# DETECTING STUDENTS AT RISK OF SUBSTANCE ABUSE BY USING DATA MINING CLASSIFICATION ALGORITHMS

by

Faruk BULUT

August 2010

# DETECTING STUDENTS AT RISK OF SUBSTANCE ABUSE BY USING DATA MINING CLASSIFICATION ALGORITHMS

By

Faruk BULUT

A thesis submitted to

the Graduate Institute of Sciences and Engineering

Of

Fatih University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

August 2010
Istanbul, Turkey

# APPROVAL PAGE

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

________________________
Assist. Prof. Tuğrul YANIK
Head of Department

This is to certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

________________________
Assist. Prof. İhsan Ömür BUCAK
Supervisor

## Examining Committee Members

Assist. Prof. İhsan Ömür BUCAK ______________________

Assist. Prof. Tuğrul YANIK ______________________

Assist. Prof. Mustafa PETEK ______________________

It is approved that this thesis has been written in compliance with the formatting rules laid down by the Graduate Institute of Sciences and Engineering.

________________________
Assoc Prof. Nurullah ARSLAN
Director

August 2010

# DETECTING STUDENTS AT RISK OF SUBSTANCE ABUSE BY USING DATA MINING CLASSIFICATION ALGORITHMS

Faruk BULUT

M. S. Thesis - Computer Engineering
August 2010

Supervisor: Assist. Prof. İhsan Ömür BUCAK

## ABSTRACT

In the recent years, the rising use of addictive drugs and substances has become one of the biggest social problems in 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.

This thesis focuses on *urgent precaution systems* which help families and educators prevent students from addiction. The aim of the study is to detect the probability of being a drug user in the future by using classification algorithms in

Machine Learning, Data Mining and Pattern Recognition. As data collection method, a questionnaire including 25 questions is applied to the elementary and high school students in Büyükçekmece and to the patients in ÇEMATEM. The data collected from the questionnaires is used to indicate the percentage of risk probability for each student with the help of classification algorithms.

The findings of the study show that if there is a computed high risk for a student, some precautions should be taken to keep this person away from substances. These precautions taken by the educators and parents can be to inform the student about the substances and to lead him/her to social activities.

**Keywords:** Substance abuse, drug dependence, drug addiction, data mining, classification algorithms**.**

# MADDE BAĞIMLISI OLMA RİSKİ ALTINDA OLAN ÖĞRENCİLERİN VERİ MADENCİLİĞİ SINIFLANDIRMA ALGORİTMALARIYLA TESPİT EDİLMESİ

Faruk BULUT

Yüksek Lisans Tezi - Bilgisayar Mühendisliği
Ağustos 2010

Danışman: Yrd. Doç. Dr. İhsan Ömür BUCAK

## ÖZ

Son yıllarda dünyada bağımlılık yapıcı maddelerin kullanımı büyük bir sosyal problem haline gelmiştir. Yasal olmayan bu maddelerin kullanımı Türkiye'deki ilköğretim ve lise öğrencileri arasında da yaygınlaşmaktadır. Bundan ötürü öğrenciler için büyük bir tehlikenin varlığından söz edilebilir: madde bağımlılığı.

Öğrencileri madde bağımlılığına iten birçok neden vardır. İlk başlarda genç bir insan zararlı maddelerin kötü yönlerini göremeyebilir ve zararlı etkilerinin farkına varamayabilir. Bağımlı olduktan sonra bağımlı kişi bu hastalıkla boğuşmakta ve hayatı daha da kötüye gitmektedir. Bilimsel çalışmalar bağımlılıktan sonra bu kötü alışkanlıktan kurtulmanın zor olduğunu göstermektedir. Bu nedenle öğrencileri bu bağımlılıktan korumak önemli bir konu haline gelmiştir.

Bu tez, öğrencileri madde bağımlılığından korumak için ailelere ve eğitimcilere yardımcı olacak bir *erken uyarı sistemi* üzerine odaklanmıştır. Bu çalışmanın amacı gelecekte bir öğrencinin madde bağımlısı olma riskini Veri Madenciliği, Makine Öğrenmesi ve Örüntü Tanıma bilim dallarında bulunan sınıflandırma algoritmaları yardımıyla hesaplamaya yöneliktir. Veri toplama metodu olarak 25 soruluk bir anket Büyükçekmece'deki bir ilköğretim okuluna, bir liseye ve ayrıca ÇEMATEM hastalarına yapıldı. Her bir öğrenci için risk yüzdesi anketlerden toplanan verilerin sınıflandırma algoritmalarında kullanılmasıyla bulundu.

Bu çalışmada bir öğrenci için yüksek bir risk değeri hesaplandıysa, bu kişi için madde bağımlılığını engelleyici bazı önlemler alınmalıdır. Eğitimciler ve ebeveynlerin alacağı bu önlemler, kendisinin bu konuda bilgilendirilmesi olabileceği gibi sosyal aktivitelere yönlendirilmesi de olabilir.

**Anahtar Kelimeler:** Uyuşturucu alışkanlığı, madde bağımlılığı, bağımlılık, veri madenciliği, sınıflandırma algoritmaları.

# ACKNOWLEDGEMENT

I would like to express my deepest appreciations to my advisor Assist. Prof. İhsan Ömür BUCAK for his guidance, valuable suggestions and creative comments throughout the research. It would not be possible to complete this thesis without his persistent help and patience. I consider myself very fortunate for being able to work with a very considerate and encouraging advisor like him.

I also want to show my sincere gratitude to the computer engineering faculty members for their lectures I have taken during my master education. I owe many thanks to Assist. Prof. Tuğrul YANIK, Assist. Prof. Zeynep ORHAN, Assist. Prof. Atakan KURT, Assist. Prof. Veli HAKKOYMAZ, Assist. Prof. Nahit EMANET and others.

Assist. Prof. İdil IŞIK, the director of the psychology department of Fatih University, has given me some valuable suggestions while preparing the "*predefined training dataset*". She has made some determinations on giving big effects to some particular attributes in calculations. I owe many thanks to her for her helps and support.

I am also grateful to those below for the support and help they have given during my research.

- Doç.Dr. E.Cüneyt EVREN, Klinik Şefi, Bakırköy AMATEM
- Ayşe ÜNLÜ ÇİLELİGİL, Baş Hemşire, Bakırköy ÇEMATEM
- Musa ÇALIŞKAN, Müdür, Tepecik Hüsnü M.Özyeğin Lisesi
- M.Enes TAN, PDR Öğretmeni, Tepecik Hüsnü M.Özyeğin Lisesi
- M.Vedat ARABACI, Müdür, Büyükçekmece İlköğretim Okulu
- Mehmet COŞKUN, Md.Yrd., Büyükçekmece İlköğretim Okulu
- Birol TİLKİ, MYO Öğr.Görevlisi, Fatih Üniversitesi

- Zuhal KAŞMER, Öğr.Görevlisi, Yabancı Diller Yüksek Okulu, Marmara Üniversitesi
- Büyükçekmece Kaymakamlığı
- Büyükçekmece İlçe Milli Eğitim Müdürlüğü
- İstanbul İl Sağlık Müdürlüğü

In addition, I express my thanks and appreciation to my dear family members Şule Zeliha, Betül, Mehmet Ali and Ahmet Safa for their understanding, motivation and patience. I dedicate this study to them.

Last but not least, I am thankful to all colleagues and friends who made my stay at the university a memorable and valuable experience.

