ANALITICAL CLASSIFIER TECHNIQUES APPLIED ON STUDENT ALCOHOL CONSUMPTION

DIVYA R. JARIWALA,
Research Scholar,
JJT University, Chudela,
District-Jhunjhunu,
Rajasthan State, India

HETA DESAI,
Assistant Professor,
UCCC & SPBCBA & UACCAIT,
Udhna-Navsari Road,
Surat, Gujarat, India

ABSRACT:

Alcohol use among students either they are school or college students are very extensive. Many students may consider drinking as a prevalent part of a social life during college. Many students sense negative reaction of alcohol utilization; however, most analysis considers these consequences for students who drink. This analysis inspects whether the positive organization between college pupils' current and high-school drinking is due to bias formation or the collision of unobserved factors of single taste. Conclusive the System underlying the resolution in alcohol use has significant policy implications. If bias formation exists, then guidelines that decrease alcohol need in single period should also reduce alcohol use in future era. The data was analyzed by using WEKA software. If however, perseverance follows unmeasured personal features, then protocol targeting teens will have no force on their large period drinking attitude. Data Mining is a materialized technique with the help of this own can efficiently learn with ancient data and use that knowledge for predicting future attitude of uneasy areas.

Index Terms: *WEKA, Alcohol consumption, Data Mining.*

I. INTRODUCTION

This analysis consider whether the positive organization between college students' current and high-school drinking is due to bias creation or the control of clandestinely parts of individual taste [1]. Observational research has established that needless drinking is organized with an increased likelihood of drinking and driving, violence, crime, and poor labor market and educational outcomes. The costs organized with drinking connected behaviors are not borne completely by those who charter in excessive drinking. This paper discriminate itself by taking an intertemporal angle to examining college students' drinking behavior [1]. Although the most of college student are below the legal drinking age, alcohol continues to be widely used on most college campuses today [2].

Students make decisions about how enough alcohol to consume. The utilization of alcohol can be normal to have a negative force on schooling both straight through its future force on subjective ability and indirectly through its collision on study bias. A negative interacting between alcohol consumption and schooling also may be noticed, however, due to the reality that singles who look high costs and/or place a minor value on future gains may invest limited in schooling and at the equivalent time these singles may be much likely to appoint in excessive drinking behavior [3].

II. RELATED WORK

Researchers presents research work on classification data mining approach is applied on the training data set to recognize the problem. The work is depend on data classification by using many different classifiers such like Multilayer Perceptron, Kstar, LWL, RepTree, Linear Regression and SMOreg algorithms.

A. Data set used

Researcher use "STUDENT ALCOHOL CONSUMPTION" Data Set massive from UCI machine open access database for finding results. The investigation also contributes the communication between alcohol usage and the social, gender and study time attributes for each student. Total 395 Instances are used on this database as training set and other details of this data displayed in table 1.

Data Set Characteristics:	Multivariate	**Number of Instances:**	1044	**Area:**	Social
Attribute Characteristics:	Integer	**Number of Attributes:**	32	**Date Donated**	2016-03-03
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of**	29846

				Web Hits:	

TABLE 1 Represent dataset values and instances.

B. WEKA as a Tool

Researchers select WEKA (Waikato Environment for Knowledge Analysis) software that was generated at the University of Waikato in New Zealand. WEKA is open source software expressed under the GNU General Public License. It accommodates tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It is portable & platform independent berceuse's it is fully implemented in the Java programming language and thus runs on almost any modern computing platform and is now recycled in more different application areas, in particular for education & research [6]. The software WEKA is suitable because it is open source. WEKA can effectively work with limited data [11]. WEKA also provides convenient data preprocessing, cleaning and handling missing values. It takes data from excel file in Comma Separated Values (CSV) format, which is a very common application software to be used in each school for initial collection of data [12]. This software includes form for an entire area of data mining tasks like Data pre-processing, Classification, Clustering, Association and Visualization.

C. Data Mining Architecture

Data Mining is the technique of locating non-obvious and conceivably valuable designs in big data warehouses. A Data mining construction is a conceptual representation of the arrangement of interconnections between, the hardware and software devices complicated in the mining methods. In the past less years, digits of architectural figures have been advanced for the aim of data mining. Information was collected about a fair number of models, from the web [4].

