# USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT ALCOHOL CONSUMPTION

Fabio Pagnotta Mat:-093579 Mohammad Amran Hossain Mat:-093192
Department of Computer science, University of Camerino
Advanced Database

In this project, we use a data set about Portuguese student on two courses ( Mathematics and Portuguese ) which was collected and analysed by *Paulo Cortez and Alice Silva, University of Minho,Portugal*. Our work intends to approach student addiction on alcohol in secondary level using business intelligence (*BI*) and Data Mining (*DM*) techniques. The result shows that a good predictive accuracy can be achieved, provided that addiction of alcohol can impact to the student performance. In addition,the result also provides the correlation between alcohol usage and the social, gender and study time attributes for each student. As a direct outcome of our project, more efficient prediction tools can be developed in order to pay more attention to the student and share how the alcohol impact so badly in his life.

## Introduction

Alcohol had lots of bad impact in our life. Drinking too much on a single occasion or over time can take a serious toll on our health. If who drinks alcohol it's likely he is experienced first-hand at least some of its short-term health effects, be it a hangover or a bad night's sleep. Alcohol have many short-term and long-term health effects. Taking alcohol as teenager age, reduce a child's mental and physical abilities, affecting judgment and co-ordination which can lead to trouble. The level of alcohol gets so high that the brain's vital functions, which include breathing control, are blocked. Alcoholics were more likely to get injured or have accidents than non-drinkers. More worrying still, they're more likely to be a passenger in a drink-driving incident. When children drink, their decision making skills are affected and they're more likely to take big risks like having unprotected sex. That can lead to sexually transmitted diseases and unwanted pregnancy. While excessive drinking by adolescents is a problem in its own right, it is at times linked to other harmful behaviours like taking illicit drugs. Alcohol can be a one-way ticket to feeling and looking downright grotty ! Underage drinkers are more likely to suffer from a range of health issues including major weight gain or weight loss, bad skin, disturbed sleep, headaches. During childhood and teenage years, the brain is still developing. Adding alcohol to that process is asking for trouble. It can affect memory function, reactions, learning ability and attention span all especially important during their school years. Drinking could affect child's performance at school and prevent them from reaching their full potential. Young people who drink excessively may be more likely to also have disturbed mental health, even self-harm. Every parent wants their child to make the best of themselves and performing well at school plays a big part in that. The stats show underage drinking makes that less likely. Children who start to drink by age 13 are more likely to go on to have worse grades, to skip school and, in the worst case scenario, to be excluded from school. They have less self-control and their brains struggle to recognise warning signs. This can lead to aggression and fights. Their risk of being involved in violence and serious vandalism increases directly in line with alcohol consumption, which could lead to arrest and a criminal record. Their natural tendency to experiment and take risks is increased. Adding alcohol to the mix is not a good idea; it can put them in vulnerable or dangerous situations. [1] [2] [3]

In this work, we will analyze recent real-world data from two Portuguese secondary schools. Two different sources were used: mark reports and questionnaires. It allowed the collection of several demographic social and school related attributes. Two DM algorithms (e.g.

Decision Trees, Random Forest) will be tested. Moreover, an explanatory analysis will be performed over the best models, in order to identify the most relevant features.

## Materials and Methods

**Data set.** We use a data set about Portuguese student which was composed by *Paulo Cortez and Alice Silva, University of Minho,Portugal*[4] . In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. Most of the students join the public and free education system. This study will consider data collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal. Hence, the database was built from two sources: school reports, based on paper sheets and including few attributes ; and questionnaires, used to complement the previous information. They designed the latter with closed questions related to several demographic (e.g. mother's education, family income), social/emotional (e.g. alcohol consumption) and school related (e.g. number of past class failures) variables that were expected to affect student performance. The questionnaire was reviewed by school professionals and tested on a small set of 15 students in order to get a feedback. The final version contained 37 questions in a single A4 sheet and it was answered in class by 788 students. Latter, 111 answers were discarded due to lack of identification details (necessary for merging with the school reports). Finally, the data was integrated into two data-sets related to Mathematics (with 395 examples) and the Portuguese language (649 records) classes.
During the preprocessing stage, some features were discarded due to the lack of discriminative value. For instance, few respondents answered about their family income ( probably due to privacy issues ), while almost 100 % of the students live with their parents and have a personal computer at home. The remaining attributes are shown in Table 1, where the last four rows denote the variables taken from the school reports.

