"1st International Symposium on Computing in Informatics and Mathematics (ISCIM 2011)"
in Collaboration between EPOKA University and "Aleksandër Moisiu" University of Durrës
on June 2-4 2011, Tirana-Durres, ALBANIA

The Evaluation of Risk of Substance Abuse Among The Youth through Bayesian Classification

İhsan Ömür Bucak[1], Faruk Bulut[2]

[1]*Department of Computer Engineering, Mevlana University, Konya, Turkey*
[2]*Department of Computer Engineering, Fatih University, İstanbul, Turkey*

ABSTRACT

In recent years, the rising use of addictive drugs and substances has become one of the biggest social problems around the world. The illicit use of a variety of drugs appears to be increasing among elementary and high schools students in Turkey. Therefore, it can be said that there is a big rising risk for the youth: substance abuse and addiction. There are many reasons leading students to be an addicted user. At first an adolescent cannot see the bad sides and realize the harmful effects of the substances. After being a drug abuser, this person struggles with the addiction and his/her life gets worse. Scientific studies show that it becomes very difficult for a person to get rid of this habit after being a user. Hence, preventing students from being addicted becomes an important issue. The aim of this study is to determine a young person's probability of becoming a drug user in the future by means of Bayesian classification algorithm. The study is focused on informing the educators and families about the students who entertain high risk, and taking precautions and counter measures before it is too late. As data collection method, a questionnaire is asked the elementary and high school students in Büyükçekmece district of Istanbul and to the patients of substance abuse and addiction in the hospitals. The data collected from the questionnaires are used to indicate the percentage of risk probability for each student with the aid of Bayesian classification algorithm.

Keywords: Substance abuse, Substance addiction, Bayesian classification, Machine learning

1. INTRODUCTION

Substance abuse is a very common problem in the world. Recent studies conducted in Turkey show that there is a gradual increase in the use of illicit drugs among elementary and high school students. Even some teenagers at the age of 10 are becoming drug users [1]. It is obvious that it is getting a big problem for the Turkish youth. Because of that some precautions have to be taken to prevent the adolescents from using and addicting to drugs. The use of *Bally* (a glue brand) and *thinner* (a diluting agent) are very common in Turkey. Especially the poor and illiterate teenagers tend to consume it because they are very cheap and easy to find. These

inhalants are regarded as drugs by Research, Treatment and Training Centre for Alcohol and Substance Addiction (AMATEM in Turkish abbreviated form). In AMATEM, Research, Treatment and Training Centre for Volatile Substance Addiction (UMATEM in Turkish abbreviated form), and Research, Treatment and Training Centre for Child and Adolescent Substance Addiction (ÇEMATEM in Turkish abbreviated form) clinics, there are many inpatient teenagers who are addicted to those inhalants.

The following steps explain how the system has been put into work. All the steps are aimed to determine the risk ratio in percentage of a student:

1. Collecting the data with the questionnaire,
    a. From The Tepecik M. Özyeğin High School
    b. From The Büyükçekmece Elementary School
    c. From AMATEM, UMATEM, and ÇEMATEM.
2. Converting all the nominal data from the questionnaires to numeric values for the C programming language and Weka,
    a. To the file `"DB.txt"`
    b. To the array `"DB[][]".`
3. Building the model,
    a. Arranging the predefined training dataset,
    b. Preparing the model,
    c. Applying Naïve Bayes Classifier.
4. Implementing the algorithm in the C programming language,
5. Discussing the results, and making comparisons if necessary,
6. Reporting,
7. Informing the educators and families about the students who are high at risk and taking precautions.

1.1. Significance of family and education to keep adolescents clear of drugs

The anti-drug education that children are getting in school is one of the major precautions that can be taken. The education ought to be given by the specialists. Otherwise while giving information about drugs to the adolescents; they feel curious and passionate towards using these drugs.

Parents are the most important role models in children's lives. What they say and do about drugs matters significantly when it comes to the choices children make. Families are responsible for their children. Thus, they should pay attention to their children. For this purpose, they can educate themselves by reading relevant books and articles.

In today's complex and busy world, parents do not always find enough time to spend time with children to talk about drugs. However, it is necessary to ensure that the family has regular gatherings with their children and to schedule regular parent-child rituals and family meetings.

Rituals, like having meals together, playing games, going to the library together once a week can be opportunities to help the family catch-up and establish better and more open communication that is essential to raising drug-free children. Family meetings held once a week can also be extremely valuable. Another way to talk with children about drugs is to take advantage of everyday "teachable moments" [2]. Some other tips which can be useful for the family and educators can be found in the literature such as the one in [3].

2. RECORDING QUESTIONNAIRES TO THE DATABASE

After collecting more than 700 questionnaires from the schools and the hospital, they have been recorded to the database `db0.txt` one by one. All the answers in the questionnaire have a nominal value. The questions are asked the youth as based on the inclusion of the followings in the questionnaire in this order: Age, gender (two choices), parents being alive (four choices), the person stayed with and the place stayed at (six choices for this combination), Monthly income of the family (four choices), meeting or visiting frequency of relatives (four choices), playing an instrument (two choices), doing physical exercises (two choices), frequency of reading a book (four choices), kind of the music pleasing (five choices), frequency of going to movies (four choices), spent hours over the internet (three choices), the action taken in case of a friend's offer known to be harmful (three choices), which one of the parent cares in case of running into a problem (four choices), guessing own life in the future as compared to today (four choices), closest person to share in case of a problem (six choices), with whom leisure time is spent mostly (five choices), what to do when facing a tough and get-over situation (three choices), the feel of self-confidence (three choices) [4].