Data is now saved in databases and/or data warehouse organization so should we design a data mining scheme that decouples or couples with databases and data warehouse systems? This question advantages to quarter possible architectures of a data mining system as follows:

No-coupling: in this construction, data mining system does not apply any performance of a database or data warehouse system. A no-coupling data mining organization recover data from an appropriate data source and stores results into the file system. The no-coupling data mining construction does not take any influences of database or data warehouse that is already actual efficient in organizing, storing, accessing and retrieving data. The no-coupling construction is deliberate a destitute architecture for data mining system; however, it is used for simple data mining processes.

Loose Coupling: in this construction, data mining system helps the database or data warehouse for data retrieval. In loose coupling data mining construction, data mining arrangement recovers data from the database or data warehouse, developments data using data mining algorithms and collection the result in those systems. This construction is mostly for memory-based data mining system that does not desire huge scalability and huge performance.

Semi-tight Coupling: in semi-tight coupling data mining constructions, beyond connecting to database or data warehouse engineering, data mining system usage several appearance of database or data warehouse systems to execute some data mining assignment include sorting, indexing, aggregation…etc. In this construction, little standard result can be saved in database or data warehouse system for better performance.

Tight Coupling: in tight coupling data mining construction, database or data warehouse is conducted as an information retrieval part of data mining system using combination. All the appearance of database or data warehouse is recycled to perform data mining tasks. This construction provides organization scalability, high performance, and integrated information.

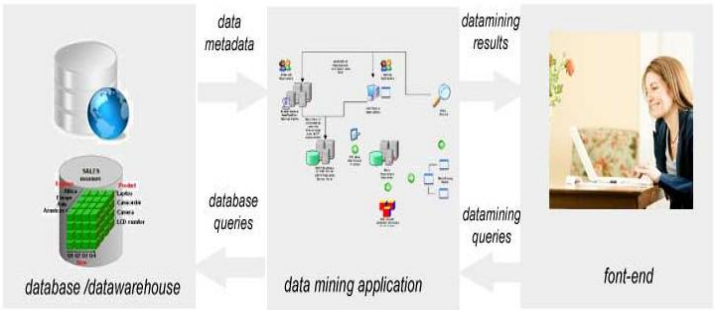

FIG. 1 Data Mining Architecture

There are 3 tiers in the tight-coupling data mining architecture:

Data layer: as mentioned above, data layer can be a database and/or data warehouse systems. This layer is a consolidated for all data sources. Data mining analysis are saved in data layer so it can be granted to end-user in the form of reports or another kind of measurements.

Data mining application layer is used to recover data from the database. Few transformation habitual can be accomplished here to convert data into the wanted format. Then data is processed using various data mining algorithms.

Front-end layer provides perceptive and friendly user consolidated for end-user to interact with data mining system. Data mining analysis presented in measurements form to the user in the front-end layer.

III. METHODOLOGY

Researchers applied data mining classification techniques for finding best results among following algorithm

A. Multilayer Perceptron

Multilayer Perceptron (MLP) is a non-linear classifier based on the perceptron. The learning rule for the Multi-layer Perceptron is named as Back Propagation Rule also called as Generalized Delta Rule. MLP is a back multiplication neural network with single or many layers. The following diagram illustrates a perceptron network with three layers.

In this model network, the neurons will be arranged into 3 or more layers which include an input layer, a resulting layer and single or many hidden layers. The back propagation rule continuously calculates a value based on an error function for each input and back propagates the error from one layer to the previous layer and the weights of the nodes are adjusted accordingly. MLP can be used to solve non-linear problems by connecting a number of neurons in the form of layers. Each of the perceptron identifies linearly separable inputs and the outputs of the perceptions are combined into a new perceptron to get the final output [5].

B. K-Star algorithm

K-star is a situation-based classifier that is the kind of a test case is based upon the kind of those training occurrence similar to it, as resolved by some similarity function. It distinct from other case-based beginners in that it uses an entropy-based distance function. The K* algorithm can be represented as an approach of cluster analysis which mainly aims at the partition of „n‟ observation into „k‟ clusters in which every observation belongs to the cluster with the shortest mean. We can describe K* algorithm as an instance based learner which uses entropy as a distance measure. The benefits are that it administers a consistent technique to handling of real valued attributes, symbolic attributes and missing values [8] [7].