## Data Mining Models

We use decision tree for classification. Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifier, predict categorical (discrete, unordered) class labels. Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data). classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "earning from" a training set made up of database tuples and their associated class labels.
Several DM algorithms, each one with its own purposes and capabilities, have been proposed for classification tasks. The Decision Tree (DT) is a branching structure that represents a set of rules, distinguishing values in a hierarchical form [5]. This representation can translated into a set of IF-THEN rules, which are easy to understand by humans. The Random Forest (RF)[6] is an ensemble of T unpruned DT. Each tree is based in a random feature selection from bootstrap training samples and the RF predictions are built by averaging the outputs of the T trees. The RF is more (*dicult*) to interpret when compared with the single DT, although it is still possible to provide explanatory knowledge in terms of its input variable relevance.
We use *KNIME* [7] Analytics Platform as a tools to perform our work. Konstanz Information Miner(KNIME) is a modular, open source platform for data integration, processing, analysis and exploration. The visual representation of the analysis steps enables the entire knowledge discovery process to be intuitively modeled and documented in a user-friendly and comprehensive fashion. KNIME build work-flows. Work-flows consist of nodes that process data; the data are transported via connections between the nodes. A workflow starts with nodes that read the data from some data sources, which are usually databases that can be queried by special nodes. Imported data are stored in an internal table-based format consisting of columns with a certain data type (integer, string, image, molecule, etc.) and an arbitrary number of rows conforming to the column specifications. These data tables are sent along the connections to other nodes, which first pre-process the data, e.g. handle missing values,filter columns or rows, partition the table into training and test data, etc. and then for the most part build predictive models with machine learning algorithms like decision trees, Naive Bayes classifier or support vector machines. For inspecting the

Table 1
*The preprocessed student related variables.*

| Attribute | Description (Domain) |
|---|---|
| sex | student's sex (binary: female or male) |
| age | student's age (numeric: from 15 to 22) |
| school | student's school (binary: *Gabriel Pereira or Mousinho da Silveira*) |
| address | student's home address type (binary: urban or rural) |
| Pstatus | parent's cohabitation status (binary: living together or apart) |
| Medu | mother's education (numeric: from 0 to 4[a]) |
| Mjob | mother's job (nominal[b] close to home, school reputation, course preference or other) |
| traveltime | home to school travel time (numeric: 1-< 15 min., 2-15 to 30 min., 3-30 min. to 1 hour or 4 -> 1 hour). |
| studytime | weekly study time (numeric: 1-< 2 hours, 2- 2 to 5 hours, 3-5 to 10 hours or 4 -> 10 hours) |
| failures | number of past class failures (numeric: n if $1 \leq n < 3$,else 4) |
| schoolsup | extra educational school support (binary: yes or no) |
| famsup | family educational support (binary: yes or no) |
| activities | extra-curricular activities (binary: yes or no) |
| paidclass | extra paid classes (binary: yes or no) |
| internet | Internet access at home (binary: yes or no) |
| nursery | attended nursery school (binary: yes or no) |
| higher | wants to take higher education (binary: yes or no) |
| romantic | with a romantic relationship (binary: yes or no) |
| freetime | free time after school (numeric: from 1- very low to 5- very high) |
| goout | going out with friends (numeric: ffrom 1- very low to 5- very high) |
| health | current health status (numeric: from 1- very bad to 5- very good) |
| absences | number of school absences (numeric: from 0 to 93) |
| G1 | first period grade (numeric: from 0 to 20) |
| G2 | second period grade (numeric: from 0 to 20) |
| G3 | final grade (numeric: from 0 to 20) |
| alc | Alcohol consumption between week (numeric: from 1- very low to 5- very high) |

a 0-none, 1- primary education (4th grade), 2- 5th to 9th grade, 3- secondary education or 4 - higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

results of an analysis workflow several view nodes are available, which display the data or the trained models in various ways.