The first integer value is the ID number of the person filling out the questionnaire. ID numbers are used to keep the names secret for the purpose of hiding the personal information. The second one is the age value. For the third one which is gender, 1 means male; 2 means female. In a similar manner, all the answers to the questions are converted to integer values (See Table 1). In the database, -1 means there is a missing value. In other words, the corresponding question has not been answered in the questionnaire. In the 20th question, students are asked to write their five best friends. They have written the names yet it has been too difficult to find the ID numbers of the "*best friends*". For example, the student with the ID number 100 likes the students whose IDs are 96, 95, 93, 99 and 0 very much (the first row in the table). 0 in this section indicates void value. Therefore, he has written down only four best friends. The answer to the 20th question is very important in this study, because peer pressure plays a major role to lead a person to drug addiction [5]. The answers of the 21st, 22nd, 23rd, 24th and 25th questions have been taken from their teachers. For these questions, 1 and 0 means "yes" and "no" respectively. For example in the first row, the student with the ID number 100 has some misbehavior. The student with the ID number 101 is alcoholic. The student with the ID number 102 has a smoking habit. The student with the ID number 103 is definitely a drug user. The student with the ID number 120 is alcoholic, smoker, and drug user, even though his GPA is 5. The value of 100 which is at the end of this line indicates that he is 100% addicted.

Table 1. An extracted part from the `db.txt` file

```
100 12 1 1 3 1 1 1 1 1 2 1 2 1 4 5 1 3 2 2 96 95 93 99 0 4 1 0 0 0

101 13 2 1 3 2 1 0 -1 1 3 3 1 2 3 3 1 1 3 2 116 103 105 108 115 5 0 1 0 0

102 13 2 1 3 -1 1 1 1 1 3 2 1 2 3 3 1 3 2 3 117 110 114 116 0 5 0 0 1 0

103 13 2 1 3 -1 1 0 1 1 3 1 2 2 3 2 1 3 3 2 116 113 108 101 0 4 0 0 0 100

104 14 2 1 3 -1 2 0 0 4 3 4 1 3 4 4 6 5 2 3 0 0 0 0 0 4 0 0 0 0

105 13 2 1 3 2 1 1 1 1 3 2 2 1 3 3 1 3 3 2 116 101 108 114 0 3 0 0 0 0

106 13 1 1 3 1 1 0 0 1 3 3 2 2 3 3 2 3 3 2 102 116 0 0 0 -1 0 0 0 0

107 13 2 1 3 2 2 1 1 1 3 2 1 2 3 3 1 1 3 2 109 112 115 0 0 -1 0 0 0 0
```

In this table, there are 25 attributes of each tuple for Bayesian classification algorithm. In each line, the first value is the ID number; the following 19 integer values are the answers to the questions in the questionnaire. Then, the next five values are the IDs of best friends. The last 5 values are the private and individual information taken from the school.

3. PREPARING DATA

As mentioned previously, in the questionnaire, all the answers are represented with integer values ranging from 0 to 6. Only the age attribute is between 10 and 20. Because all the data is in that format, there is no need to calculate the *arithmetic mean* and the *standard deviation*.

Figure 1 below represents the steps of data preparation for further calculations in this study. `DB0.txt` is the database as a text file; `DB0[][]` is the database as a two dimensional array used in the implementations.

