Indian Journal of Science and Technology, Vol 9(32), DOI: 10.17485/ijst/2016/v9i32/100716, August 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Alcoholic Behavior Prediction through Comparative Analysis of J48 and Random Tree Classification Algorithms using WEKA

Navdeep Kaur* and Williamjeet Singh

Punjabi University Patiala, Punjab, India; 24gill@gmail.com, williamjeet@gmail.com

Abstract

Objectives/Background: Addiction of alcohol is a complex disease which results from diversity of social, genetic and environmental influences. A report by World Health Organization, WHO (2014) estimates that most of the deaths are from alcohol related causes.The objective of this study is to analyze the alcoholic behavior of different age group people on the basis of risk factors. In this paper, we construct a comparative model of different classification techniques to analyze the best algorithm for predicting the alcoholic behavior of a person. **Methods:** Under this context, random tree and J48 that are decision tree algorithms have been exercised on the dataset of 600 people that is collected through a structured questionnaire by visiting de addicted centers, colleges, villages, government offices, old age homes of Patiala, Punjab. **Findings:** Results conclude that the random tree provides more precise results than J48 for all the age group people. Risk factors that come out to be most effective are impulsive nature, sensation seeking nature, financial loss, family conflict, depression, child abuse, alcoholic shop near home distance.The overall accuracy of random tree is 75.94% and for J48 is 71.26%. **Applications/Improvement:** There is a need to develop some intelligent tools in this area and the rules extracted from this analysis can be further used for designing the tool. More attributescan be incorporated to achieve the optimal results for predicting the behavior of an alcoholic person.

Keywords: Addiction, Classification, Data Mining, Prediction

1. Introduction

Alcohol consumption by the human beings has a great impact on their lives. Consumption of alcohol is a social prohibition in the most regions of India. Most of the societies encounter the extreme challenge of alcohol consumption, which is usually related with the social problems. In[1] most of the youth set up life-long model of alcohol utilize during the age of rising adulthood. It is closely associated to historical, social, cultural, religious, economic and environmental aspects of a society and is simulated by various factors such as family, quantitative, physical, medical and environmental ones. In[2,3] early use of alcohol increases the chances of alcohol abuse. In[4] report states that 38.3% of the world's population consumed alcohol regularly. In[5] report estimates that alcoholism increased by about 55% between 1992 and 2012. Predicting alcoholic behavior of a person becomes a crucial task to overcome the problem of alcohol consumption.In this paper, two different techniques of data mining are used to predict whether the person is alcoholic or not on the basis of risk factors that prestige the people to take alcohol. This paper explores that the cause of alcoholism is different for every addict person and predicts the alcoholic behavior of a person by considering the main risk factors that cause humans to drink alcohol. In[6] data mining provides the two models for storing large data stores in databases. It can be either predictive or descriptive. In[7] predictive data mining is used to build the predictions based on stored data. It can be further categorized into: Classification, Neural Networks, Decision Tree, etc. descriptive data mining deals with the general character-

*Author for correspondence

istics of current data. This can be classified into: Feature Extraction, Clustering, Association Rule Mining, etc.The main goal of this paper is to analyze the most important risk factors that indulge the people to take alcohol. These factors help to differentiate the alcoholic and non alcoholic person by using data mining technique. Two different algorithms of decision tree that are: random tree and J48 are used to construct the model and the motive of this study is to compare the different models and estimating the influence of various risk factors on alcoholic person.This paper is organized as follows: Section 2 highlights the prior work followed by section 3 which explains the experimental settings. Section 4 confers the result and section 5 discusses the conclusion.