## Project Description

In project description we discuss how we prepare the data before start work, learning outcome of our project and what is the prediction of this work. We also provide information about, how perform testing on our data-set.

### Preprocessing

Real-world databases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results. We use several data pre-processing techniques[8]: 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 "January 1" displayed for birthday). This is known as disguised missing data. In our case we use a data-set which is collect and store by Portuguese professor, perform the test for missing data or value on this data-set and find no missing value or data in those data-set.

Our goal is finding alcohol consumption by secondary school student. In this data-set, there are two different attribute on alcohol. The former is alcohol taking in work day(Dalc) and the latter is alcohol taking in weekend(walc). Only one index can be predicted, so we made an attribute that represents the total alcohol taking by a specific student in a whole week. So we merge those two attribute the following equation 1

$$Alc = \frac{Walc \times 2 + Dalc \times 5}{7} \tag{1}$$

The new attribute also changes between one and five. In order to check if a student is a drinker or not. Alc becomes a binary value if is lower than 3 is 0, that means you are not a drinker, otherwise 1. An important contribution was made by using Weka tools [9]. As we can see in figure .1

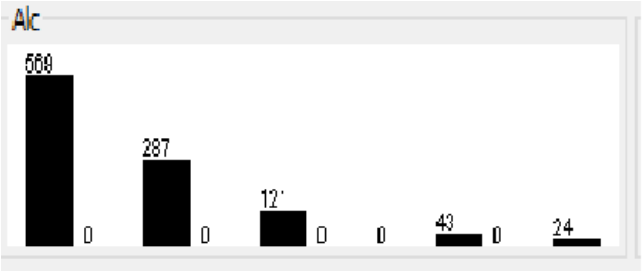


*Figure 1*. Alcohol Consumption

This allows to see which is the most interesting features. As we can see in full image .2 Thanks to this screen, we can manipulate data in order to get a distribution.

We try to find student absence rate at school. If a student frequently make absence in school, he takes more alcohol than others. Therefore this attribute becomes a

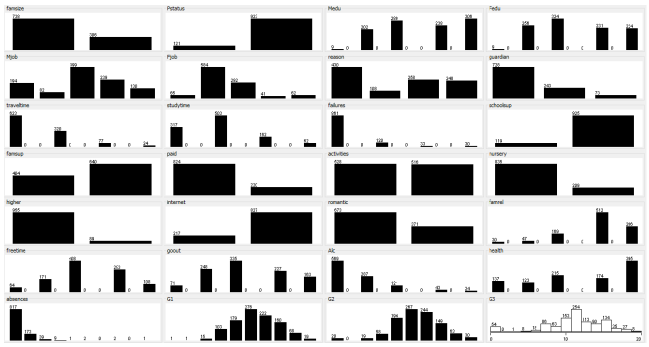

*Figure 2*. Weka full view

binary value, if he frequently make absence ( Over 10 days) is 0, otherwise value is 1. We can predict absence rate using the image 2.

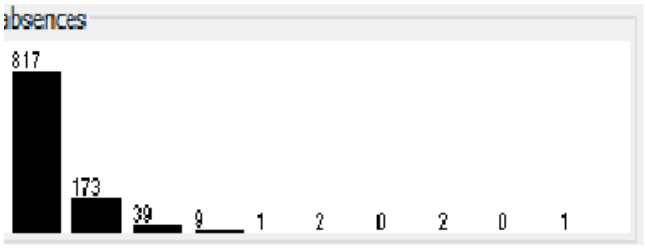


*Figure 3*. Absences Rate

We use rule engine in *KNIME* work-flow to perform this data reduction, Concatenate component for merge two data-set: *student-mat* and *student-por* into a new data-set and convert binary value of *Alc* and *Fabs* to nominal value to perform correlations. Show in figure 4

