Prediction of Secondary School Students' Alcohol Addiction using Random Forest

B. Hariharan
B.E Computer Science and Engineering.
SSN College of Engineering
Kalavakkam, Chennai, Tamil Nadu, India.

R. Krithivasan
B.E Computer Science and Engineering.
SSN College of Engineering
Kalavakkam, Chennai, Tamil Nadu, India.

Angel Deborah
M.E
Assistant Professor B.E., M.E.
SSN College of Engineering
Kalavakkam, Chennai, Tamil Nadu, India.

ABSTRACT
Teenage alcohol addiction poses a major problem to the well-being of the individual as well as the society. Prevention of this requires identifying the factors causing this addiction. The existing systems mainly rely on decision trees and are able to isolate the factors causing the addiction. The proposed system will be able to predict whether a student with a set of conditions will get addicted to alcohol or not with high accuracy and thereby verify the extent to which the isolated factors are correct.

General Terms
Data mining, Student characteristics dataset

Keywords
Student alcohol behavior Prediction, data mining, patterns, Knowledge patterns

1. INTRODUCTION

1.1 Alcoholism and associated problems
Alcoholism or alcohol abuse might have adverse effect on one's physical as well as mental well-being. The effect of this is more profound on teenagers as they lack the maturity of the adults and they are not fully ready to face the consequences both mentally and physically. During teenage many adolescents tend to experiment with alcohol yet only very few of them are aware of its consequences. This is the time when most of their brain development takes place. This addiction to alcohol and other drugs inhibits the proper neural and endocrine system maturation. It also affects their decision making ability and most of them tend to get involved in violence and resort to other unacceptable behavior like vandalism, etc. This also reflects badly on their academic performance and health.

1.2 Approach
Alcohol addiction is due to a variety of interconnected factors including genetics, your social environment, how you were raised and your emotional health. Most of the time the core factors contributing are stress and inability to handle emotional problems. Most students resort to alcohol and drugs as an escape from reality or a relaxation through which they forget these problems. This addiction must be curtailed in early stages. The most promising route to effective strategies for prevention of teenage alcohol addiction is through a cause-focused approach. This approach requires identification of causal factors and the proper preventive measures to prevent addiction.

1.3 Literature
Several researches have already been carried out in this domain. Among them the most notable one is the isolation of the key factors contributing to alcohol addiction of Secondary school students by Fabio Pagnotta, Hossain Mohammad Amran. In their work they aim at isolating the core factors responsible for secondary school students' alcohol consumption using several machine learning and data mining techniques. Apart from that the effects of alcohol consumption on adolescents is very well described by Hawkins, J. David; Catalano, Richard F.; Miller, Janet Y.

2. PROPOSED SYSTEM
In this work, the main aim is to predict whether a student will get addicted to alcohol given a set of parameters and isolate the core factors and the percentage to which they contribute to this addiction by analyzing a real world dataset from two Portuguese schools using machine learning and data mining techniques. A couple of machine learning algorithms like random forest, support vector machine (SVM) have been used. This was achieved with the help of a software tool called weka.

2.1 Dataset
The dataset consists of details of Portuguese secondary school students gathered by Paulo Cortez and Alice Silva from University of Minho, Portugal. The education system of Portugal is unique. Unlike most countries, the secondary education consists of three years of schooling, preceding nine years of basic education followed by higher studies. This data was gathered from two public schools in Alentjo region of Portugal. It consists of the data about students of 2005-2006 batch. This data was gathered mainly from two sources, a questionnaire and school report.

This data was checked for inconsistencies and 111 entries were discarded as certain students didn't want to reveal their personal information. It also includes demographic details like mother's occupation, father's occupation, family income, etc. Finally, the data was integrated into two data-sets related to Mathematics (with 395 examples) and the Portuguese language (649 records) classes which was further integrated into a single dataset. Certain attributes were discarded during the preprocessing stage. The various attributes are listed in table 1.

Table 1. Sample attributes and their descriptions

Attribute	Description
Sex	Denotes the student's gender
famsize	Denotes the family size (GT3,LT3)
Medu	Student's Mothers education.(no,high,primary,5-9)
Fjob	Student's father's job falls under categories.(services,other,etc)
Internet	Student has internet connection or not.
traveltime	Student's travel time.(<15,15-30,etc)
studytime	Student's study time .(<2,2-5hrs,etc)
schoolsup	School support either yes or no.
Famsup	Family support is there or not.
activities	Other activities yes or no.
higheredu	Student wishes to pursue higher studies or not.
familyrel	Student's relation with his family(bad or good or very bad, etc).
freetime	Student's free time is high.(high, low, very high)
Gout	Student goes out .(high,low,etc)
Health	Student's health condition(bad,good,ok)
Alc	Student is addicted or not.