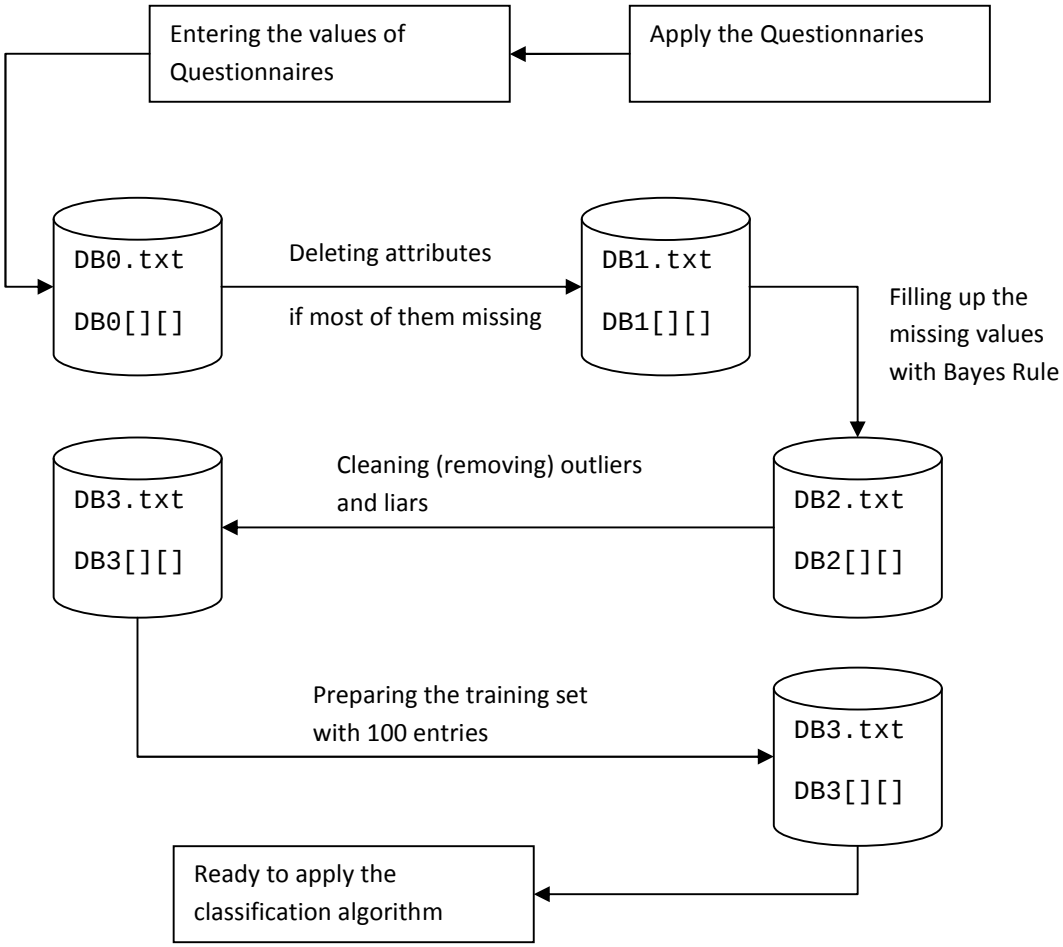

Figure 1. **Steps of data preparation from questionnaires.**

Some students have filled out the questionnaires randomly without reading. Some even intentionally have given wrong answers. These may have caused to lower the reliability of the study, as well. Therefore, some corrections and removals have been needed to be done to validate the study by the establishment of Sections 3.1 through 3.4 as the following:

3. 1. Removing Outliers

Erroneous and noisy data have been corrected and some of them even removed, whereas missing data must be supplied or predicted by using data mining tools. In the questionnaire, there are some entries that do not fit nicely into our derived model. These outliers should be eliminated before calculations. The 21st question in the questionnaire according to the answers is regarded as outlier on the basis of the answers given by the students and thus removed for the further steps.

3. 2. Removing Liars

Some students have filled out the questionnaires randomly without reading. Furthermore, some of them intentionally have given wrong answers. They have been regarded as *liars* and removed from the database to keep this study more reliable. There are some examples to wrong answers below:

- A primary school student has written down his age as 22.
- Someone has answered the 7th question "*Do you play any music instrument?*" as "*yes*". But in the 10th question "*Which kind of music do you like?*" he has given the answer "*I don't like music*". This is obviously a conflict. This record has been extracted from the database.
- In the 3rd question, "*Are your parents alive?*" a student has said "*No*". But later, in the 14th question "*If you have a problem, who are you talking with?*" he has said "*With both my mother and father.*" This is also a conflict. This entry has also been removed from the database.

Table 2. **The Pseudo code finding the liars.**

```
function FindLiar (int N, int DB[MAX][MAX])
declare i, int
begin
for i ← 1 to i<= N
        if DB[i][3]==4  AND  DB[i][4]==3  // Parents Liars
        makeliar(i)
        if DB[i][1]>20  OR  DB[i][1]<10        // Age Liars
        makeliar(i)
        if DB[i][3]==2  AND  DB[i][4]==2  // Parents Liars
        makeliar(i)
        if DB[i][3]==3  AND  DB[i][4]==1   // Parents Liars
        makeliar(i)
        if DB[i][3]==4  AND  (DB[i][4]==1 OR  DB[i][4]==2
        OR  DB[i][4]==3) // Parents Liars      makeliar(i)
        if DB[i][14]==1  AND  DB[i][3]==3  // Parents Liars
        makeliar(i)
        if DB[i][14]==2  AND  DB[i][3]==2  // Parents Liars
        makeliar(i)
        if DB[i][14]==3  AND  DB[i][3]!=1  // Parents Liars
        makeliar(i)
        if DB[i][7]==0  AND  DB[i][10]==5  // Music Liars
        makeliar(i)
        if DB[i][16]==1  AND  DB[i][3]==4  // Parents Liars
        makeliar(i)
        if DB[i][25]>=1  AND  DB[i][25]<=5  // GPA Liars
        makeliar(i)
end
```

The liars have been removed from the database to have more reliable, true and better results. The implemented codes generate an output text file "OutLiers ID List.txt" as below to show the liars.