# TABLE OF CONTENTS

# LIST OF TABLES

**TABLE**

# LIST OF FIGURES

**FIGURE**

# LIST OF SYSMBOLS AND ABBREVIATIONS

| SYMBOL/ABBREVIATION | |
|---|---|
| 1R | One Attribute Rule, oneR |
| ANSI | American National Standard Institute |
| AMATEM | Alkol ve Madde Bağımlıları Tedavi Merkezi, (Research, Treatment and Training Centre for Alcohol and Substance Addiction) |
| BFS | Breath First Search |
| ÇEMATEM | Çocuk ve Ergen Alkol ve Madde Bağımlıları Tedavi Merkezi |
| DB | Data Base |
| DT | Decision Tree |
| GPA | Great Point Average |
| ID | Identification |
| kNN | k Nearest Neighbors |
| PDR | Psikolojik Danışma ve Rehberlik |
| TUBİM | Türkiye Uyuşturucu ve Uyuşturucu Bağımlılığı İzleme Merkezi (Turkish Monitoring Centre for Drugs and Drug Addiction) |
| UMATEM | Uçucu Madde Araştırma ve Tedavi Merkezi |

# CHAPTER 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 (Hürriyet, 2010). 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 drugs. This thesis aims to perform this task.

Before discussing the precautions to be taken, drug types, specifications and effects of them should be examined.

## 1.1. Drug Types, Their Features and Effects

Several types of drugs are susceptible to abuse by the youth. These drugs range from most common and less expensive such as cigarettes and alcohol to expensive and more deadly such as cocaine and heroin.

There is a list of the most common abused drugs and substances below (US Government, 2010), (Roberts, 2010), (Drugfree, 2010), (Listverse, 2010). Their names, types, features, ways of consumption, and their effects on people are listed below.

Drug Name: **Amphetamines**
Drug Type: Stimulant
Facts: Chronic use can induce psychosis with symptoms.
How Consumed: Orally, injected, snorted, or smoked.
Effects: Addiction, irritability, anxiety, increased blood pressure, paranoia, depression, aggression, dizziness, sleeplessness, loss of appetite.

Drug Name: **Methamphetamines**
Drug Type: Stimulant
Facts: Some users avoid sleep 3 to 15 days.
How Consumed: Orally, injected, smelled, or smoked.
Effects: Addiction, irritability, aggression, stroke, paranoia, psychosis, heart and blood vessel toxicity, hallucinations, arrhythmia.

Drug Name: **Ecstasy**
Drug Type: Stimulants
Facts: Ecstasy is popular at all-night underground parties and is the most common designer drug.
How Consumed: Orally
Effects: Psychiatric disturbances, including panic, anxiety, depression, and paranoia, blurred vision, sweating, increased heart rate, hallucinations, and sleep problems.

Drug Name: **Ritalin**
Drug Type: Stimulant
How Consumed: Tablet is crushed, and the powder is snorted or injected.
Effects: Loss of appetite, fevers, convulsions, and severe headaches. Increased risk of exposure to HIV, hepatitis, and other infections. Paranoia, hallucinations, excessive repetition of movements and meaningless tasks, muscle twitching.

Drug Name: **Cocaine**
Drug Type: Stimulant
Facts: A powerfully addictive drug. Heavy use may produce paranoia, hallucinations, aggression, insomnia, and depression.
How Consumed: Smelled, dissolved in water and injected.
Effects: Addiction, pupil dilation, high blood pressure and heart rate, increased heart attack, insomnia, anxiety, restlessness, irritability, increased body temperature, death from overdose.

Drug Name: **Heroin**
Drug Type: Opiates
Facts: Heroin users quickly develop a tolerance to the drug and need more and more of it. They want to get the same effects.
How Consumed: Injected generally from arm.
Effects: Addiction, slurred speech, constricted pupils, impaired night vision, nodding off, respiratory depression or failure, and skin infections. Increased risk of exposure to HIV, hepatitis, and other infectious diseases if injected.

Drug Name: **PCP**
Drug Type: Hallucinogens
Facts: Marijuana joints can be inserted into PCP.
How Consumed: Snorted, smoked, orally, or injected.
Effects: Hallucinations, fear, panic, aggressive behavior, impaired motor coordination, inability to feel physical pain. Increased risk of exposure to HIV, hepatitis, and other infectious diseases if injected.

Drug Name: **LSD (Lysergic Acid Diethyl amide)**
Drug Type: Hallucinogen
Facts: LSD is the most common hallucinogen. LSD tabs are often decorated with colorful designs or cartoon characters.
How Consumed: Tabs taken orally or gelatin/liquid put in eyes.
Effects: Elevated body temperature and blood pressure, suppressed appetite, sleeplessness, tremors, chronic recurring hallucinations.

Drug Name: **Mushrooms**
Drug Type: Hallucinogens
Facts: Purchasable via internet.
How Consumed: Eaten or brewed and drunk in tea.
Effects: Increased blood pressure, sweating, nausea, hallucinations.

Drug Name: **Marijuana**
Facts: Can be smoked using homemade pipes.
How Consumed: Smoked or eaten.
Effects: Bloodshot eyes, paranoia, reduced comprehension, reduced ability to perform tasks requiring concentration and coordination. Impairments in learning, perception, and judgment; difficulty speaking.

Drug Name: **Tobacco**
How Consumed: Cigarettes, cigars, pipes, smokeless tobacco, water pipe (*nargile*).
Effects: Addiction, heart disease. Cancer of the lung, larynx, esophagus, bladder, pancreas, kidney, and mouth.

Drug Name: **Alcohol**
Drug Type: Depressant
How Consumed: Orally
Effects: Addiction (alcoholism), dizziness, vomiting, hangovers, slurred speech, impaired motor skills, violent behavior, fetal alcohol syndrome, respiratory depression and death (high doses).

Drug Name: **Inhalants**
Other Names: Solvents, thinners, glues, laughing gas, sprays, cleaning fluids, chemical liquids.
Facts: Many products used at homes can be sniffed. All inhalants are toxic.
How Consumed: Sniffed in a nylon bag or inhaling vapors.
Effects: Addiction, headache, vomiting, impaired motor skills, violent behavior, cancer, bloodshot eyes; liver, lung, and kidney damage.

Thinner is used as a solvent to dissolve dyestuff. *Bally* (a glue brand) and thinner 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 AMATEM stuff. In AMATEM, UMATEM and ÇEMATEM clinics there are many inpatient teenagers who are addicted to those inhalants.

### 1.2. To Stay Adolescents Drug Free

How can a person at high risk be prevented from being addicted? The answers vary depending on each student’s attributes, social and economical status, and family type. What the school management and the family can do to help the adolescents to stay drug free is listed below as some precautions (US Government, 2010), (Köknel, 2010).

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 a curiosity and passion 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" (Gürol, 2007). Also, there are some other tips below which can be useful for the family and educators (US Government, 2010b).

- Leading them to social events such as sport activities and playing instruments
- Reading anti-drug books and articles together and having a discussions
- Talking to them about the dangers of substances
- Listening to what they are saying, looking at them when listening
- Finding out how their day was, what happened in school or with their friends
- Going to their events, i.e., sports games, plays, school shows
- Knowing who their friends are and knowing where they are
- Setting clear expectations for their behavior
- Being consistent in your training and discipline towards them
- Creating ways to have meaningful participation in their lives.

### 1.3. The Purpose of the Thesis

After being a user, a person struggles with the addiction problem and his/her life gets worse. Scientific studies indicate that it becomes very difficult for a person to get rid of this habit after being addicted. Because of that, preventing teenagers from being addicted becomes an important issue to be discussed. This study does not focus on addicted people. This study focuses on people at risk whose ages rage from 10 to 20.

The purpose of the thesis is to detect the risk ratio for a student. If the risk ratio is higher than normal it is strongly recommended families and educators that some precautions should be taken for this student to stay him/her drug free.

Hence, we can call this project "*Urgent Warning System*" for those teenagers who are highly at risk. With the help of this project it will be possible to take some precautions for the teenagers to stay them drug free.