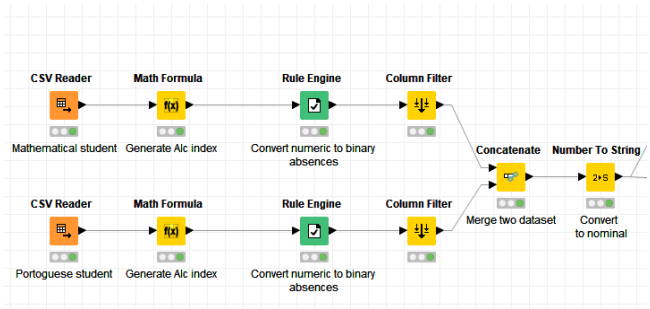


*Figure 4*. Data Preprocessing by KNIME

We perform linear correlation and filter low correlation value where correlation threshold value is less then 0.35. It filtered four column attribute($Fedu, G1, G2, G2$) that is not correlated to Alc. We also perform elimination of backward feature. Show in figure 5. We use a loop with cross validation for elimination. Cross validation is used to test data

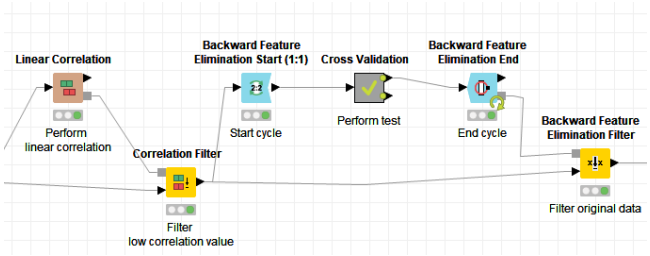


*Figure 5*. Data Correlation using KNIME

by using *Random Forest*. This procedure again used to perform prediction and testing the result. The last procedure of data Preprocessing is filter backward elimination feature data to rescue original data. Here 11 column attribute is filtered which is not correlated with Alc attribute.

### Learning and Prediction

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. A decision tree is composed by several IF-THEN in cascade. When the algorithm creates a decision tree, it needs to decide which attribute are involved in the splitting phase. There is an index called split criterion for this purpose. KNIME proposes Information gain and Gini impurity. Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset and Information gain is a measure based on entropy. We tried both and we chose Gini index because we got significant decision trees with a good accuracy grade.

$$GI(f) = \sum_{i \neq k}^{m} f_i f_k \tag{2}$$

C4.5 was the first test. The decision tree classifier is built in two phases: a growth phase and a prune phase. In the former, the tree is built by recursively partitioning the data until each partition is either "pure" (all members belong to the same class) or sufficiently small (a parameter set by the user). The splitting partition depends on the attribute that the algorithm select. For a continuous attribute, the splitting is made using a threshold. For nominal attribute instead, it creates n sets where n is the number of possible value that can takes the attribute. The second phase involves pruning about the generated tree. This pruning removes dependence on statistical noise or variation that may be particular only to the training set. In our case we tried, with pruning and without pruning. The error rate in average, was too high and we discarded. Therefore we tried the Random forest method. This is an ensemble method. It uses so many tree classifier and generate a forest of decision tree. Each time the algorithm selects a random partition and performs the classifier. We find that 25 partition is good trade-off based on accuracy and a good result. Since we got a high value of accuracy, we used this kind of classifier. After the learner generates the tree models, the prediction tries to guess the alcohol attribute. The guess uses the majority vote between all decision trees: this technique is also called bagging.

### Test

For test our result we use cross validation. In cross validation data are randomly partition into subset or folds. Training and testing are perform several time. One partition is reserved as the test set and remaining partition are collectively used for training the model figure 6. Each sample is used the same number of

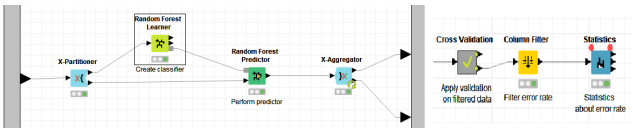


*Figure 6*. Cross validation & Testing Procedure using KNIME

times for training and once for testing. The estimate accuracy is the overall number of correct classifications from the iteration, divided by total number of tuples in the initial data. Random forest is used to classify data. To construct a decision tree classifier randomly select each node. Predictor provides new attribute of prediction with our target attribute. By using X-Aggregator we find the prediction table and error rates from cross validation. The data get from cross validation is filtered(Error rate) by column filter to measure the accuracy of our model. We use statistics to find and view the statistic value. The figure 6 show testing part of our model.