2. Literature Review

In past years, researchers have paid a vast deliberation at determining the various factors that indulge humans to take drugs. In[8] considered 1023 students to evaluate the various risk factors that affect the addictive behavior of youth. In[8] found that the peer pressure, antisocial behavior, parental monitoring turns the youth into drink. In[9] considered 1025 teenagers to predict the young adulthood non smoker and smoker. Findings conclude that the unmarried in adulthood, less education, lower family support cause the person to smoke. In[10] estimate the response rate for doing survey in concern with the health of alcoholic person. Result shows that the respond rate of men, young ones and the people in the deprived areas is less than the women who live in those areas. In[11] evaluate the significant protective mediation which helps the parents to weaken the exposure of alcohol use issues by their children. In[11] conclude that the web based preventive intervention for parents has a big prospective as a family friendly component in the scale of involvement which is required to gear the problem of alcohol misuse over the society. In[12] identify the parenting strategies associated with adolescent alcohol consumption. Finding provides the factors which are associated with adults that use alcohol are: parental modeling, provision of alcohol, parental monitoring, parent-child relationship, family conflict, parental support. In[13] frankness in communication by parents to their children reduces the chances of alcohol consumption by their children. In[14] predict the alcohol use by the association and suggest that hidden relations with optimistic stimulation plays an important role in alcohol consumption activities, and explain the strength of the IAT-RF as an estimate of inherent alcohol relations. In[15] examine the alcohol utilize commencement and periodic drinking among various students. Risk factors: domestic violence, physical abuse, sexual abuse is the factors for indulging into addiction. In[16] investigate the relationship between different professional stressors and heavy consumption of alcohol among male employees. Results show that professional class, conjugal status, smoking and work load are different factors that have taken to find the relation between occupational stressors and heavy drinking. In[17] build an exploratory model for the different risk factors of alcohol. Results provide that the supposed affiliation with friends, female care giver, and general self-respect are the factors that are associated with physical aggression. In[18] determine the psychological and social influences on rising adult drinking behavior. Finding provides the factors that influence the adults to drink and they are: gender, race/ethnicity, marital status, employment, family influence.

3. Experimental Settings

The main purpose of this approach is to construct a model of classification that codifies the alcoholic and non alcoholic person. Data mining process is used to build the classifiers by linking the steps which include: data understanding, data preparation, modeling and the application of data mining technique that is chosen for the proposed work.

3.1 Data Understanding

The data of different age group people was composed by the means of well-regulated questionnaire by visiting various colleges, government offices, de addicted centers, old age homes, villages etc. Every age group (0-24, 25-40, 41-60, >60) has different questionnaire. A dataset of 600 people was collected having 30 factors which include family factors, physical factors, quantitative factors, environmental and social factors, medical factors, religious factors as shown in Table 1.

3.2 Data Preprocessing

Excel sheets are used to saved the collected data. Data cleaning process is applied to eliminate the missing values in data, analyzing outliers and remove the inconsistent

data. Data consist of all the factors that affect the alcoholic person and the final class on which the result is based consisting of two categories alcoholic and non alcoholic.

3.3 Modeling

WEKA is a data mining tool which is used for the classification technique. It is an open source tool and has integral algorithms that can be used for any type of data set.

3.4 Classification

The mechanism follows the tree like structure which classifies the instances by arranging them in descending order from top node to some bottom node and also classifies the every instance of tree. In[19] every node of tree determines a test of few attributes of the instance and every branch declining from the node equal to one of the probable values for this attribute. J48 is a decision tree algorithm which creates both unpruned and pruned decision trees whereas Random tree generates an unpruned tree that examines N attributes at every node which are chosen randomly. Cross Validation method is chosen to test the dataset because it provides the perfect calculation of error and is relevant to confined dataset.