### 1.4. Steps of the Work in this Thesis

There is a list below which has been done step by step while studying on this thesis. All the steps are aimed to figure out 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 B.Çekmece Elementary School
    c. From the ÇEMATEM (Çocuk ve Ergen Alkol ve Madde Bağımlıları Tedavi Merkezi)
2. Converting all the nominal data from the questionnaires to numeric values
    a. To the file "`DB.txt`"
    b. To the array "`DB[][]`"
3. Building the model

   a. Arranging the predefined training set
   b. Preparing the model and applying it to the algorithms. The algorithms are:
      i. kNN (k- nearest neighbors) Classifiers
      ii. Naïve Bayes Classifiers
      iii. Decision Tree Algorithms
      iv. One-Attribute-Rule and PART Algorithms

4. Implementing the algorithms in the C programming language
5. Discussing the results and making comparisons
6. Reporting
7. Informing the educators and families about the students who are highly at risk and taking precautions.

To illustrate these steps there is a flow char below. This flow chart shows the system how it works. As it is seen, after applying the questionnaire to a student, some extra information such as GPA, smoking habit, alcoholism, and misbehavior is taken from school. This data is collected and applied to the classifier algorithms so as to find out the risk ratio about the student.

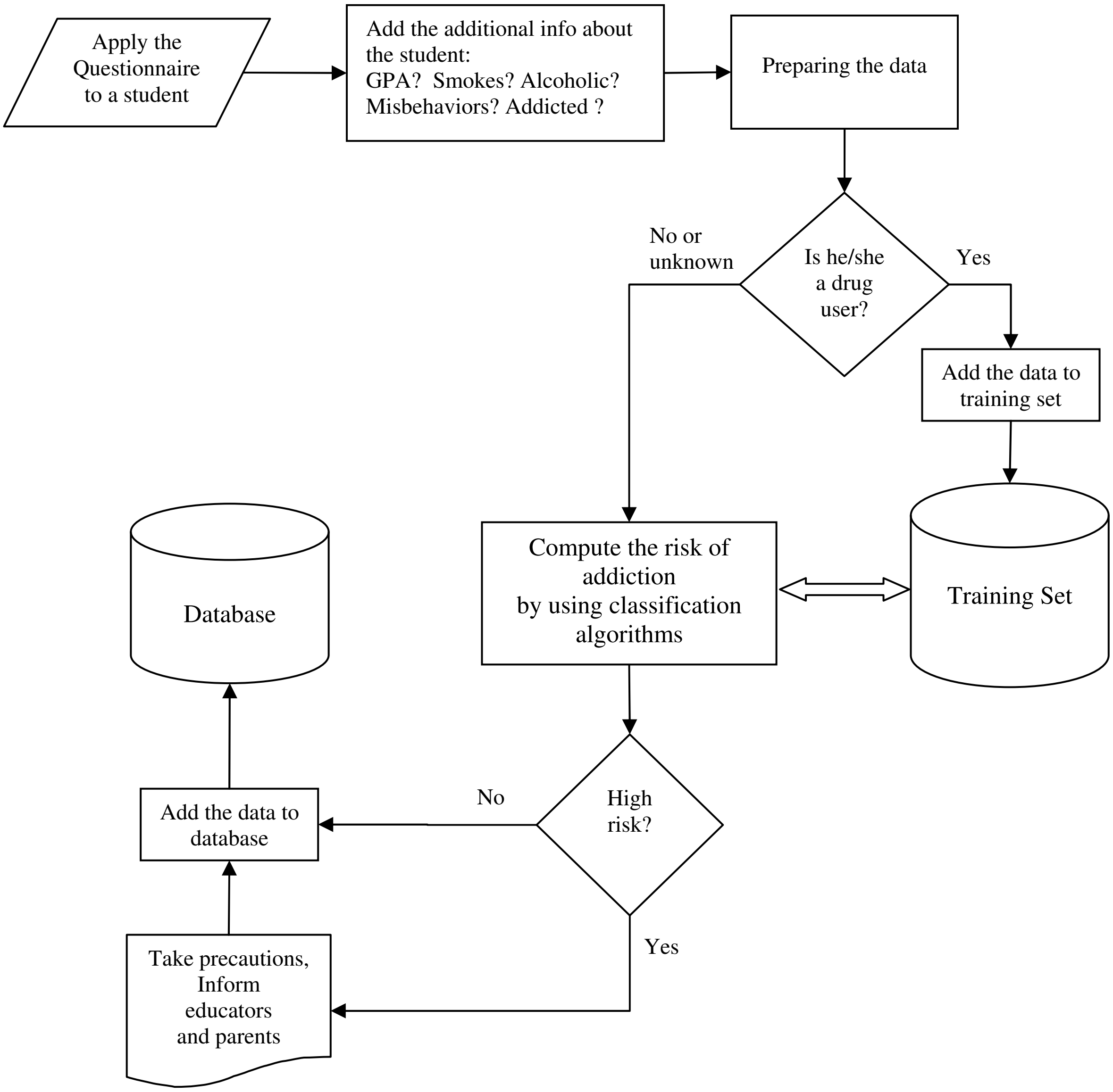

Apply the Questionnaire to a student
Add the additional info about the student: GPA? Smokes? Alcoholic? Misbehaviors? Addicted ?
Preparing the data
Is he/she a drug user?
No or unknown
Yes
Add the data to training set
Training Set
Compute the risk of addiction by using classification algorithms
High risk?
No
Yes
Add the data to database
Database
Take precautions, Inform educators and parents

**Figure 1.1** How the system in this thesis works?

# CHAPTER 2

# PREPARING THE QUESTIONNAIRE

In order to find out the percentage of risk value for each student, the reasons and factors leading to addiction should be identified. A questionnaire which aims to find out the reasons and factors of addiction ought to be utilized so as to gather relevant data to conduct this study.

## 2.1. The Reasons of Addiction

There are many harmful effects of drug abuse including changes in the user's brain, body, and spirit. This thesis is not dealing with the results, but on dealing with the reasons of being dependent. There are some reasons below leading adolescents to addiction:

- Family related reasons: dissatisfaction with family relationships, antisocial family members, stress in the family, poverty or welfare in the family, illiterate parents, divorced parents, loss of one or two parents, lack of people who could be a positive role model for the adolescent (Ögel, 2004),
- Physical/sexual abuse or violence,
- Some genetic factors, birth related problems, physical or psychological problems (Ögel, 2001; cited in Kircan, 2006),
- Social variables can be peer pressure, cultural effects, and acceptance of substance use in the society, low socioeconomic status, and unemployment. Peer relationships

influence adolescent substance use like establishing closer companionship with substance user friends, and need for approval from friends (Kimmel & Weiner, 1995; Windle & Windle, 2003),

- In general, adolescents first experience alcohol or cigarette before heroin or hard substances (Bukstein, 1995).

Possible risk factors for the adolescence are all stated above. They can be predictors of substance use.

### 2.2. Questions in the Questionnaire

There are 20 questions in the questionnaire which is given to the students at schools and the inpatients in ÇEMATEM to fill out. The answers in the questionnaire are regarded as the attributes of the records in the database. The questionnaire has been prepared on the basis of the reasons mentioned above. The questions have been derived from these scientific articles and related books (Tuncer, 2007), (Can, 2007), (Zor, 2005), (Gülkan, 1994), (Erdem et al, 2006), (Aydin, 2006), (Seyman, 2000), (Gök, 2010).

Each of the questions tries to determine an effect leading the person to be a drug abuser. Some of them have higher effects; some of them lower effects on addiction. For example, smoking cigarette, being a member of problematic family, being a friend of drug users have higher effects on the individual to use substances.

