independent Variables

Independent Variables	LR	DA	ANN	DT	RF	SGB
Frequency of Alcohol		X	X	X	X	X
Frequency of Tobacco Use		X	X	X	X	
Illicit Drug Use		X		X	X	
Legal Medicinal Drug Use		X		X	X	X
Peers Taking Drugs for Fun		X		X	X	X
Exposure of Alcohol at Parties		X			X	X
Exposure to Pornographic Material		X			X	X
Friends Consuming Alcohol	X	X			X	

CONCLUSION AND RECOMMENDATIONS

There has been an emerging need to reduce the prevalence of adolescent alcohol consumption in India. Studies have shown that psychosocial factors, such as those significant independent variables identified in this report contribute to the ongoing issue of adolescent alcohol use. The recursive techniques addressed in this article are becoming useful predictive instruments not only in the context of alcohol misuse; however, for other socio-health problems such as drug abuse, adolescent sex behaviour and burden of disease. Identifying associated risk factors for adolescent alcohol consumption provides information to develop interventional programs and frameworks to potentially change legislative policy surrounding adolescent alcohol consumption.

ACKNOWLEDGEMENTS

Authors are highly grateful to the both reviewers and Editor for helpful comments on the earlier version of the paper and pointing out few mistakes which we overlooked in the original draft.

REFERENCES

1. National Institute of Alcohol Abuse and Alcoholism. Alcohol Altert. 2000.
2. Gopiram P, Kishore MT. Psychosocial attributes of substance abuse among adolescents and young adults: A comparative study of users and non-users. Indian journal of psychological medicine. 2014 Jan;36(1):58.
3. Tiwari VK, Piang LL, TP SR, Nair KS. Correlates of Social, Demographic and Behavioral Factors affecting Adolescent Sexuality in a Traditional Society in India: Perspectives and Challenges. Indian Journal of Youth & Adolescent Health. 2015 Oct 14;2(3):44-57.
4. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology. 1996 Nov 1;49(11):1225-31.
5. Gepp A, Kumar K, Bhattacharya S. Business failure prediction using decision trees. Journal of forecasting. 2010 Sep 1; 29(6):536-55.
6. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.
7. Salford Systems. TreeNet. 2016.