2.2 Preprocessing

Real world datasets are always prone to noisy, inconsistent and missing data values. This is due to the fact that they might have been collected from a variety of heterogeneous sources. The quality of data is extremely vital. Lower the quality of data poorer the results of data mining process. Perform data cleaning to remove noise and correct inconsistencies in data and merge two different data-set by using data integration. There are many possible reasons for inaccurate data (i.e., having incorrect attribute values). The data collection instruments used may be faulty. There may have been human or computer errors occurring when data entry. Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information (e.g., by choosing the default value). This is known as disguised missing data. Certain attributes were discarded during the preprocessing stage. They were either irrelevant or consisted of very low quality data.

2.3 Learning and Prediction

Random forest and SVM algorithms have been used primarily for classifying the data. Random forest is an ensemble learning technique. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against over fitting. SVM or Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. The region is split into three, one region lies above the plane, one below the plane and one on the plane. The mathematical formulation is an equation taking all the factors into account and based on this the classification occurs.

In addition, it is very user-friendly in the sense that it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values. Decision trees are the heart of this work. They were used for having a good prediction, find correlation between features and as we saw before for pre-processing the data set.

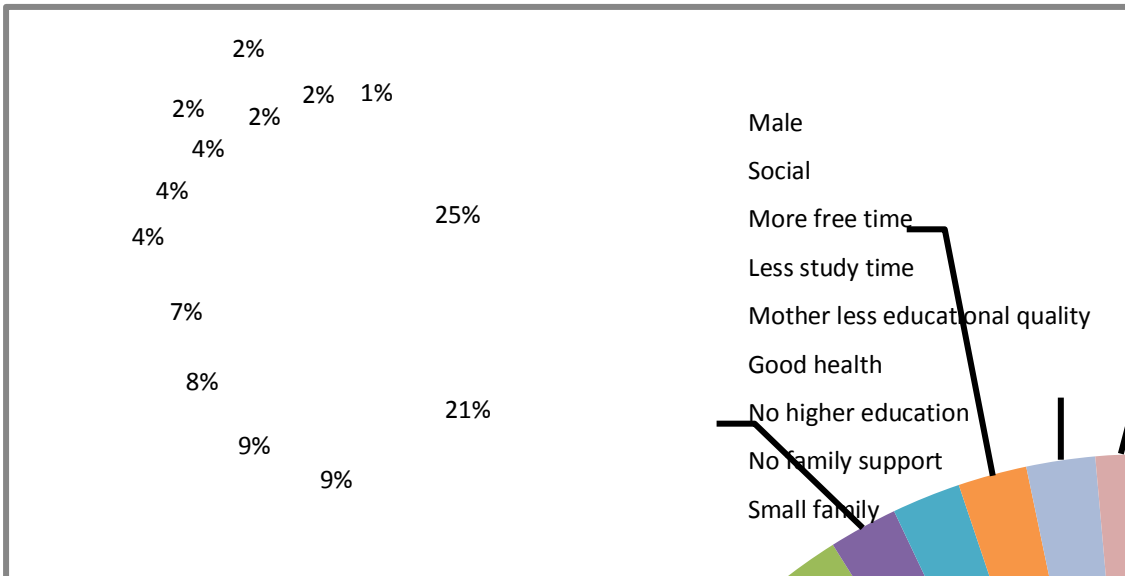

Fig 1: Attributes and their contribution

Random forest algorithm used is:

1. For b = 1 to B:

(a) Draw a bootstrap sample Z* of size N from the training data.

(b) Grow a random-forest tree Tb to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size nmin is reached.

i. Select m variables at random from the p variables.

ii. Pick the best variable/split-point among the m.

iii. Split the node into two daughter nodes.

2. Output the ensemble of trees {T b}13.

To make a prediction at a new point x:

Regression: (x) = EbB_i Tb(x).

Classification: Let O b(x) be the class prediction of the bth random –forest tree. Then Cg(x) = majority vote {Cb(x)}P.

2.4 Experiment and discussion

The primary goal of the project is to predict whether a student will get addicted to alcohol based on a set of parameters. First a set of 14 parameters from the set of attributes in the dataset are selected. These 14 attributes contribute the most towards a student's addiction to alcohol. These also contain few demographic details. The 14 attributes and the percentage they contribute is shown in figure 1 and table 2.Now the percentages are scaled down to a scale of ten and the various attributes are assigned weightage depending on that. In the preprocessing stage the 14 attributes are converted to binary data from categorical data. The predictor field is the alc field. The value of alc is calculated and it ranges between 0-5.Further it is converted into categorical data and divided into two sets 0-2 -A and others B.