The first twenty questions are asked directly to the students. The answers of the last five questions are taken from the educators in their school. It is not preferred to ask these last five questions to the students directly because they avoid answering these private and individual questions. They do not want to reveal their bad habits. That is why the answers of the last five questions are taken from the educators in the schools. Twenty five questions in the questionnaire are listed below:

1. Yaşınız: ............

2. Cinsiyetiniz? (Tosun, 2005)

   a. Erkek b. Kız

3. Anne baba sağ mı?

   a. Evet, ikisi de sağ
   b. Sadece annem sağ
   c. Sadece babam sağ
   d. İkisi de sağ değil

4. Kiminle kalıyorsunuz? (Saatçioğlu,2002)

   a. Sadece annemle (annem babam ayrı)
   b. Sadece babamla (annem babam ayrı)
   c. Annem ve babamla
   d. Başka bir akrabamda
   e. Yetimhanede
   f. Yurtta

5. Ailenizin aylık ortalama geliri?

   a. 1000 TL'den az
   b. 1000–3000 TL
   c. 3000–6000 TL
   d. 6000 TL'den fazla

6. Akrabalarınızla (halanızla, amcanızla, teyzenizle, dayınızla, büyük annenizle, büyük babanızla, kuzenlerinizle) ne sıklıkta görüşürsünüz?

   a. Haftada bir b. Ayda bir c. Yılda bir d.
   Hiç

7. Enstrüman çalar mısınız?

   a. Evet b. Hayır

8. Düzenli olarak spor yapar mısınız?

   a. Evet b. Hayır

9. Ne sıklıkla kitap okursunuz?

   a. Haftada bir
   b. Yılda bir
   c. Ayda bir
   d. Okumam

10. Hangi tür müzik seversiniz?

   a. Rock – Heavy Metal

   b. RAP
   c. Türkçe veya yabancı Pop
   d. Türkü
   e. Müzik sevmem

11. Sinemaya ne sıklıkla gidersiniz?

    a. Haftada bir
    b. Ayda bir
    c. Yılda bir
    d. Neredeyse hiç

12. Günde kaç saatinizi Internet başında geçiriyorsunuz?

    a. Bazen hiç, bazen bir saatten az
    b. 1–2 saat
    c. En az 3 saat

13. Kendiniz için zararlı olduğunu bildiğiniz bir şeyin çok sevdiğiniz bir arkadaşınız tarafından size teklif edilmesi durumunda...

    a. Bir defaya mahsus kabul ederim
    b. Kesinlikle reddederim
    c. Bilemiyorum

14. Anne veya babanız bir probleminiz olduğunda sizinle ilgilenir mi?

    a. Sadece annem ilgilenir
    b. Sadece babam ilgilenir
    c. Annem ve babam ilgilenir
    d. Bana pek karışmazlar

15. Gelecekte hayatınızın bugüne göre nasıl olacağını tahmin ediyorsunuz?

    a. Daha kötü
    b. Aynı
    c. Daha iyi
    d. Fikrim yok

16. Bir sorununuz olduğunda paylaşacağınız ve kendinize en yakın hissettiğiniz kişi kimdir?

    a. Annem veya babam
    b. Kardeşim
    c. Akrabam
    d. Öğretmenim
    e. Arkadaşım
    f. Kendimi hiç kimseye yakın hissetmiyorum

17. Boş zamanlarınızı en çok kiminle paylaşıyorsunuz?

    a. Ailemden biriyle

b. Akrabalarımdan biriyle
c. Arkadaşlarımla
d. Bilgisayarımla
e. Hiç kimseyle

18. Aşmak zorunda olduğunuz fakat aşılması oldukça güç bir durumla karşılaştığınızda...

a. Zor olduğunu biliyorsam hiç uğraşma gereği duymam
b. Biraz uğraşırım olmazsa olayları kendi akışına bırakırım
c. Ne kadar zor olursa olsun aşmam gerekiyorsa elimden geleni yaparım

19. Kendime güvenim yok diye düşünüyorum.

a. Evet yok
b. Hayır var
c. Bilemiyorum

20. Okulunuzda en çok sevdiğiniz arkadaşlarınız kimlerdir? Lütfen soy isimleriyle birlikte yazın

a. . . . . . . . . . . . . . . . . . . . . . . . . .
b. . . . . . . . . . . . . . . . . . . . . . . . . .
c. . . . . . . . . . . . . . . . . . . . . . . . . .
d. . . . . . . . . . . . . . . . . . . . . . . . . .
e. . . . . . . . . . . . . . . . . . . . . . . . . .

21. Beşlik sisteme göre bir önceki dönem not ortalamanız nadir. (Gürol, 2007):

22. Öğrencinin o ana kadar disiplin suçu işleyip işlemediği

23. Alkol bağımlılığının olup olmadığı

24. Sigara içip içmediği

25. Madde bağımlısı olup olmadığı

There are some possible questions that could have place in the questionnaire. These are:

- How many hours do you sleep?
- How much do you eat?
- Are you aggressive?
- Is there any tattoo on your skin?
- Is there any cut or scratch on your arm?
- Are you eyes red?
- How is your wearing style?

These questions have not been asked because, these are abuse symptoms. The aim of the questions is not to figure out whether the person at that time is a drug abuser or not. These answers do not show the leading reasons, but the results of addiction. Our project deals with the reasons.

For the purpose of piloting, the questionnaire including 25 questions has been checked and examined by the PDR teachers before applying. Necessary changes in it have been done.

### 2.3. Data Collection

With the official permission taken from B.ÇEKMECE İLÇE MİLLİ EĞİTİM MÜDÜRLÜĞÜ, the questionnaires with the 20 questions have been distributed to the students to fill out in Tepecik M.Hüsnü Özyeğin High School and B.Çekmece Primary School. The answers of the last five questions have been given by the PDR teacher, assistant principals and the teachers in these schools. 671 questionnaires in total, have been completed. In these schools, only two drug users have been identified.

In addition, with the official permission taken from İSTANBUL İL SAĞLIK MÜDÜRLÜĞÜ, the same questionnaire has been applied to the 35 ÇEMATEM inpatients. The ÇEMATEM clinic is the health center for the drug abused youth in the hospital, “Prof.Dr. Mazhar Osman Bakırköy Ruh ve Sinir Hastalıkları Hastanesi”. In ÇEMATEM, some of the drug abusers have filled out the questionnaires by themselves and some of them have only replied orally. Few of them have avoided giving their names and surnames.

In the database there are more than 700 records so at to use in the implementations of the algorithms.

# CHAPTER 3

# DATA PREPARATION PROCEDURE

## 3.1. Recording Questionnaires to the Database

After collecting more than 700 questionnaires from the schools and the hospital, they have been recorded to the database one by one. All the answers in the questionnaire have a nominal value shown in the figure 3.1. 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, 1 means male; 2 means female. For the forth one, there are four choices. 1, 2, 3, or 4 are equal to “a”, b”, “c” or “d” respectively. All the answers to the questions have been converted to integer values by this method. But in the 7th and 8th questions, 0 means “No”, 1 means “Yes”.

To illustrate, there are some randomly selected questions from the questionnaire below. There are some integer values in the parenthesis at the end of each line. They represent the numeric values of the nominal answers. All the integer values for a person are then recorded to a line in the database.

**Table 3.1** Scoring the questionnaire with numeric values according to the answers

**1.** Yaşınız: ............(An integer value)

**2.** Cinsiyetiniz?

a. Erkek (1)    b. Kız (2)

**4.** Kiminle yaşıyorsunuz?

a. Sadece annemle (annem babam ayrı) (1)
b. Sadece babamla (annem babam ayrı) (2)
c. Annem ve babamla (3)
d. Başka bir akrabamda (4)
e. Yetimhanede (5)
f. Yurtta (6)

**7.** Enstrüman çalar mısınız?

a. Evet(1)    b. Hayır (0)

**22.** Öğrencinin o ana kadar disiplin suçu işleyip işlemediği (0: işlemedi, 1: işledi)

**23.** Alkol bağımlılığının olup olmadığı (1: var, 0:yok)

**24.** Sigara içip içmediği (1:tiryaki, 0:değil)

**25.** Madde bağımlısı olup olmadığı (100:Bağımlı, 0:kesinlikle değil)

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. He has written only four best friends. The answer to the 20th question is very important in this thesis, because peer pressure plays a major role to lead a person to drug addiction (Windle & Windle, 2003).