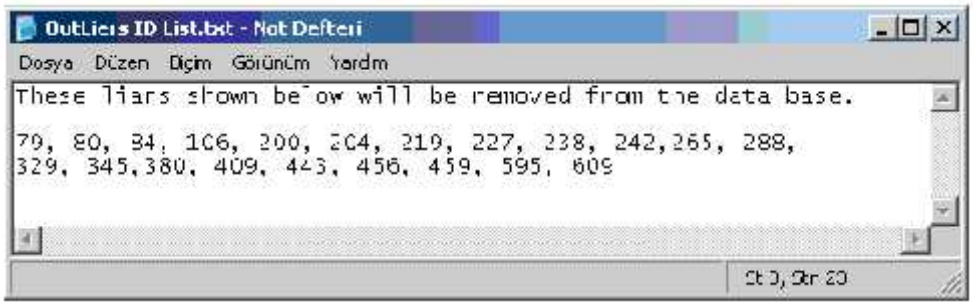

Figure 2. The "`OutLiers ID List.txt`" file to show the liars.

3.3. Missing Data

In the questionnaire, there are few questions that have not been answered by the students. While converting the questionnaires to numeric values, it is put as -1 to demonstrate the missing data in the database. The missing data is replaced with Bayesian Probability Technique although there are some other methods [6].

If there is a missing value of an attribute of a tuple, the whole probabilities of the attribute in all the tuples ought to be examined. The one which is the biggest gets the value that takes the place of the missing data. The Bayesian formula used in the codes is shown below:

$$P(h \mid D) = \frac{P(D \mid h).P(h)}{P(D)} \qquad (1)$$

Some outputs that show the frequencies of each attribute are given in Figure 3. For example, in the first column, there are 4 students at the age of 11. There are 95 students at the age of 12. In the second column, it is presented that 320 students are male and 350 students are female.

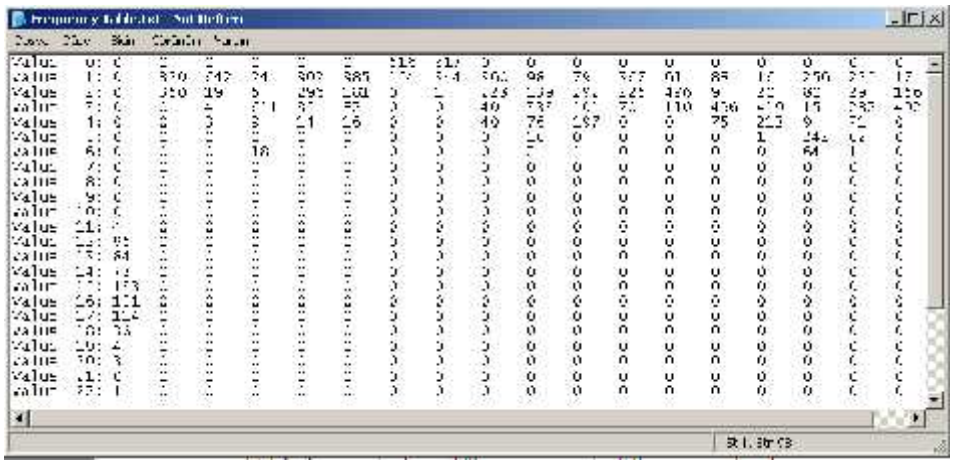

Figure 3. The "`Frequency Table.txt`" showing the frequencies of each attribute.

3. 4. Deleting Attributes

In this step, some attributes (columns) are deleted. If missing values of an attribute are in majority, the attribute is canceled and not contaminated to the calculations. For example, only a few students have answered the question which is asking their GPA. At least 1/3 is missing. Therefore, this column is canceled and not added to the calculations. In the database, the GPA

attribute has been changed with the value of -2. In the implementation, -2 is regarded as a void attribute.

Omitting some attributes in the `DB0[][]` and converting it to `DB1[][]` and `DB1.txt` is performed in further steps.

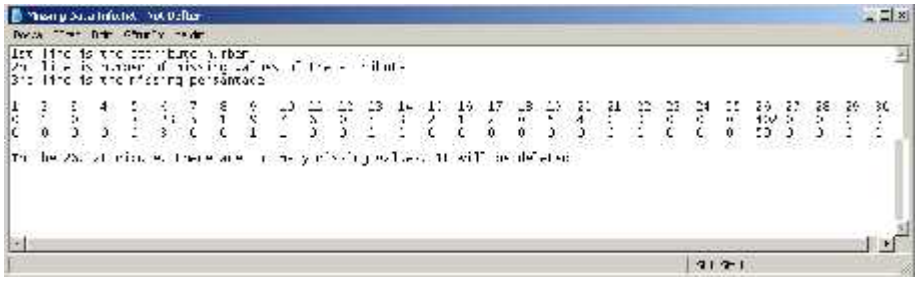

Figure 4. The "`Missing Data Info.txt`" file showing the missing values.

4. PREDEFINED TRAINING DATASET

For classification problems, training dataset is obligatory. All approaches to performing classification require some predefined training dataset. A training set is used to develop the specific parameters required by the technique. Training data consist of sample input and output data as well as the classification assignment for the data [7]. The training dataset should be defined well before applying the algorithms.

In this study, there are 110 tuples in the predefined set. 37 of them are absolutely 100% drug users. 35 -out of 37- are from ÇEMATEM clinic. 2 -out of 37- are from M. Hüsnü Özyeğin Tepecik High School. 63 students are chosen by the Psychological Counseling and Guidance (PDR in Turkish common usage) Department. 43 of them are at risk ratios ranging from 10% to 80%. 20 students are at the risk ratio %0.