Table 2. Most impacted attribute

Attribute	Percentage
Male	25.35%
Social	21.13%
More free time	9.39%
Less study time	8.45%
Mother less educational quality	7.98%
Good health	7.04%
No higher education	4.23%
No family support	3.76%
Small family	3.76%
High travel time	1.88%
Less activities	1.88%
No support school	1.88%
Father work	1.88%
Internet connectivity	1.41%

The weka tool is used to perform the classification of the dataset. Thanks to weka tool for enabling us visualize interesting information about the data. The visualizations are shown in figure 2 and figure 3.The random forest algorithm is applied for classification of the data. The dataset was classified using both SVM and Random Forest. The system and set of selected attributes work fine for SVM with considerable accuracy of 96.6%(approximately).The system works with a higher accuracy when random forest algorithm is applied. An accuracy of 98.5% was achieved.

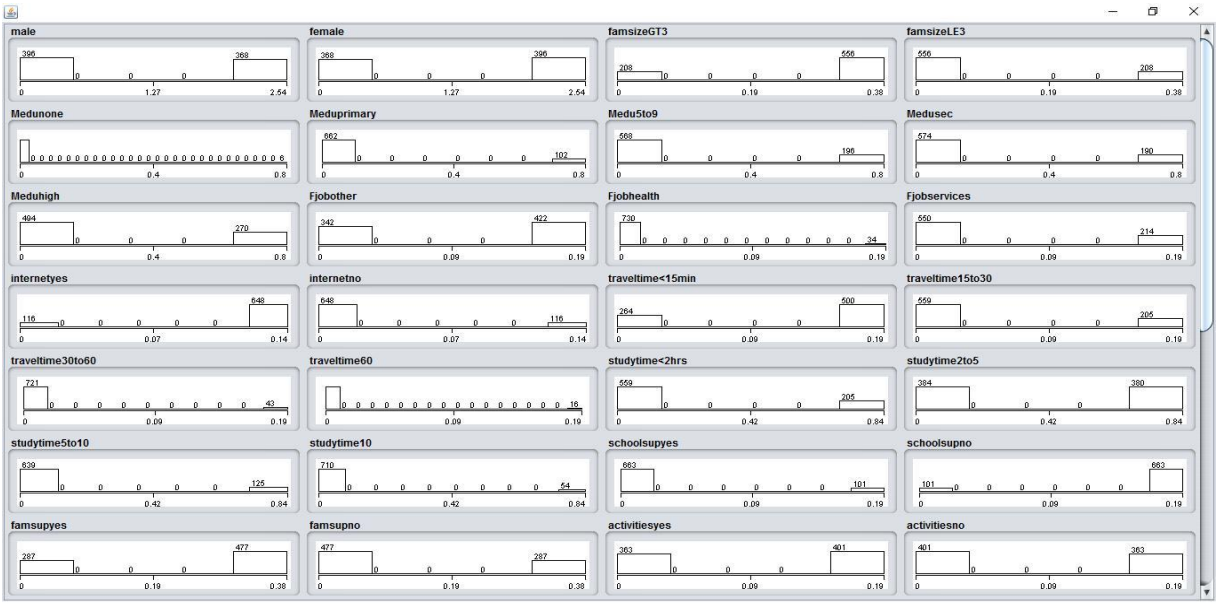

Fig 2: Weka full view 1

Fig 3: Weka full view 2

3. RESULT

The data was successfully classified and predicted with an accuracy of nearly 98.5%(98.429 to be precise).The result is shown in the screen snapshot figure 4.

Through this the validity of assigning of weightage and impact of each attributes has been verified to be correct.