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 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 3.2** 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
108 13 2 1 3 2 1 1 0 2 2 1 3 1 3 2 5 4 2 1 105 101 114 116 0 4 0 0 0 0
109 13 2 1 3 1 2 1 1 1 2 1 2 2 3 3 1 3 3 2 112 114 107 108 110 -1 0 0 0 0
110 13 2 1 3 3 1 1 0 1 5 1 1 1 3 2 4 3 2 2 104 114 115 100 0 -1 0 0 0 0
111 14 1 1 3 2 1 0 1 2 2 2 1 2 3 4 2 1 2 3 0 0 0 0 0 3 0 0 0 0
112 15 2 1 3 1 1 1 1 1 2 1 3 2 3 3 1 3 3 2 100 114 107 108 0 -1 0 0 0 0
113 13 2 1 3 3 1 0 1 1 2 2 2 2 3 2 5 4 3 2 103 0 0 0 0 -1 0 0 0 0
114 13 2 1 3 1 1 1 1 2 3 4 2 2 3 3 5 1 3 2 102 108 100 110 115 4 0 0 0 0
115 13 2 1 3 2 1 0 1 1 3 2 1 2 3 3 5 1 3 2 103 117 108 116 107 4 0 0 0 0
116 13 2 1 3 2 1 0 1 2 3 1 1 2 3 3 1 3 3 2 101 108 115 105 115 4 0 0 0 0
117 13 2 1 3 2 3 1 0 1 3 2 1 2 3 3 1 3 3 2 102 116 110 0 0 -1 0 0 0 0
118 13 1 -1 3 1 1 0 1 2 2 2 1 3 3 4 1 1 3 2 120 131 0 0 0 -1 0 0 0 0
119 17 1 1 3 2 1 0 1 2 2 2 2 2 4 2 6 1 1 1 120 0 0 0 0 -1 0 0 0 0
120 13 1 1 3 2 3 0 1 2 2 2 1 2 3 4 1 1 2 1 118 0 0 0 0 5 0 1 1 100
```

As it is seen above, there are 25 attributes of each tuple for those classification algorithms. 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 five values are the IDs of best friends. The last 5 values are the private and individual information taken from the school.

**Table 3.3** The structure of the database

| | Answers in the questionnaire | | | | | | | | | | | Additional data taken from school | | | | | Output |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | | | 18 | 19 | 20 | | | | | | |
| Student ID | Age | Gender | Are your mother and father alive? | Who do you live with? | Your family's income? | | | | | | | GPA | Smoking habit? | Alcoholic? | Misbehaviors? | Drug addiction? | The risk ratio |
| 1 | 14 | 1 | 1 | 1 | 1 | . | . | . | . | . | . | 3 | 0 | 0 | 0 | 0 | % 10 |
| 2 | 14 | 2 | 1 | 2 | 4 | | | | | | | 4 | 0 | 0 | 1 | 0 | % 5 |
| 3 | 13 | 2 | 1 | 2 | 3 | | | | | | | 5 | 1 | 0 | 1 | 0 | % 65 |
| 4 | 15 | 1 | 2 | 1 | 2 | | | | | | | 5 | 0 | 0 | 0 | 0 | % 20 |
| 5 | 16 | 1 | 3 | 1 | 2 | | | | | | | 3 | 0 | 0 | 0 | 0 | % 10 |
| 6 | 17 | 2 | 4 | 1 | 2 | | | | | | | 2 | 0 | 1 | 0 | 0 | % 0 |
| 7 | 14 | 1 | 2 | 4 | 1 | | | | | | | 4 | 0 | 0 | 0 | 0 | % 0 |
| 8 | 12 | 2 | 1 | 1 | 1 | | | | | | | 4 | 0 | 0 | 0 | 0 | % 0 |
| 9 | 14 | 1 | 1 | 6 | 1 | | | | | | | 2 | 1 | 1 | 1 | 1 | % 100 |
| 10 | 15 | 1 | 1 | 3 | 2 | | | | | | | 2 | 1 | 1 | 1 | 1 | % 100 |
| 11 | 17 | 1 | 1 | 1 | 4 | | | | | | | 4 | 1 | 1 | 1 | 0 | % 40 |
| | . | . | . | . | . | | | | | | | . | . | . | . | . | . |
| | . | . | . | . | . | | | | | | | . | . | . | . | . | . |
| | . | . | . | . | . | | | | | | | . | . | . | . | . | . |
| 501 | 17 | 1 | 2 | 1 | 1 | | | | | | | 3 | 0 | 0 | 0 | 0 | %5 |
| 502 | 14 | 1 | 2 | 1 | 1 | | | | | | | 2 | 1 | 0 | 1 | 0 | %60 |
| 503 | 15 | 1 | 1 | 2 | 4 | | | | | | | 4 | 0 | 0 | 1 | 0 | %15 |
| | . | . | . | . | . | | | | | | | . | . | . | . | . | . |

### 3.2. Preparing Data

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*.

There is a figure below representing the steps of data preparation for further calculations in this thesis. DB0.txt is the database as a text file; DB0[][] is the database as a two dimensional array used in the implementations.

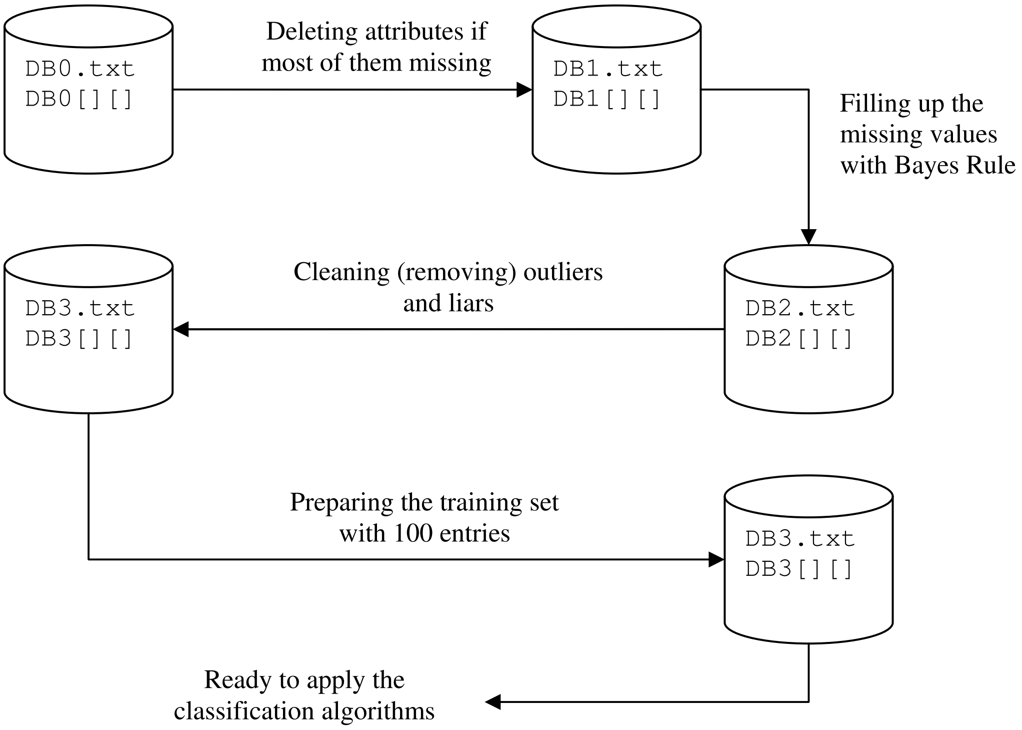

DB0.txt
DB0[][]
Deleting attributes if most of them missing
DB1.txt
DB1[][]
Filling up the missing values with Bayes Rule
DB2.txt
DB2[][]
Cleaning (removing) outliers and liars
DB3.txt
DB3[][]
Preparing the training set with 100 entries
DB3.txt
DB3[][]
Ready to apply the classification algorithms

**Figure 3.1** Steps of data preparation from questionnaires

#### 3.2.1. Removing Outliers

Erroneous and noisy data has been corrected and some of them have been 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.

### 3.2.2. Removing Liars

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

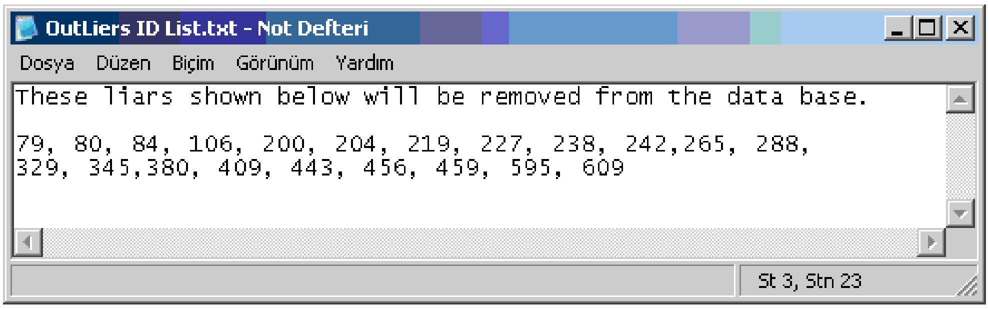
- A primary school student has written his age as 22.
- Someone has answered the 7th question "*Do you play any music instrument?*" as "*yes*". But in the 10th question "*Which kind of music do you like?*" he has given the answer "*I don't like music*". This is obviously a conflict. This record has been 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 been removed from the database.

**Table 3.4** 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 below, “OutLiers ID List.txt” to show the liars.

**Figure 3.2** The “OutLiers ID List.txt” file showing the liars