Table 1. Risk Factors of alcoholic person

Major risk factor	**Sub factor**
Family factors	1. Heredity
	2. Nuclear family
	3. Ethnicity
	4. Conflict in family
	5. More home responsibilities
	1. Work load
	2. Retirement
	3. Participation in recreational activities
Occupation or Physical factors	4. Financial loss
	5. Peer influence
	6. Hostility
	7. Dropping School
	8. Satisfaction with work
	1. Impulsive nature
	2. Introvert nature
Quantitative factors	3. Loneliness
	4. Depression
	5. Sensation Seeking nature
	1.Domestic violence
Environmental and social factors	2. Low neighborhood attachment
	3. Child abuse
	4. Easy affordability
	5. Sexual Assault
	1. Sleeping period
Medical factors	2. Improper diet
	3. Suffer from any disease
	4. Monthly checkup
	1.Less traditional education support
Religious factors	2. Racial
	3. Atheist
	4. Less religious education support

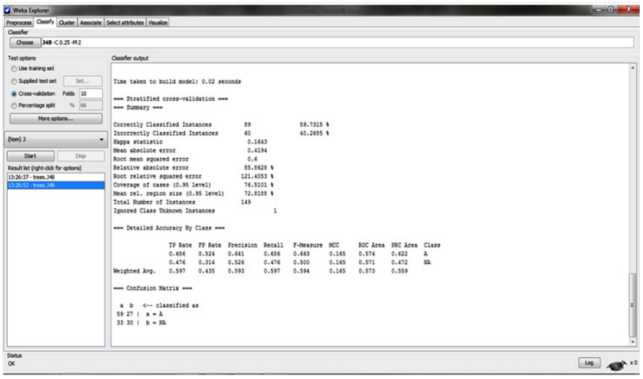

Figure 1. J48 result summary of age group 0-24

4. Result and Discussion

Random tree and J48 were implemented on the dataset of 600 people via 10 fold cross validation method. The summary and rules of age group 0-24 generated by J48 are listed in Figure 1 and Table 2, same for age group 25-40 is shown in Figure 2 and Table 3, for age group 41-60 is shown in Figure 3 and Table 4 and for age group >60 is shown in Figure 4 and Table 5 while the summary of random tree for age group 0-25 is shown in Figure 5, for age group 25-40 is shown in Figure 6, for age group 41-60 is shown in Figure 7 and for age group >60 is shown in

Figure 8. The efficacy of both the algorithms is estimated through three factors that are: precision, True Positive (TP) rate and recall. Recall is the division of significant instances that are recovered. Precision is the division of recovered instances that are significant. In[20] true positive rate is the number of examples predicted positive that are actually positive. If precision is high then it means that the algorithm gives more accurate results and if recall is high then it means that most of the results are relevant that the algorithm returns. The performance comparison of random tree and J48 for all the age groups is shown in the Table 6-9.

Table 2. Rules obtained from J48 for age group 0-24

1. If (not impulsive) and (suffer from disease) and (no sensation seeking nature) and (no child abuse) : Non alcoholic
2. If (not impulsive) and (suffer from disease) and (no sensation seeking nature) and (always go through child abuse) : Alcoholic
3. If (not impulsive) and (sometimes go through child abuse) and (no sensation seeking nature) : Alcoholic
4. If (not impulsive) and (no child abuse) and (no sensation seeking nature) and (not love to do thrilling events) : Non alcoholic
5. If (not impulsive) and (child abuse) : Alcoholic
6. If (impulsive) and (high peer pressure or friends influence) : Alcoholic
7. If (impulsive) and (little peer pressure) and (suffer from disease) : Alcoholic

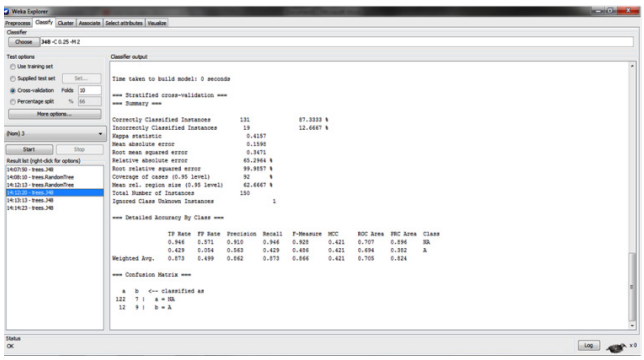

Figure 2. J48 result summary of age group 25-40

Table 3. Rules obtained from J48 for age group 25-40

1. If (no family conflict) and (no depression) : Non alcoholic
2. If (no family conflict) and (occur from trauma) and (never give more effort to do simple task): Non alcoholic
3. If (no family conflict) and (occur from trauma) and (never give more effort to do simple task) and (always bother about things) : Alcoholic
4. If (family conflict) and (never love to live alone) and (no peer pressure) : Non alcoholic
5. If (family conflict) and (never love to live alone) and (highly influenced by friends) : Alcoholic
6. If (family conflict) and (love to live alone) : Alcoholic