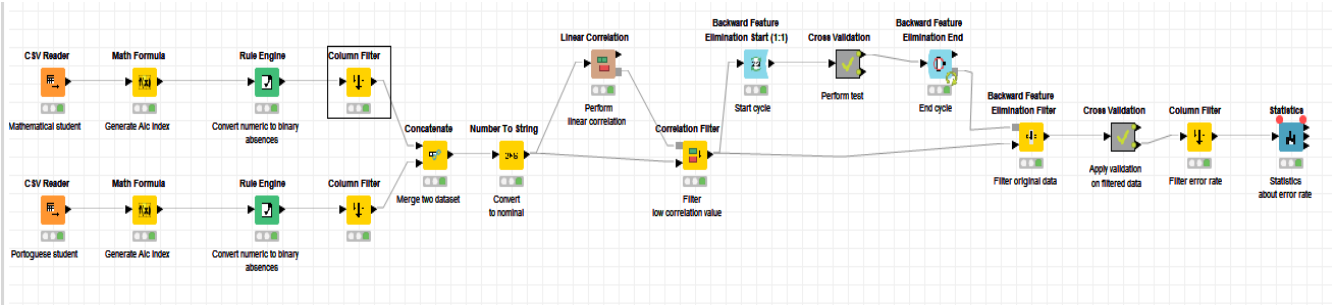


*Figure 7*. KNIME work flow

## Result

After complete the data preprocessing, we select Random Forest to get the best accuracy and got our results. We have 25(there are more tree available) trees that represent our drinker identikit. We take only those tree which have good value related to our prediction. From those trees, we found several common pattern, group the features and try to find how they are impact (we ignore some trees for significant value). The percentage of impact value are shown on by graph 8 and also by table 2.

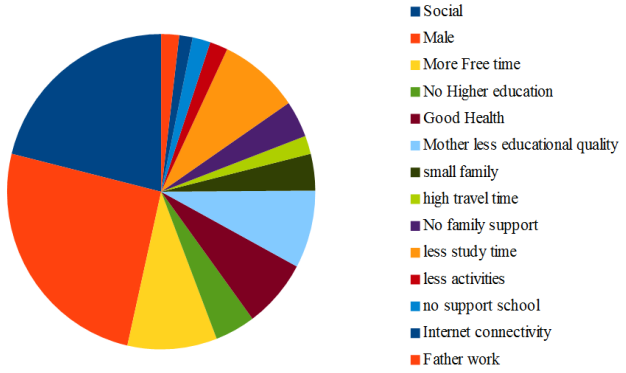


*Figure 8*. Attribute Show by graph

The percentage is calculate using the following equation:

$$\forall a \in A, \forall l \in \{1,2,3,4,5\} : C(T,l,a) * (6-i) \qquad (3)$$

A is the set of attributes, l represents the number of level, T is the set of tree, the C function counts the times that the attribute a is present in any trees in a specific level and 6-i represents the weight.
As we can see from the data graph and table the gender is one of most important feature for a drinker. Men are

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 most involved with alcohol. To confirm this,we also search on website and find some related work by the world health organization [10] . Which shows that male drink more than female in a 2014 alcohol report. Another important aspect of a alcoholics is the social activity. The person who goes out frequently with friends and relative take more alcohol. This is because drinking, became a way to celebrate a good news or in our society also to know people since usually they are more available when they are drunk. Some important characteristics are also: more free time and also less study time because usually they find in alcohol a way to relax and escape from problems). Since our main topic are teenagers, they will decide their future: university or work. According to our result, who won't frequent

university will drink more. This characteristic is also correlated somehow with the time spent for study. Factor that cannot be forget is the family. Support for study, big family size and a good father job are important for the child growth. Also the mother education is a fundamental stone in their life. Our study shows that, lack of the some characteristic have more chance to addict with alcohol for child.