```
These liars shown below will be removed from the data base.

79, 80, 84, 106, 200, 204, 219, 227, 238, 242,265, 288,
329, 345,380, 409, 443, 456, 459, 595, 609
```

### 3.2.3. Missing Data

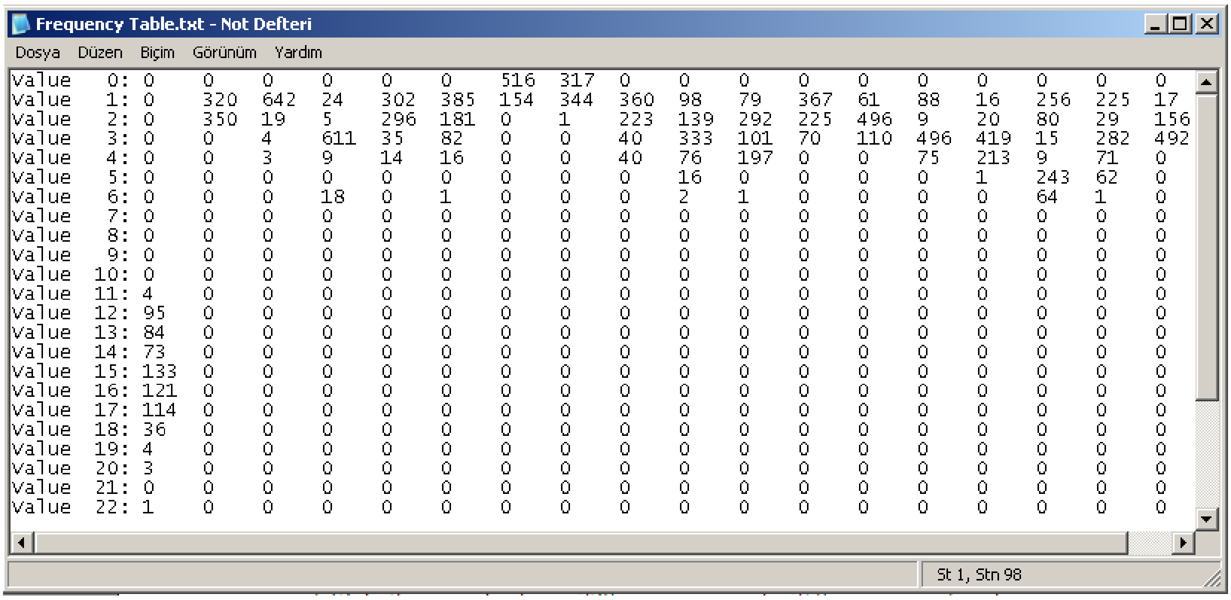
In the questionnaire, there are few questions that have not been answered by the students. While converting the questionnaires to numeric values, it is put as -1 to demonstrate the missing data in the database. The missing data is replaced with Bayesian Probability Technique although there are some other methods (Dunham, 2005a).

If there is a missing value of an attribute of a tuple, the whole probabilities of the attribute in the all 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:

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

The explanation of the formula is in the Chapter 4, under the title Naïve Bayes Classifier.

**Table 3.5** The "Frequency Table.txt" showing the frequencies of each attribute


Frequency Table.txt - Not Defteri
Dosya Düzen Biçim Görünüm Yardım
St 1, Stn 98

```
Value   0: 0    0    0    0    0    0    516  317  0    0    0    0    0    0    0    0    0    0
Value   1: 0    320  642  24   302  385  154  344  360  98   79   367  61   88   16   256  225  17
Value   2: 0    350  19   5    296  181  0    1    223  139  292  225  496  9    20   80   29   156
Value   3: 0    0    4    611  35   82   0    0    40   333  101  70   110  496  419  15   282  492
Value   4: 0    0    3    9    14   16   0    0    40   76   197  0    0    75   213  9    71   0
Value   5: 0    0    0    0    0    0    0    0    0    16   0    0    0    0    1    243  62   0
Value   6: 0    0    0    18   0    1    0    0    0    2    1    0    0    0    0    64   1    0
Value   7: 0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value   8: 0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value   9: 0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  10: 0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  11: 4    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  12: 95   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  13: 84   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  14: 73   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  15: 133  0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  16: 121  0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  17: 114  0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  18: 36   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  19: 4    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  20: 3    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  21: 0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
Value  22: 1    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

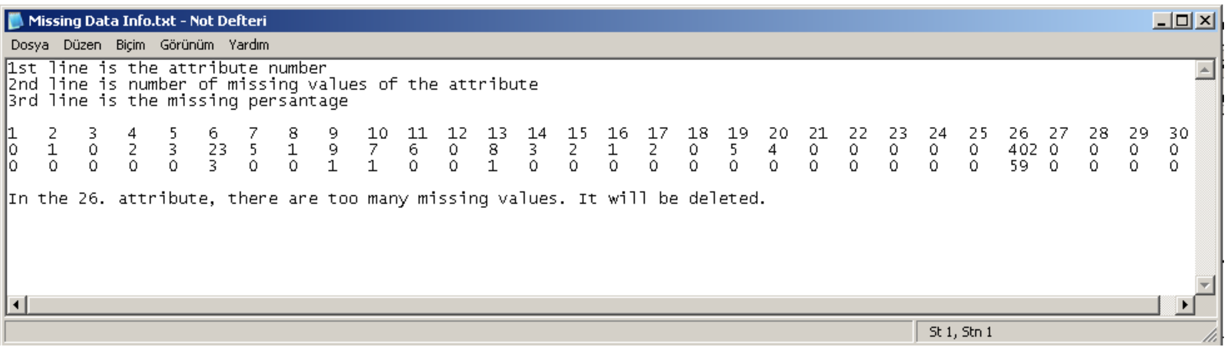
Some outputs that show the frequencies of each attribute are given In the figure above. For example, in the first column, there are 4 students at the age of 11. There are 95 students at the age of 12. In the second column, it is presented that 320 students are male and 350 students are female.

### 3.2.4. 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 asks 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 3.3** The "Missing Data Info.txt" file showing the missing values



```
1st line is the attribute number
2nd line is number of missing values of the attribute
3rd line is the missing persantage

1    2    3    4    5    6    7    8    9    10   11   12   13   14   15   16   17   18   19   20   21   22   23   24   25   26   27   28   29   30
0    1    0    2    3    23   5    1    9    7    6    0    8    3    2    1    2    0    5    4    0    0    0    0    0    402  0    0    0    0
0    0    0    0    0    3    0    0    1    1    0    0    1    0    0    0    0    0    0    0    0    0    0    0    0    59   0    0    0    0

In the 26. attribute, there are too many missing values. It will be deleted.
```

The implemented codes which removes outliers, extracts liars, tolerates missing data, deletes attributes are presented in a file at the attachment.