C. LWL Algorithm

LWL methods are non-incidental and the current forecast is done by local functions which are using only a subgroup of the data. The primary concepts behind LWL is that instead of building a global ideal for the entire function space, for each point of interest a local model is created based on neighboring data of the query point. For this purpose every data pixels becomes a weighting factor which expresses the influence of the data point for the prediction. In general, data points which are in the close neighborhood to the current query point are receiving a higher weight than data points which are far away. LWL is also called lazy education, because the processing of the training data is shifted until a query point needs to be answered. This approach makes LWL a very accurate function approximation method where it is easy to add new training points [9].

D. REPTree Algorithm

RepTree recycles the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be studied as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree. Basically Reduced Error Pruning Tree ("REPT") is rapid decision tree learning and it frames a decision tree based on the information gain or reducing the deviation. REP Tree is frequent decision tree beginners which frames a decision/regression tree utilizing information advance as the dividing test, and reduce it utilizing diminished error shearing. It only character code for numeric characteristics

individual. Missing values are dealt with using C4.5's approach of using fractional situation. The illustration of REP Tree algorithm is enforced on UCI repository and the confusion matrix is generated for kind gender having six possible values. [13] [14] [15] [10]

E. Linear Regression

Regression is the effortless approach to help, but is also doubtless the first powerful. This approach can be as simple as one received variable and single result variable. Of course, it can earn many complicated than that, including dozens of received variables. In effect, regression figure all suitable the identical general designs. There are a digits of independent variables, which, when captured composed, produce an analysis — a dependent variable. The regression model is then recycled to anticipate the analysis of an unknown helpless variable addicted the values of the independent variables

F. SMO reg

SMO reg appliances the support vector machine for regression. The criterion can be beginners utilizing different algorithms. The algorithm is selected by setting the Reg Optimizer [16]. This application universally restores all missing values and revolutionizes theoretical attributes into twice individuals. It also distributes all attributes by default [17].

IV. RESULT

The research work has chosen kstar algorithm for approximate study and to analysis the attributes name of the school, sex, age of the student, address, family size, parent's cohabitation status, mother's education, father education, mother's job, father's job, reason to choose school, guardian, travel time, study time, number of past class failure, Extra education support, family education support, extra paid class, activity, attended nursery school, want for higher education, internet use at home, romantic relation, family relationship, free time after school, going out with friends, workday alcohol consumption, weekend alcohol consumption, health, and attendance. We applied multilayer perceptron, k star, LWL, REP Tree, Linear Regression and SMO reg algorithm on this data set. We found that k star algorithm is best among this entire algorithm because it clearly shows that time taken to build model is 0.01 seconds, Time taken to test model is 1.9 seconds, co-relation co-efficient is 1, Mean absolute error is 0, Root mean squared error is 0.0001, relative Absolute error is 0.0002% and Root Relative Squared error is 0.0022%. In kstar algorithm co-relation co-efficient value is high and other errors values are nearest to 0. Multilayer Perceptron is worst for this dataset because all possible errors nearest to 1. This research clearly showed in table 2 with numerical as well as figure 2. So, bases of this result researcher suggested Kstar algorithm is best for this "Student Alcohol Consumption" dataset and it is generate best result among them.

Algorithm Used	Time taken to buid model	Time taken to test model	Correlation Coefficient	Mean Absolute error	Root mean squared error	Relative absolute error	Root Relative Squared error
MultilayerPerceptron	9.8 seconds	0.02 seconds	0.9971	0.4864	0.5652	14.1856%	12.3519%
Kstar	0.01 seconds	1.9 seconds	1	0	0.0001	0.0002%	0.0022%
LWL	0 seconds	1.41 seconds	0.7641	2.3661	2.9525	69.0038%	64.5270%
REPTree	0.06 seconds	0 seconds	0.9472	0.839	1.4675	24.4678%	32.0721%
Linear Regression	0.07 seconds	0 seconds	0.9165	1.1872	1.83	34.6222%	39.9944%
SMOreg	1.23 seconds	0.01 seconds	0.9054	0.9659	2.0026	28.168%	43.7670%