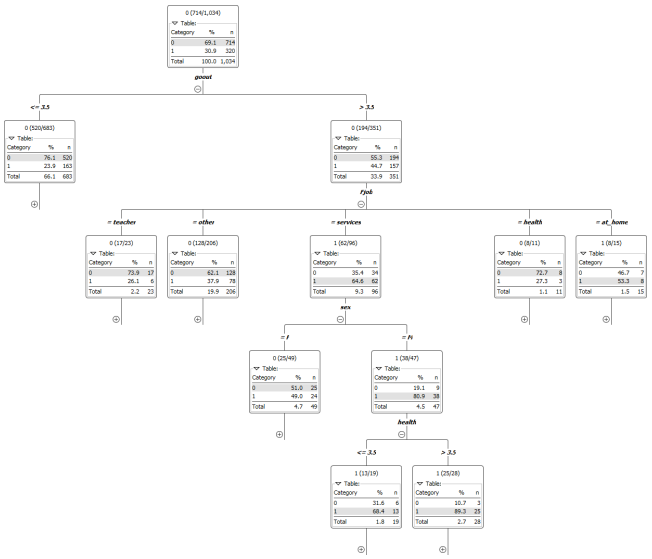

*Figure 9*. Decision Tree 1

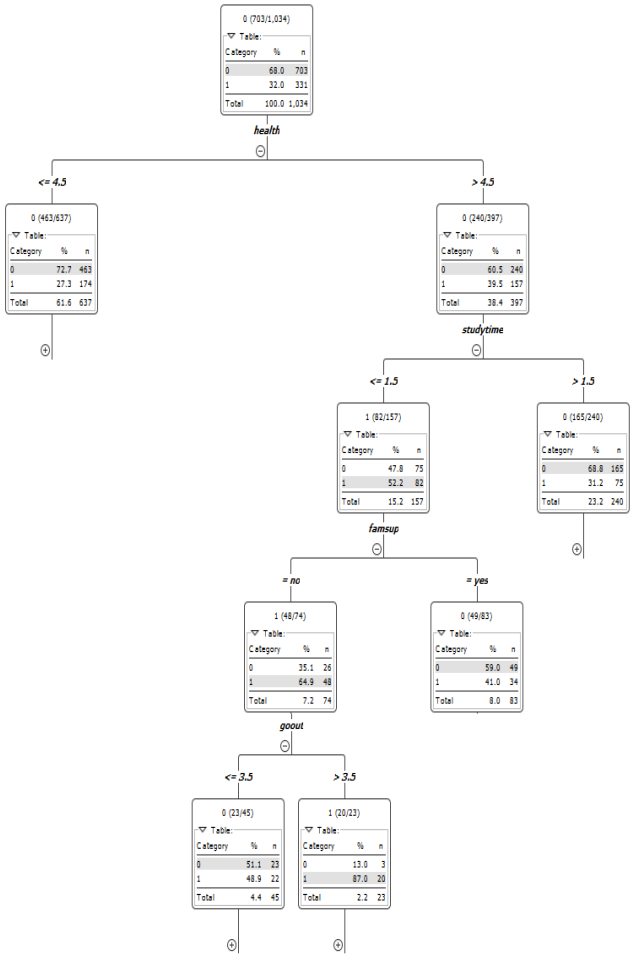

*Figure 10*. Decision Tree 2

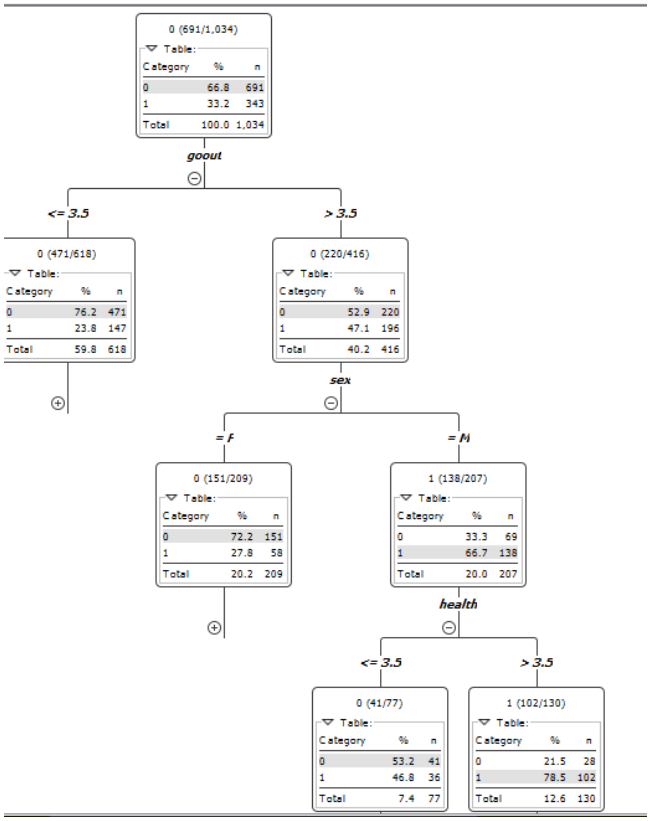


*Figure 11*. Decision Tree 3

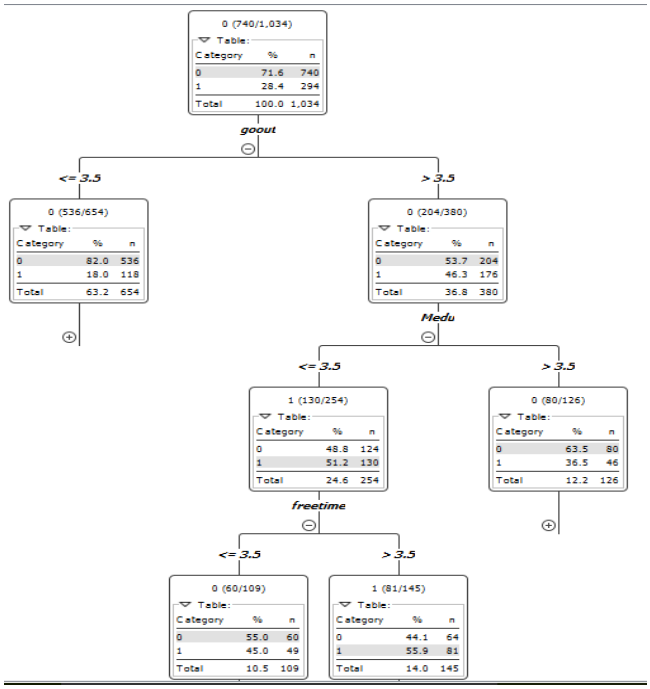


*Figure 12*. Decision Tree 4

Moreover less activities and high travel time between the school and the home are stress source, since this imply sedentary and no rest time for the person. Another problem coming in the last century, Internet. This can influences the guy in a negative way. Some combination between features can be risky and increase the possibility to have child addicted by alcohol. In figure 9, 10,11& 10 we saw some decision tree as image and can find the all the factor and rate of students are addicted by alcohol.

At this part of our result analysis, we use statistics formula to measure the error rate/ accuracy of our work. Below figure 13 also show mean and histogram of or result.

The average error rate is 8.018%. Which means our accuracy is almost 92%, that is acceptable. Also statistics results shows us, there is no missing value.

## Conclusion

Education is a crucial element in our society. Business Intelligence (BI)/Data Mining (DM) techniques, which allow a high level extraction of knowledge from raw data, offer interesting possibilities for the education domain. In particular, several studies have used BI/DM methods to reduced the alcohol addiction rate to teenagers and enhance lifestyle for child.

In this work, we have addressed the prediction of teenagers alcohol addiction by using past school records, demographic, family and other data related to student. Several DM goals and DM method were tested. Here we are have some limitations, like as number of folds we used is only 10. Because for large number folds, it takes more time and need powerful computer to process data. Also we takes only 25 models for test, because some model value is very lower. So we ignore them for best analysis. The obtained results reveal that it is possible to keep the child away from alcohol. This confirm the conclusion found that, child behavior is highly affected by friends or group. Nevertheless, an analysis to knowledge provided by the best predictive models has shown that, in some case, there are other relevant features, such as: school related, demographic and social variables.