## 3.3. Predefined Training Dataset

For classification problems, training dataset is an 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 (Dunham, 2005b). The training set of data should be defined well before applying the algorithms.

In the thesis, 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 PDR department. 43 of them are at risk ratios ranging from 10% to 80%. 20 students are at the risk ratio %0. The former group and the latter group of students are put into these risk intervals on the basis of the reasons of addiction discussed in chapter 2.

# CHAPTER 4

# APPLYING CLASSIFICATION ALGORITHMS

## 4.1. What is classification?

Classification is the most familiar and the most popular data mining technique. Classification is also an important subject in Machine Learning, in Pattern Recognition and in Artificial Intelligence. It is used in a variety of areas.

Typically, given a set of training instances with corresponding class labels, a classifier is learned form these training instances and used to predict the class of a new instance (Jiang et al, 2006).

## 4.2. Applied Classifiers in This Thesis

In this thesis, four types of classification method are applied to categorize the new coming records. These algorithms are kNN as a distance based algorithm, Naïve Bayes Classifier as a statical based algorithm, ID3 and C4.5 as decision tree based algorithms, One-Attribute-Rule (OneR) and PART as rule based algorithm. All four types of classification methods are applied, discussed and criticized in this thesis. The definitions of some notations used in these methods are as follows:

$D$ is the database. $t_i$ is the $i^{th}$ tuple in the database.

$T = \{ t_1, t_2, t_3, \ldots t_n \}$

$D = \{ t_1, t_2, t_3, \ldots t_n \}$

$C = \{ Class_1, Class_2, Class_3, Class_4, Class_5 \}$

*C* is the set of classes including 5 types. It is defined in the Title 4.3. Risk Group Classes.

$f : D \rightarrow C$

*f* is the mapping function, in other words *f* is the classifier.

$C_j = \{ t_i \mid f(t_i) = C_j, 1 \leq i \leq N, \text{and } t_i \in D \}$

Each tuple in the database is assigned to exactly one class by the classifier algorithms.

### 4.3. Risk Group Classes (Output Classes)

All the classifiers applied in this thesis give outputs for each record in the database. The output indicates the risk ratio for this person. There are five risk classes as outputs including $Class_1$, $Class_2$, $Class_3$, $Class_4$, and $Class_5$.

The risk ratio in $Class_1$ is between 0% and 20%. Since it is the lowest risk section, there is no need to take any precautions. Students in this section are not at risk at all.

The risk ratio in $Class_2$ is between 20% and 40%. Students in this class are not potential drug users. Bu that does not mean there is no risk at all.

The risk ratio in $Class_3$ is between 40% and 60%. Students in this section tent to be potential drug users in the future. it will be better to inform the students about the harmful effects of drugs and to take some precautions.

The risk ratio in $Class_4$ is between 60% and 80%. There is a very big risk for the students in this section. It is compulsory to take urgent precautions.

The risk ratio in *Class*$_5$ is between 80% and 100%. Since the risk rate is the highest in this section, some precautions must be taken immediately. Educators and the family have to be very careful about the student. If he/she is a drug user, it is compulsory to consult to an AMATEM clinic immediately.

These risk groups can be seen in the figure below:

**Table 4.1** Risk group classes

| | | %10 | %20 | %30 | %40 | %50 | %60 | %70 | %80 | %90 | %100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Class 1 | Very low risk | ▨ | ▨ | | | | | | | | |
| Class 2 | Low risk | | | ▨ | ▨ | | | | | | |
| Class 3 | Normal risk | | | | | ▨ | ▨ | | | | |
| Class 4 | High risk | | | | | | | 🟥 | 🟥 | | |
| Class 5 | Very high risk | | | | | | | | | 🟥 | 🟥 |

## 4.4. k-NEAREST NEIGHBORS ALGORITHMS (kNN)

As a type of an instance-based learning and a lazy learning algorithm, kNN is one of the best and the most usable classification algorithms which is used largely in different applications. The *k*-Nearest Neighbors algorithm is a common method for classifying new tuples (entries, records, or instances) based on closest training examples in the feature space (Jiang et al, 2006).

A kNN classifier determines the class label of a record by looking at the labels of its *k* nearest neighbors in the training dataset and puts the record into the class that most of its neighbors belong to (Aggarwal et al, 2008a).

The neighbors are taken from a set of objects for which the correct classification is known. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

### 4.4.1. *k* value

*k* is a positive integer, typically small. In the classification phase, *k* is a user-defined constant and an unlabelled vector (a query or test point). It is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point. Usually Euclidean distance is used as a distance metric. In this algorithm, the value of *k* yields some problems (Wikipedia, 2010a):

- If the *k* value is high, the prediction of the output class for the new coming tuple gets difficult.
- Larger values of *k* reduce the effect of noise on the classification.
- If the *k* value is low, the number of output classes for the new tuple increases.
- To determine the *k* value depends on the users. Users sometimes can not decide the appropriate value of *k*.
- If *k* is equal to 1, then the object is simply assigned to the class of its nearest neighbor. If there is more than one nearest neighbor classes, the new object is assigned to a class according to the average of these nearest classes.

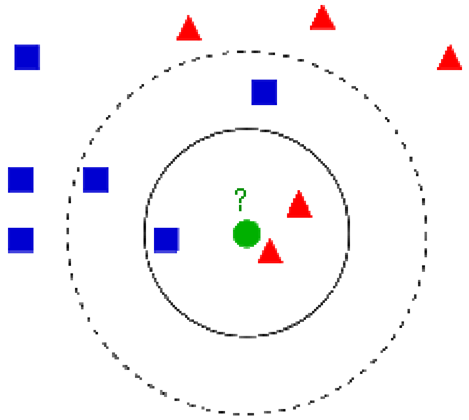


**Figure 4.1** Sample kNN classification (Wikipedia, 2010a)

The test sample (the green circle in the middle) should be classified either to the first class of the squares or to the second class of the triangles. If *k* is equal to 3 it is classified to the second class because there are 2 triangles and only 1 square inside the inner circle. If *k* is equal to 5 it is classified to the first class (3 squares vs. 2 triangles inside the outer circle). There is a special case here where the class is predicted to be the class of the closest training sample, in other words when *k* is equal to 1, is called the nearest neighbor algorithm (Wikipedia, 2010a).

Determining the *k* value in this method usually seems to be problematic for users. To avoid the *k* value problem and to have better results, the *distance weighted kNN* algorithm has been applied in this study.

The accuracy of the kNN method can be degraded by the presence of irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve classification. A particularly popular approach is the use of evolutionary algorithms to optimize feature scaling (Nigsch et al, 2006).

In this thesis, five kinds of kNN algorithms have been applied: *Simple kNN, Absolute Distance kNN*, *Euclidian Distance kNN*, *Distance Weighted kNN* and finally *Distance and Attribute Weighted kNN*. These algorithms are listed below from the least to the most efficient.

### 4.4.2. Simple kNN

This type of method is the simplest and easiest one of kNN techniques. Only the average of the closest tuples in the training set to the new record is assumed enough.

For instance, *x* and *y* are considered as two tuples. *y* is in the training set and *x* is the new entry waiting to be placed into a class. *x* and *y* are composed of *N* features (attributes), such that

$x = \{x_1, x_2, x_3, \dots x_N\}$,
$y = \{y_1, y_2, y_3, \dots y_N\}$.

In order to implement the simple kNN, it will be enough to check whether there is a difference between the values of each corresponding attributes of these two tuples. If there is a difference between these two values of the corresponding attributes, it is regarded as 1. If the values of the attributes are the same, it is regarded as 0.