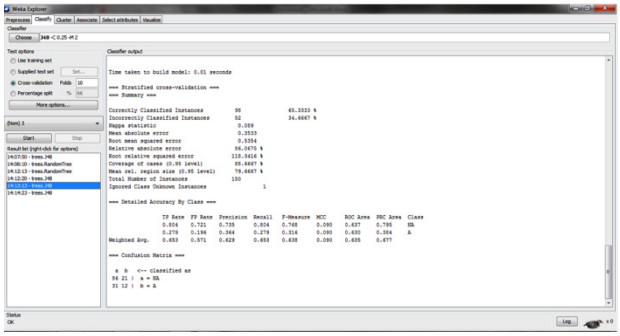

Figure 3. J48 result summary of age group 41-60

Table 4. Rules obtained from J48 for age group 41-60

1. If (not sensation seeking) and (exercise daily) and (not occur from any trauma) and (like trilling events) : Alcoholic
2. If (not sensation seeking) and (exercise daily) and (not occur from trauma) and (not like trilling events): Non alcoholic
3. If (not sensation seeking) and (not exercise daily) and (occur from trauma) : Non alcoholic
4. If (not sensation seeking) and (not exercise daily) and (feel loneliness) : Alcoholic
5. If (not sensation seeking) and (not exercise daily) and (not feel lonely) and (high work load) and (little neighborhood attachment) : Alcoholic
6. If (not sensation seeking) and (not exercise daily) and (not feel lonely) and (high work load) and (more neighborhood attachment) : Non alcoholic
7. If (sensation seeking) and (love to live alone) and (lives in nuclear family) and (suffer from domestic violence) and (no hereditary issues): Alcoholic
8. If (sensation seeking) and (never love to live alone): Non alcoholic
9. If (sensation seeking) and (never love to live alone) and (no family conflict): Non alcoholic
10. If (sensation seeking) and (never love to live alone) and (family conflict) and (lives in nuclear family): Alcoholic

Table 5. Rules obtained from J48 for age group >60

1. If (no financial loss) and (do not like the substances that make him feel high) : non alcoholic
2. If (no financial loss) and (like the substances that make him feel high) and (not feel difficulty in doing quiet tasks) : Alcoholic

3. If (financial loss) and (like its daily routine) : Non alcoholic
4. If (financial loss0 and (like its daily routine) and (never satisfies with its work) : Alcoholic
5. If (financial loss) and (bored from daily routine) and (do not like the substances that make him feel high) and (have alcohol shop near home distance) and (not like multitasking) and (feeling uncomfortable while hearing scripture reading) : Alcoholic
6. If (financial loss) and (bored from daily routine) and (do not like the substances that make him feel high) and (have alcohol shop near home distance) and (not like multitasking) and (feeling interested while hearing scripture reading) and (not studied any religious subject) and (feel loneliness) and (little neighborhood attachment) : Alcoholic
7. If (financial loss) and (bored from daily routine) and (like the substances that make him feel high) and (not studied traditional subject) : Alcoholic
8. If (financial loss) and (bored from daily routine) and (like the substances that make him feel high) and (studied traditional subject) and (no burden of home responsibilities) : Alcoholic