Our work is based on offline study. All the techniques we used to a data-set collected by other people. However, there is a potential for an automatic on-line learning environment, by using Internet can find more data

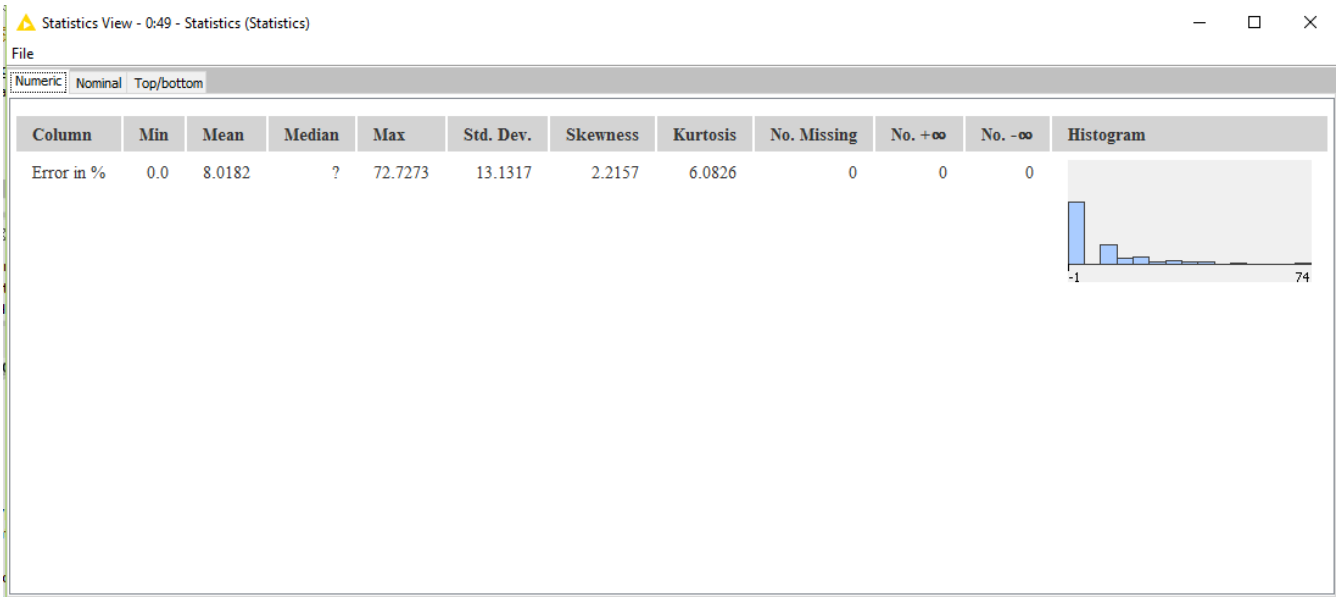


*Figure 13*. Statistics view:error rate, Mean, Histogram

about underage alcohol addiction. Moreover, we decide if someone want work more deeply on this topic and overcome the limitations of our work. We can provide some idea and our work file to them. More research and bigger data-set are also needed in order to understand why and how some variables affect on child's and they are addicted to alcohol.

## References

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

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

[3] National Institute on Alcohol Abuse and Alcohlism(NIH). Alcohol's effects on the body.

[4] 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 (*fubutec*2008) pp:5-12 porto portugal]. 2008.

[5] Breiman L. Friedman J. Ohlsen R. and Stone C. *classification and regression trees*. 1984.

[6] Leo Breiman. Random forests. 2001.

[7] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[8] Jiawei Han Micheline Kamber, Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, third edition.

[9] Geoffrey Holmes Bernhard Pfahringer Peter Reutemann Ian H. Witten Mark Hall, Eibe Frank. The weka data mining software: An update; sigkdd explorations, volume 11, issue 1., 2009.

[10] Com (*S rl*). Villars sous Yens. Global status report on alcohol and health. Technical report, World Health Organization, 2014.