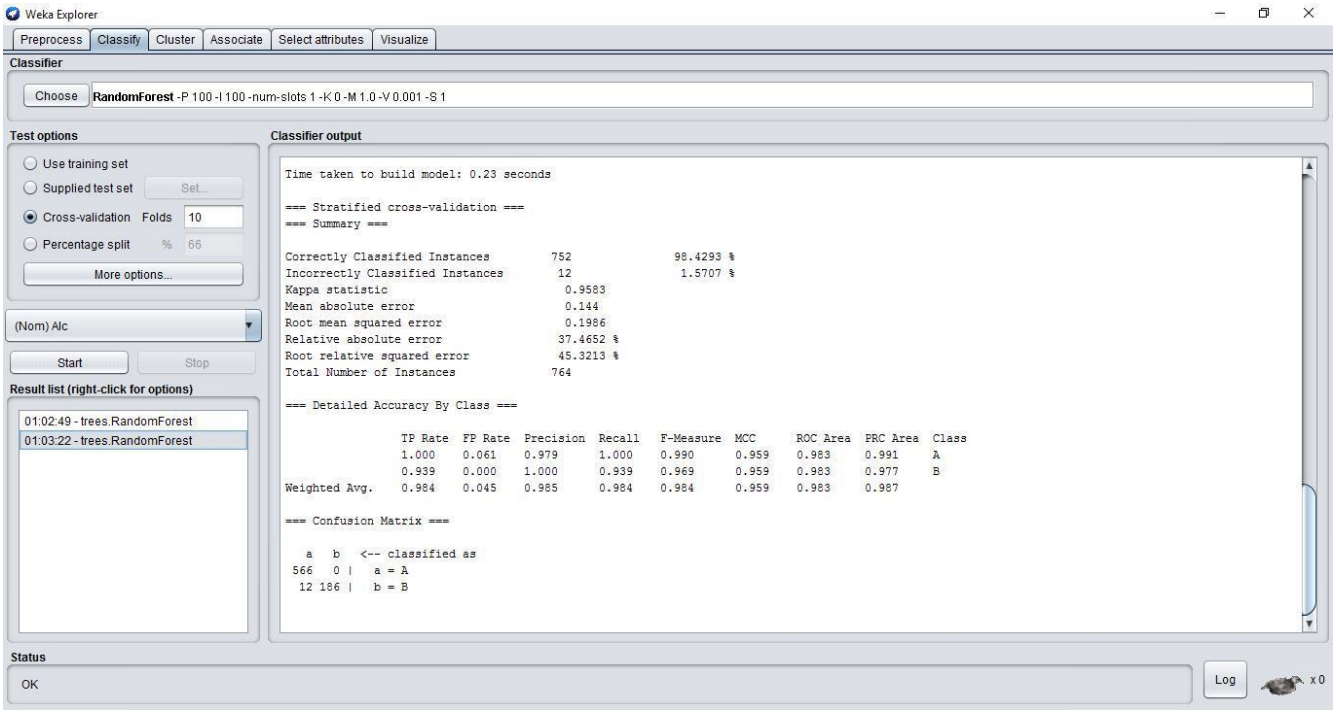

Fig 4: Weka output snapshot

4. CONCLUSION

From earlier works in this domain the factors responsible for students' alcohol addiction were identified. They had used a number of algorithms and arrived at an accuracy of about 93%.But a weightage was assigned for each factor depending upon the amount of impact each factor has on a person's alcohol addiction and this system was able to predict whether a student will be addicted to alcohol or not given a set of attribute values with an accuracy of 98.5%.The random forest algorithm was used for this and a high accuracy rate was achieved by careful inspection of the factors responsible and the extent to which they were responsible and considering it in the Prediction process. The full potential of the project can be felt only when the process is done on a larger set of current data. The first step in reduction of student consumption of alcohol is isolation of the key factors responsible. This has been implemented with considerable accuracy. With this

information it is possible to find out which demographic or non-demographic factor causes alcohol addiction and to what extent it influences the addiction. For example, more free time for students, encouraging them to take part in co-curricular activities, spending more time with family, reducing study time and providing a stress free environment etc would reduce alcohol addiction because most students seek alcohol as an escape from reality or a way of relaxation. This will be highly useful in controlling the alcohol addiction rate. Technology touches people's life only when it has an impact on their day to day lives and addresses their common problems. One such attempt has been made with the intention to uplift the society and thereby the world as a whole.

5. REFERENCES

[1] P. Cortez and A. Silva. Using data mining to predict secondary school student performance. [In a. brito and j. teixeira eds. proceedings of 5th future business technology conference (fubutec2008) pp: 5-12 porto portugal]. 2008.BRITO, A. ; TEIXEIRA, J., eds. lit. – "Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008". [S.l. : EUROSIS, 2008]. ISBN 978-9077381-39-7. p. 5-12..

[2] Breiman L. Friedman J. Ohlsen R. and Stone C. classification and regression trees. 1984.ISBN 978-1-118-44714-7

[3] Using Data Mining To Predict Secondary School Student Alcohol Consumption. Fabio Pagnotta, Hossain Mohammad Amran, Department of Computer Science, University of Camerino

[4] Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention. Hawkins, J. David; Catalano, Richard F.; Miller, Janet Y. Psychological Bulletin, Vol 112(1), Jul 1992, 64-105

[5] Classification and Regression by random Forest Andy Liaw and Matthew Wiener , R News Vol. 2/3, December 2002. ISSN 1609-3631 ... 0.1842105. Veh. 7. 4. 6. 0. 0. 0. 0.6470588. Con. 0. 2. 0 10. 0. 1. 0.2307692.

[6] http://www.statsoft.com/Textbook/Support-Vector-Machines

[7] Drinkaware.co.uk. Why underage drinking is a risky business.

[8] Drugfreeworld.org. The truth about alcohol.

[9] ISBN:0387848584 The Elements of Statistical Learning: Data Mining, Inference, and Prediction.By Trevor Hastie, Robert Tibshirani, Jerome Friedman.