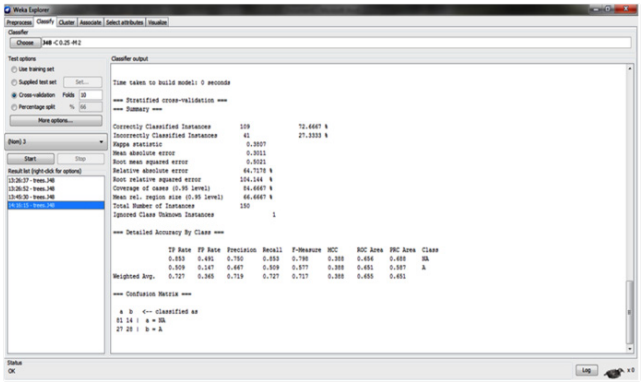

Figure 4. J48 result summary of age group >60

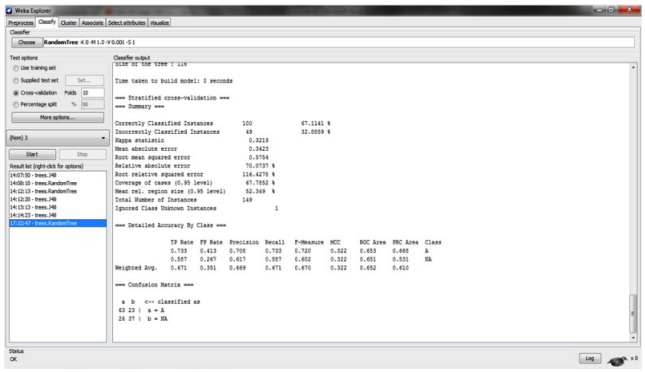

Figure 5. Random tree result summary of age group 0-24

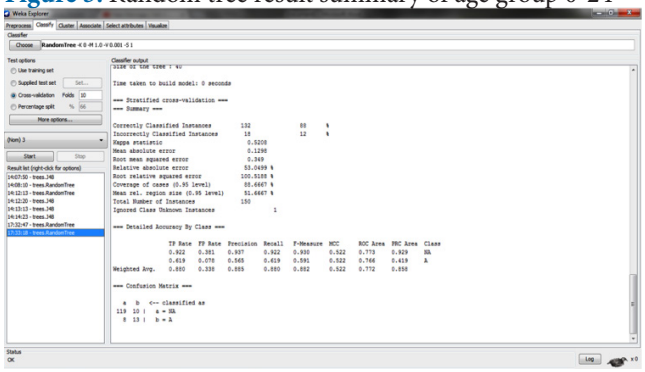

Figure 6. Random tree result summary of age group 25-40

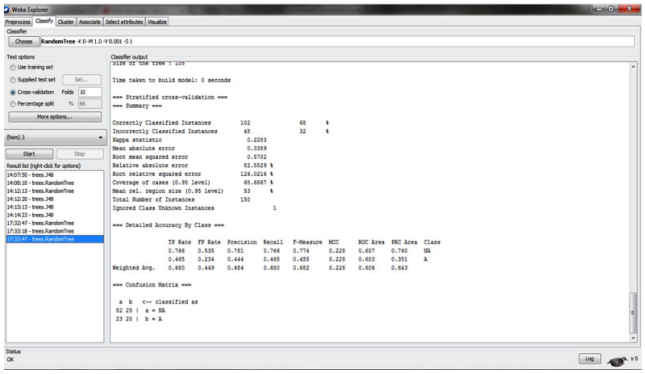

Figure 7. Random tree result summary of age group 41-60

Table 6. Performance Comparison of J48 and Random Tree for age group 0-24

	J48			Random tree		
	TP rate	Recall	Precision	TP rate	Recall	Precision
Alcoholic	0.686	0.686	0.641	0.733	0.733	0.708
Non alcoholic	0.476	0.476	0.526	0.587	0.587	0.617
Weighted Average	0.597	0.597	0.593	0.671	0.671	0.669
Correctly classified instances	59.73%			67.11%		
Incorrectly classified instances	40%			32.88%		