The Department of Psychology at Fatih University, and the Psychological Counseling and Guidance Unit at the High school has provided us with valuable insights throughout the preparation of the "*predefined training dataset*".

5. BAYESIAN CLASSIFICATION

Bayesian classification is based on Bayes Rule of conditional probability. In this method, each of the attributes has a part in calculation. Bayes Rule will be defined before examining the steps of the classifier algorithm.

5.1. Bayes Rule (Bayes Theorem)

In classification problems and machine learning, determining the best hypothesis (the most probable hypothesis) from some space *H* given the training data *D* is very important. Bayes Rule provides a way to calculate the probability of a hypothesis by this formula:

$$P(h \mid D) = \frac{P(D \mid h).P(h)}{P(D)}, \qquad (2)$$

where $P(h)$ denotes the initial probability that hypothesis *h* holds before observing the training data. $P(h)$ is usually called the prior probability of *h* and may reflect any background knowledge we have about the chance that *h* is a correct hypothesis. If there is no such prior knowledge, then the same prior probability might be simply assigned to each candidate hypothesis. Similarly, $P(D)$ denotes the prior probability that the training data *D* will be observed (i.e., the probability of *D* given no knowledge about which hypothesis holds). $P(D \mid h)$ denotes the probability of observing data *D* given some world in which hypothesis *h* holds.

More generally, writing $P(x \mid y)$ denotes the probability of *x* given *y*. In Machine learning and classification problems, we are interested in the probability $P(h \mid D)$ that *h* holds given the observed training data *D*. $P(h \mid D)$ is the posterior probability of *h*, because it shows our confidence that *h* holds after we have seen the training data *D*. It is important that the posterior probability $P(h \mid D)$ reflects the influence of the training data *D*, in contrast to the prior probability $P(h)$, which is independent of *D*.

Bayes theorem is the development of Bayesian learning methods because it enables a way to calculate the posterior probability $P(h \mid D)$, from the prior probability $P(h)$, together with $P(D)$ and $P(h \mid D)$ [8].

There is an example to the posterior probability. In our experiments we have found that 94.4% of the drug users are smokers. On the other hand 8.3% of the smokers are drug users. *P* is the probability, *S* is the smokers, and *D* is the drug users in the set. In this case, we can write down $P(S \mid D) = 0.944$ and $P(D \mid S) = 0.083$.

5.2. Bayesian Classifier

It is also known as Naïve Bayes Classifier. In this technique, the contributions of the whole attributes are independent and each contribution is held equally to the classification problem. This method has a big relation to the conditional probability, the Bayes Rule.

By analyzing the contribution of each "independent" attribute, a conditional probability is determined at the beginning. A classification is done by combining the impact that the different attributes have on the prediction. This approach is called "Naive" because it assumes the independence between the various attribute values. Given a data value x_i the probability that a

related tuple, t_i, is in class C_j is described by $P(C_j \mid x_i)$. Training data can be used to determine $P(x_i)$, $P(x_i \mid C_j)$, and $P(C_j)$.

For each attribute x_i, the number of occurrences of each attribute value x_i can be counted to determine $P(x_i)$. Similarly, the probability $P(x_i \mid C_j)$ can be estimated by counting how often each value occurs in the class in the training data. When classifying a new record, the conditional and prior probabilities generated from the training set are used to make prediction. This is done by combining the values of the different attributes from the tuple. Tuple t_i has p independent attribute values $\{ x_{i1} , x_{i2} , x_{i3} , \ldots. x_{ip} \}$. It is known that $P(x_{ik} \mid C_j)$ will be calculated for each class C_j and attribute x_{ik}. And finally it is easy to estimate $P(t_i \mid C_j)$ [9]:

$$P(t_i \mid C_j) = \prod_{k=1}^{p} P(x_{ik} \mid C_j). \qquad (3)$$

To calculate $P(t_i)$, it is easy to find the likelihood that t_i is in each class. The probability that t_i is in a class is the product of the conditional probabilities for each attribute values. Therefore, it gets easy to classify the new tuple as the highest probability of all.

As an example to the Bayesian classification, Table 3 shows a predefined training dataset including 14 tuples where some entries of this dataset have been selectively picked and some attributes eliminated to simply improve the understanding by bringing plainness into the scene. The risk values in the table have been determined together with The Department of Psychology and the Psychological Counseling and Guidance experts who have provided us with some valuable suggestions while preparing the "*predefined training dataset*". The new coming tuples will be put into a proper class according to the 14 tuples of the set by using the Bayesian classification algorithm. In this dataset, there are five attributes as the inputs. These are age, gender, family income, sports activity, and smoking habit.

The risk groups have been classified into five classes as follows: Class 1 (C_1) is the first 20 percent slice from 0% up to 20% of the whole risk and carries "Very low risk". Class 2 (C_2) is the second 20 percent slice from 20% up to 40% of the entire risk and carries "Low risk". Class 3 (C_3) is the third 20 percent slice from 40% up to 60% of the entire risk and carries "Normal risk". Class 4 (C_4) is the fourth 20 percent slice from 60% up to 80% of the entire risk and carries "High risk". Finally, Class 5 (C_5) is the remaining 20 percent slice from 80% up to 100% of the entire risk and carries "Very high risk".