TABLE 2 Different classification techniques applied on student alcohol consumption dataset

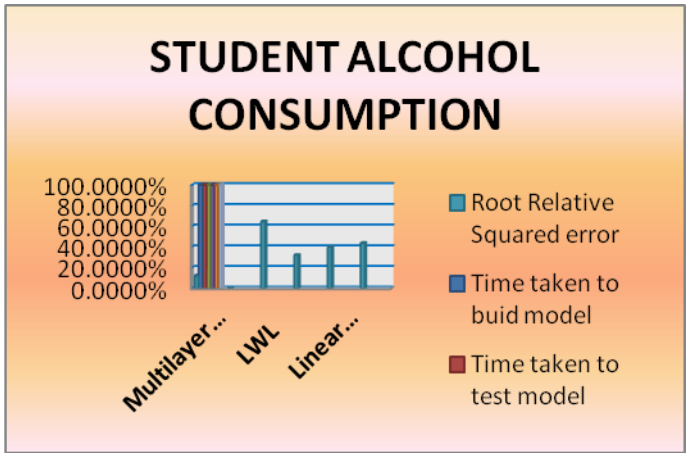

FIG. 2 Different classification techniques applied on student alcohol consumption charts

V. CONCLUSION

Preventing alcohol related harm is a critical health priority. It requires a combination of legal and regulatory interventions, enforcement, and community based programmers and actions, better health and social services which focus on alcohol, personal behavior change and shifts in community attitudes. In this paper it is found that, as with the general population of young adults, the persistence exhibited in the sample of college students 'drinking is attributable to habit formation.

REFERENCES

[1] Jenny Williams (2002)," Habit and Heterogeneity in College Students' Demand for Alcohol"
[2] Shadd Cabalatungan," THE CONSEQUENCES OF ALCOHOL CONSUMPTION FOR DRINKING AND NON-DRINKING STUDENTS"
[3] Lisa M. Powell, Jenny Williams, Henry Wechsler (2002)," Study Habits and the Level Of Alcohol Use among College Students Lisa M."
[4] Thomas Thomas, Sanjeev Jayakumar, B.Muthukumaran, "DATA MINING ARCHITECTURES – A COMPARATIVE STUDY"
[5] "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification"
[6] S. Lakshmi Prabha and Dr.A.R.Mohamed Shanavas (2015) "Application of Educational Data mining techniques in e-Learning- A Case Study" , International Journal of Computer Science and Information Technologies, Vol. 6, No. 5, pp.- 4440-4443.
[7] Ms S. Vijayarani, Ms M. Muthulakshmi (2013), "Comparative Analysis of Bayes and Lazy Classification Algorithms" International Journal of Advanced Research in Computer and Communication Engineering, vol 2, pp.3118-3124.
[8] Trilok Chand Sharma, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm".
[9] Peter Englert, "LocallyWeighted Learning"
[10] Sushilkumar Kalmegh(2015)," Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News" IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2, pp. 438-446
[11] "WEKA Data Mining Book" (n.d.) http://www.cs.waikato.ac.nz/~ml/weka/book.html.
[12] "WEKA 3: Data Mining Software in Java" (n.d.) Retrieved March 2010 from http://www.cs.waikato.ac.nz/ml/weka/.
[13] Ian H. Witten, Eibe Frank & Mark A. Hall., "Data Mining Practical Machine Learning Tools and Techniques, Third Edition." Morgan Kaufmann Publishers is an imprint of Elsevier.
[14] Dr. B. Srinivasan, P.Mekala, "Mining Social Networking Data for Classification Using REPTree", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 10, October 2014 pp- 155-160
[15] Payal P.Dhakate, Suvarna Patil, K. Rajeswari, Deepa Abin, "Preprocessing and Classification in WEKA Using Different Classifier", Int. Journal of Engineering Research and Applications, Vol. 4, Issue 8(Version 5), August 2014, pp- 91-93
[16] http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMOreg.html
[17] http://www.dbs.ifi.lmu.de/~zimek/diplomathesis/implementations/EHNDs/doc/weka/classifiers/functions/SMOreg.html