Table 7. Performance Comparison of J48 and Random Tree for age group 25-40

	J48			Random tree		
	TP rate	Recall	Precision	TP rate	Recall	Precision
Alcoholic	0.429	0.429	0.563	0.619	0.619	0.565
Non alcoholic	0.946	0.946	0.91	0.922	0.922	0.937
Weighted Average	0.873	0.873	0.862	0.880	0.880	0.885
Correctly classified instances	87.33%			88%		
Incorrectly classified instances	12%			12%		

Table 8. Performance Comparison of J48 and Random Tree for age group 41-60

	J48			Random tree		
	TP rate	Recall	Precision	TP rate	Recall	Precision
Alcoholic	0.279	0.279	0.364	0.465	0.465	0.444
Non alcoholic	0.804	0.804	0.735	0.766	0.922	0.781
Weighted Average	0.653	0.653	0.629	0.680	0.680	0.684
Correctly classified instances	65.33%			68%		
Incorrectly classified instances	34.66%			32%		

Table 9. Performance Comparison of J48 and Random Tree for age group >60

	J48			Random tree		
	TP rate	Recall	Precision	TP rate	Recall	Precision
Alcoholic	0.509	0.509	0.667	0.691	0.691	0.76
Non alcoholic	0.853	0.853	0.750	0.874	0.851	0.83
Weighted Average	0.727	0.727	0.719	0.807	0.807	0.804
Correctly classified instances	72.66%			80.66%		
Incorrectly classified instances	27.33%			19.33%		

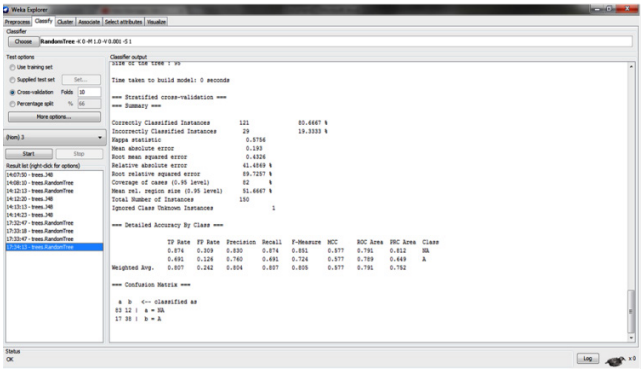

Figure 8. Random tree result summary of age group >60

The performance of both the algorithms is acceptable; but the higher accuracy for all the age group people is attained by random tree. Accuracy for age group 0-24 is 67.11% in case of random tree and 59.73% in case of J48, same for age group 25-40 is 88% in case of random tree and 87.33% in case of J48, same for age group 41-60 is 68% in case of random tree and 65.33% in case of J48 and same for age group >60 is 80.66% for random tree and 72.66% for J48. Also the precision, recall and true positive rate measures of random tree are higher than J48.

5. Conclusion and Future Work

Addiction of alcohol affects the lives of people very deeply. This paper provides a vision towards identifying the attributes that prestige people to turn into drink. In this study, the analysis is carried out to find out the accuracy of different classification algorithms to predict the alcoholic behavior of a person using WEKA. Risk factors that come out to be most effective are impulsive nature, sensation seeking nature, financial loss, family conflict, depression, child abuse, alcoholic shop near home distance. Random

tree provides the higher accuracy of prediction than J48 for all the age groups. The overall accuracy of random tree is 75.94% and for J48 is 71.26%. So as per findings, the random tree provides more precise results than J48. Future research includes the prediction of severity level of people means how prone the person is to take alcohol in future. More attributescan be incorporated to achieve the optimal results for predicting the behavior of an alcoholic person.