Table 3. **A simple training data example for Bayesian Classifier.**

ID	**Age**	(1=	(1=p nal,	**Spo ivity** (1=y	**Smo abit** (1=y	**Output**

						Risk value	**Class (C_i)**
1	14	1	1	0	0	% 35	Class 2
2	14	2	4	0	1	% 75	Class 4
3	13	1	3	1	0	% 25	Class 2
4	15	1	2	0	0	% 30	Class 2
5	16	1	2	0	0	% 30	Class 3
6	17	2	2	1	0	% 0	Class 1
7	14	1	3	1	0	% 0	Class 1
8	12	2	1	0	0	% 10	Class 1
9	14	1	1	0	1	% 100	Class 5
10	15	2	4	0	1	% 100	Class 5
11	17	1	4	1	0	% 45	Class 3
12	17	1	1	0	0	%15	Class 1
13	14	1	1	1	1	%65	Class 4
14	15	1	4	0	1	%70	Class 4

Given the training set, we can compute the probabilities:

$P(C_1) = 4/14 = 0.286$ (C_i means Class i, e.g., C_1 means Class 1)
$P(C_2) = 3/14 = 0.214$
$P(C_3) = 2/14 = 0.143$
$P(C_4) = 3/14 = 0.214$
$P(C_5) = 2/14 = 0.143$

Age Attribute Probabilities (Group1 (Ages=12-14), Group2 (Ages=15-18)):

$P(1|C_1)=2/4$ $P(1|C_2)=2/3$ $P(1|C_3)=0$ $P(1|C_4)=2/3$ $P(1|C_5)=1/2$
$P(0|C_1)=2/4$ $P(0|C_2)=1/3$ $P(0|C_3)=2/2$ $P(0|C_4)=1/3$ $P(0|C_5)=1/2$

Gender Attribute Probabilities (1=male, 0=female):

$P(1|C_1)=2/4$ $P(1|C_2)=3/3$ $P(1|C_3)=2/2$ $P(1|C_4)=2/3$ $P(1|C_5)=1/2$
$P(0|C_1)=2/4$ $P(0|C_2)=0$ $P(0|C_3)=0$ $P(0|C_4)=1/3$ $P(0|C_5)=1/2$

Family Income Attribute Probabilities (1=poor, 2=normal, 3=good, 4=rich):

$P(1|C_1)=2/4$ $P(1|C_2)=1/3$ $P(1|C_3)=0$ $P(1|C_4)=1/3$ $P(1|C_5)=1/2$
$P(2|C_1)=1/4$ $P(2|C_2)=1/3$ $P(2|C_3)=1/2$ $P(2|C_4)=0$ $P(2|C_5)=0$
$P(3|C_1)=1/4$ $P(3|C_2)=1/3$ $P(3|C_3)=0$ $P(3|C_4)=0$ $P(3|C_5)=0$
$P(4|C_1)=0$ $P(4|C_2)=0$ $P(4|C_3)=1/2$ $P(4|C_4)=2/3$ $P(4|C_5)=1/2$

Sport Activity Attribute Probabilities (1=yes, 0=no):

$P(1|C_1)=2/4$ $P(1|C_2)=1/3$ $P(1|C_3)=1/2$ $P(1|C_4)=1/3$ $P(1|C_5)=0$
$P(0|C_1)=2/4$ $P(0|C_2)=2/3$ $P(0|C_3)=1/2$ $P(0|C_4)=2/3$ $P(0|C_5)=2/2$

Smoking Habit Attribute Probabilities (1=yes, 0=no):

$P(1|C_1)=0$ $P(1|C_2)=0$ $P(1|C_3)=0$ $P(1|C_4)=3/3$ $P(1|C_5)=2/2$
$P(0|C_1)=4/4$ $P(0|C_2)=3/3$ $P(0|C_3)=2/2$ $P(0|C_4)=0$ $P(0|C_5)=0$

Table 4. **Probabilities of the class attributes.**

	C_1	C_2	C_3	C_4	C_5
Age					
Group1	2/4	2/3	0	2/3	1/2
Group2	2/4	1/3	2/2	1/3	1/2
Gender					
Male	2/4	3/3	2/2	2/3	1/2
Female	2/4	0	0	1/3	1/2
Family Income					
Poor	2/4	1/3	0	1/3	1/2
Normal	1/4	1/3	1/2	0	0
Good	1/4	1/3	0	0	0
Rich	0	0	1/2	2/3	1/2
Sport Activity					
Yes	2/4	1/3	1/2	1/3	0

No	2/4	2/3	1/2	2/3	2/2

Smoking Habit					
Yes	0	0	0	3/3	2/2
No	4/4	3/3	2/2	0	0

For example *x* and *y* are the new records waiting to be classified:

x = < Age=17, Gender=Male, Family Income=poor, Sport=No, Smoking=Yes >
y = < Age=13, Gender=Female, Family Income=Rich, Sport=No, Smoking=No >.

In order to classify *x* and *y*, the following equation is used:

$$P(t_i \mid C_j) = \prod_{k=1}^{p} P(x_{ik} \mid C_j) \qquad (4)$$

Calculations for the tuple *x* :

$P(x|C_1)=P(\text{Age=17}|C_1).P(\text{Male}|C_1).P(\text{Poor}|C_1).P(\text{Sport=No}|C_1).P(\text{Smoke=Yes}|C_1)$
$P(x|C_1)=(2/4).(2/4).(2/4).(2/4).(0)=0$
$P(x|C_2)=P(\text{Age=17}|C_2).P(\text{Male}|C_2).P(\text{Poor}|C_2).P(\text{Sport= No}|C_2).P(\text{Smoke=Yes}|C_2)$
$P(x|C_2)=(1/3).(3/3).(1/3).(2/3).(0)=0$
$P(x|C_3)=P(\text{Age=17}|C_3).P(\text{Male}|C_3).P(\text{Poor}|C_3).P(\text{Sport= No}|C_3).P(\text{Smoke=Yes}|C_3)$
$P(x|C_3)=(2/2).(2/2).(0).(1/2).(0)=0$
$P(x|C_4)=P(\text{Age=17}|C_4).P(\text{Male}|C_4).P(\text{Poor}|C_4).P(\text{Sport= No}|C_4).P(\text{Smoke=Yes}|C_4)$
$P(x|C_4)=(1/3).(2/3).(1/3).(2/3).(3/3)=0.049$
$P(x|C_5)=P(\text{Age=17}|C_5).P(\text{Male}|C_5).P(\text{Poor}|C_5).P(\text{Sport= No}|C_5).P(\text{Smoke=Yes}|C_5)$
$P(x|C_5)=(1/2).(1/2).(1/2).(2/2).(2/2)=0.125$

The result of the probability $P(x|C_5)$ is the biggest one of all. Therefore the *x* record will be put into the Class 5.

Calculations for the tuple *y*:

$P(y|C_1)=P(\text{Age=13}|C_1).P(\text{Female}|C_1).P(\text{Rich}|C_1).P(\text{Sport=No}|C_1).P(\text{Smoke=No}|C_1)$
$P(y|C_1)=(2/4).(2/4).(0).(2/4).(4/4)=0$
$P(y|C_2)=P(\text{Age=13}|C_2).P(\text{Female}|C_2).P(\text{Rich}|C_2).P(\text{Sport= No}|C_2).P(\text{Smoke=No}|C_2)$
$P(y|C_2)=(2/3).(0).(0).(2/3).(3/3)=0$
$P(y|C_3)=P(\text{Age=13}|C_3).P(\text{Female}|C_3).P(\text{Rich}|C_3).P(\text{Sport= No}|C_3).P(\text{Smoke=No}|C_3)$
$P(y|C_3)=(0).(0).(1/2).(1/2).(2/2)=0$
$P(y|C_4)=P(\text{Age=13}|C_4).P(\text{Female}|C_4).P(\text{Rich}|C_4).P(\text{Sport= No}|C_4).P(\text{Smoke=No}|C_4)$

$P(y|C_4)=(2/3).(1/3).(2/3).(2/3).(0)=0$
$P(y|C_5)=P(\text{Age}=13|C_5).P(\text{Female}|C_5).P(\text{Rich}|C_5).P(\text{Sport= No}|C_5).P(\text{Smoke=No}|C_5)$
$P(y|C_5)=(1/2).(1/2).(1/2).(2/2).(0)=0$

There is an interesting situation above; the result of all the probabilities is equal to zero. At that point classifying the tuple *y* becomes impossible.

As it is seen above, this classification method gives only the class name as an output. Sometimes the probabilities of some classes are computed as zero. Zero probabilities are always painful for the Naïve Bayes Classifier. Thus, this makes a barrier to predict the class of the new records. However, in this study, this weakness of the Bayesian classifier has never been met since the training set was rich enough. Correctly classified instances have been found 77.27% in this method.

6. EVALUATIONS AND CONCLUSIONS

The C programming language has been used on the Windows environment in order to prepare the data collected from questionnaires. Devcpp v4.9 is used as an ANSI C Compiler. After recording all the questionnaires to the database `db0.txt`, the output files as text formats have been created. The Weka v3.6.0. program has also been used so as to check the results and measure the accuracy of the algorithm. Programming in C gives us a flexible opportunity and scalability. This could be very difficult in Weka or in another program.

The most relevant attributes in the training set can be found by Weka. Attribute selection is done for the purpose of finding which subset of attributes works best for prediction. This is done by searching through all possible combinations of attributes in the training data. *Information Gain Attribute Evaluation* method is used as an attribute evaluator.

As a result of the evaluation through Weka program, the accuracy of the Naïve Bayes classifier has been obtained 77.2% as correctly classified instances and 23.8% as incorrectly classified instances. As it is shown the accuracy of this classifier seems not high enough. The low accuracy in these calculations is related with the high number of output classes in the dataset. As there are five types of output classes, the accuracy decreases. As an example, let's assume there are two tuples whose risk percentages are 39% and 41% respectively. The first tuple is placed into Class2, the second one is put into Class3 although there is a 2% differences between them.