6. References

1. Maggs JL, Schulenberg JE. Trajectories of alcohol use during the transition to adulthood. *Alcohol Research and Health*. 2004 Dec; 28(4):195–201.
2. McGue M, Iacono WG. The association of early adolescent problem behavior with adult psychopathology. *American Journal of Psychiatry*. 2005 Jun; 162(6):1118-1124.
3. Williams PS, Hine DW. Parental behavior and alcohol misuse among adolescents: A path analysis or mediating influences. *Australian Journal of Psychology*. 2002 Apr; 54(1):17-24.
4. World Health Organization (WHO). Global status report on alcohol and health. http://www.who.int/substance_abuse/publications/global_alcohol_report/en/. Date Accessed: 2014.
5. Organization for Economic Cooperation and Development (OECD). http://www.oecd.org/health/tackling-harmful-alcohol-use-9789264181069-en.htm. Date Accessed: 12/05/2015.
6. Silver M, Sakara T, Su HC, Herman C, Dolins SB, O'shea MJ. Case study: how to apply data mining techniques in a healthcare data warehouse. *Health care Information Management*. 2001 Feb; 15(2):155-164.
7. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medicine Information*. 2008 Feb; 77(2):81-97.
8. Natalie Guillen, Erick Rotha, Alhena Alfaroa, Erik Fernandez. Youth alcohol drinking behavior: Associated risk and protective factors. *Revista Iberoamericana de Psicologia Y Salud*. 2015 Jul; 6(2):53-63.
9. Jennifer Mendel R, Carla Berg J, Rebecca Windle C, Michael Windle. Predicting young adulthood smoking among adolescent smokers and nonsmokers. *American Journal of Health Behavior*. 2012 Jul; 36(4):542–554.
10. Brett Maclennan, Kypros Kypria, John Langleya, Robin Room. Non-response bias in a community survey of drinking, alcohol-related experiences and public opinion on alcohol policy. *Drug and Alcohol Dependence*. 2012 Nov; 126(1-2):189-194.
11. Marie Yap, Anthony Jorm, Renee Bazley, Claire Kelly, Siobhan Ryan, Dan Lubman. Web-Based Parenting Program to Prevent Adolescent Alcohol Misuse: Rationale and Development. *Australas Psychiatry*. 2011 Aug; 19(4):339-344.
12. Ryan SM, Jorm AF, Lubman DI. Parenting factors associated with reduced adolescent alcohol use: A systematic review of longitudinal studies. *Australian and New Zeeland Journal of Psychiatry*. 2010 Sep; 44(9):774-783.
13. Cable N, Sacker A. Typologies of alcohol consumption in adolescence: Predictors and adult outcomes. *Alcohol and Alcoholism*. 2008 Jan-Feb; 43(1):81–90.
14. Katrijn Houben, Klaus Rothermund, Reinout Wiers W. Predicting alcohol use with a recoding-free variant of the Implicit Association Test. *Addictive Behaviors*. 2009 Jan; 34(5):487-489.
15. Hamburger ME, Leeb RT, Swahn MH. Childhood maltreatment and early alcohol Use among high-risk adolescents. *Journal of Studies on Alcohol and Drugs*. 2008 Mar; 69(2):291-295.
16. Hiro H, Kawakami N, Tanaka K, Nakamura K. Association between job stressors and heavy drinking: age differences in male Japanese workers. *Japan Work Stress and Health Cohort Study Group*. 2007 Jun; 45(3):415-425.
17. Legault L, Anawati M, Flynn R. Factors favoring psychological resilience among fostered young people. *Children and Youth Services Review*. 2006 Sep; 28(9):1024-1038.
18. White HR, Jackson K. Social and psychological influences on emerging adult drinking behavior. *Alcohol Research and Health*. 2004 Dec; 28(4):182-190.
19. Anuradha C, Velmurugan T. A comparative analysis on the evaluation of classification algorithms in the prediction of students' performance. *Indian Journal of Science and Technology*. 2015 Jul; 8(15):1-12.
20. Tripti Mishra, Dharminder Kumar, Sangeeta Gupta. Mining students' data for performance prediction. *Fourth International Conference on Advanced Computing and Communication Technologies*. 2014 Feb, 255-262.