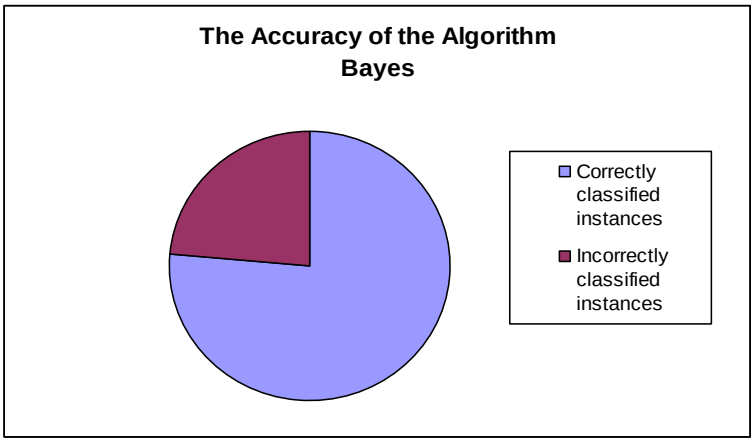

Figure 5. The accuracy of Bayes Classifier Algorithm.

The following results have been obtained after the evaluation of the questionnaries asked more than 700 people from the schools and the hospital:

	Number of the tuples corresponding to the classes	Percentage of the tuples for each class
Class 1	369	56,8
Class 2	96	14,8
Class 3	105	16,2
Class 4	27	4,2
Class 5	53	8,2

Figure 6. The number and percentage of the tuples for each class.

Naive Bayes can be used for both binary and multiclass classification problems. It is easy to implement, fast in running, and highly scalable model building and scoring. But if the training set is not rich enough, some probabilities of the attributes in the training set can be zero. This is a big barrier to the calculations. Naïve Bayes Classifier has produced 23.80% incorrectly classified outputs in the experiments. In other words, 23.80% of the students are put into the wrong class. It could not give the satisfactory results in the implementations, because it holds all the attributes equally. But some attributes has higher effect than others. For example, *smoking* attribute has a higher effect than *going to the cinema occasionally* attribute.

Additionally, this method requires a well-prepared maybe the best-prepared training data. All the probabilities and estimations in this technique depend on the training dataset. In this study the number of tuples in the training set is 110. Larger training datasets would yield better outputs. There should be more than 2000 tuples or different entries, hence different risk values, in the predefined training dataset so as to get satisfactory and feasible results. 110 tuples of the training dataset varies at different risk ratios from 0% to 100%. If the conditional probability used in the Naïve Bayes Classifier is zero, the new records cannot easily be classified. Therefore, the tuple variety is essential for the training dataset.

This study addresses a very common problem of the youth which is drug dependency, and applies Naïve Bayes Classifier to propose a feasible solution to the students at high risk of drug dependence. The results seem promising to further the study to propose solid solutions as part of taking urgent steps and precautions for such a challenging issue, and prevent the youth from being addicted to these drugs and substances.

REFERENCES

[1] Hürriyet Newspaper, (2010) published in Turkey, The "Gündem" page, retrieved in 03/24/2010, www.hurriyet.com.tr , Turkey

[2] Gürol D.T., (2007) *Uyuşturucu/Uyarıcı Maddeler ve Önleme, Öğretmen için Kitapçık*, ÇEMATEM Yayınları, Bakıköy Ruh ve Sinir Hastalıkları Hastanesi, İstanbul, p.14, pp. 15-38, Turkey

[3] US Government, (2010) *Safe and Drug Free Schools, What Can Parents Do to Help Their Children Be Drug Free,* Washington, retrieved in 07/12/2010. http://www.yic.gov/drugfree/whatparent.html, USA

[4] Bulut, F., (2010) *Detecting Students at Risk of Substance Abuse by Using Data Mining Classification Algorithms*, M.Sc. Thesis, Fatih University, Graduate Institute of Science and Engineering, Istanbul

[5] Windle. M. & Windle. (2003) R.C., *Alchohol and other substance use and abuse*. In G.R. Adams & M.D.Berzonsky (ed). *Blackwell handbook of adolescence*. Malden. MA: Blackwell Publishing, USA

[6] Dunham, M. H., (2005a) *Data Mining Introductory and Advanced Topics*, Prentice Hall Publishing, ISBN: 0130888923, p.15, pp. 77-78, USA

[7] Dunham, M. H., (2005b) *Data Mining Introductory and Advanced Topics*, Prentice Hall Publishing, ISBN: 0130888923, p76, USA

[8] Mitchell, T.M., (1997) *Machine Learning*, McGraw-Hill, ISBN: 0070428077, pp. 156-159, USA

[9] Dunham, M. H., (2005) *Data Mining Introductory and Advanced Topics*, Prentice Hall Publishing, ISBN: 0130888923, p.87, USA

Equations

$$P(h \mid D) = \frac{P(D \mid h).P(h)}{P(D)} \quad (1)$$

$$P(h \mid D) = \frac{P(D \mid h).P(h)}{P(D)}, \quad (2)$$

$$P(t_i \mid C_j) = \prod_{k=1}^{p} P(x_{ik} \mid C_j). \quad (3)$$

$$P(t_i \mid C_j) = \prod_{k=1}^{p} P(x_{ik} \mid C_j) \quad (4)$$