There is an example below in the figure 4.2 illustrating this Simple kNN method.

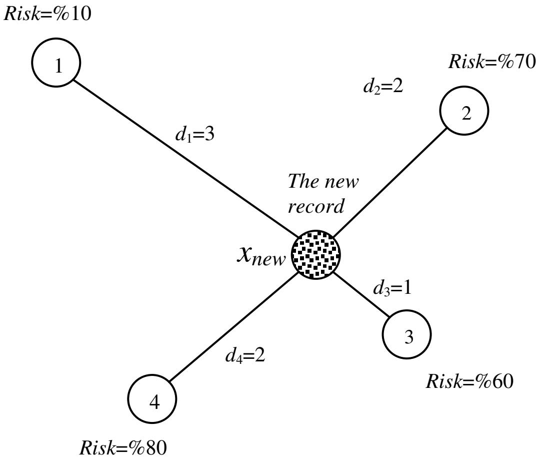


**Figure 4.2** An example to Simple kNN

*k* value is taken as 3. There are four close tuples, whose distances less than or equal to the *k* value(*k*=3), to the new record $x_{new.}$. *d* is the distance between the tuples $x_i$ and $x_{new}$. *N* is the number of tuples. *f* function finds the class of the new entry.

$$f(x_{new}) = \frac{\sum_{i=1}^{N} f(x_i)}{N} = \frac{10+70+60+80}{4} = 55$$

The average of the risk values is 55. %55 is the risk factor of the new record. The new coming record according to this technique will be put into the *Class*$_3$.

This version of the kNN method has yielded the least efficient results.

### 4.4.3. Absolute Distance kNN

This version is obviously better than the *Simple kNN* method. We can compute the absolute distances between two tuples using the absolute distance function *d*(*x*,*y*). *x* is the new tuple to be put into a class and *y* is one of the tuples in the training dataset. *x* and *y* are composed of *N* features (attributes), such that

$x = \{x_1, x_2, x_3, \ldots x_N\}$,

$y = \{y_1, y_2, y_3, \ldots y_N\}$.

$N$ refers to the number of features (attributes) of a tuple. In this thesis, the $N$ value is equal to 25. The distance function, absolute distance measuring has a formula:

$$d_A(x, y) = \sum_{i=1}^{N} \left| x_i - y_i \right|$$

To illustrate, an example is presented below illustrating this *Absolute Distance kNN* method. The 5th question in the questionnaire:

"What is the income of your family in a month?"

a. Less than 1000TL
b. Between 1000 and 3000TL
c. Between 3000 and 6000TL
d. More than 6000TL

The answers to this question are marked as 1 (for a), 2 (for b), 3 (for c), or 4 (for d) respectively in the database. If the new record $x$ has chosen "a" as an answer to this question, its value is 1. On the other hand, if $y$ record has chosen "d" as an answer, its value is 4. There is a great gap in the financial situation between the records $x$ and $y$. To find the absolute distance between $x$ and $y$, 1 is subtracted from 4. The distance is 3.

$$d_A(x_5, y_5) = \left| x_5 - y_5 \right| = \left| 1 - 4 \right| = 3.00$$

If this example is applied to the same example in the previous technique, $x$ will be the new record; $y$ will be the 4th tuple whose risk value is %80. It is assumed that the distances of 1st, 2nd, and 3rd tuples has not changed; only the 4th tuple's distance has changed. In the previous technique the distance between $x$ and $y$ was equal to 1. But now it is equal to 3. As it is seen in the figure below, the 4th tuple has gone away 2 steps from the new record.

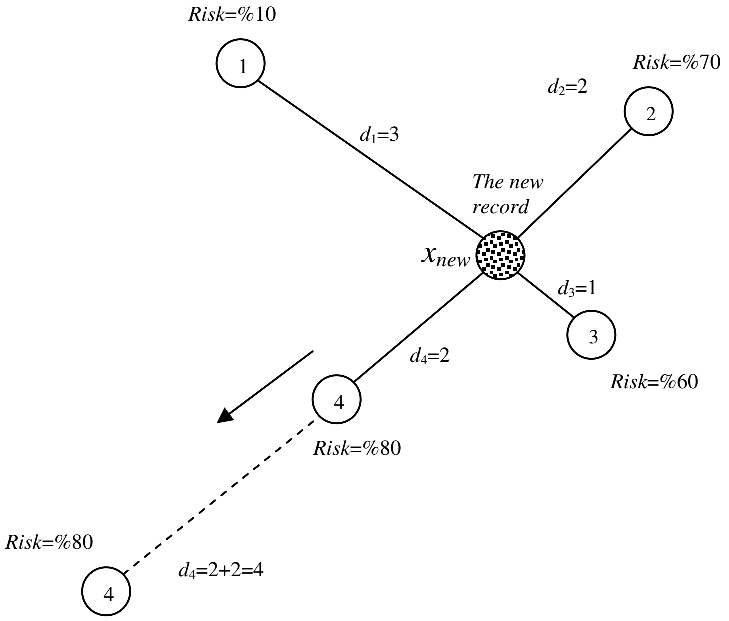


**Figure 4.3** An example to Absolute Distance kNN

*k* value is still taken as 3. There are four close tuples, whose distances less than or equal to the *k* value (*k*=3), to the new record $x_{new.}$. *d* is the distance between the tuples $x_i$ and $x_{new}$. *N* is the number of tuples. *f* function finds the class of the new entry. But the 4th tuple has gone away because of the Absolute Distance metric and not contaminated to the calculation.

$$f(x_{new})=\frac{\sum_{i=1}^{N} f(x_i)}{N}=\frac{10+70+60}{3}=46.67$$

Now the average of the risk values is 46.67. %46.67 is the risk factor of the new record. The new coming record according to this technique will be put into the same class, *Class*3. But in this calculation risk factor decreases to 46.67% from 55%. In the previous technique, the distance between these two tuples is accepted as just 1. But in this technique, the distance is equal to 3. Because of that, this method gives better results than the previous method, Simple kNN.

### 4.4.4. Euclidian Distance kNN

Euclidian distance measuring is used to find out the real distance in N dimentional space between two tuples. Because of that it can be said that it is better than Absolute Distance Measuring. Euclidian distance measuring has a formula:

$$d_E(x, y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$$

*N* is the number of attributes of a tuple. *N* is equal to 25 in our study. It seems to be in a 25 dimensional space. $d_E$ is the Euclidian distance between the tuples *x* and *y*. *N* is the number of features (attributes) of a tuple. To illustrate, the 5th question in the questionnaire used in the previous title can be examined again. The answers of *x* and *y* are 1 and 4 respectively.

$$d_E(x_5, y_5) = \sqrt{4^2 - 1^2} = 3.87$$

By *Euclidian distance measuring,* the distance is calculated as 3.87. In the previous technique, in the Ab*solute Distance Measuring*, it was 3.00. On the other hand, in the *simple kNN* version, the result was only 1.00. This example shows that *Euclidian distance measuring* has given better results when it is compared with other measuring techniques mentioned above.

### 4.4.5. Distance Weighted kNN

One obvious refinement to the k-Nearest Neighbor algorithm is to weight the contribution of each of the *k* neighbors according to their distance to the query point $x_{new}$, giving greater weight to closer neighbors. We might weight the vote of each neighbor according to the inverse square of its distance from *x*. This is done by this formula (Mitchell, 1997a):

$$f(x_{new}) = \underset{v \in V}{\arg\max} \sum_{i=1}^{N} w_i \delta(v, f(x_i)) \text{ where}$$

$$w_i = \frac{1}{d^2} = \frac{1}{d(x_{new}, x_i)^2}$$

*d* is the distance between the tuples $x_i$ and $x_{new}$. *N* is the number of tuples in the training set. *f* function finds the class of the new entry.

To accommodate the case where the query point $x_{new}$, exactly matches one of the training instances $x_i$ and the denominator $d(x_i, x_{new})^2$ is therefore zero, we assign $f(x_{new})$ to be $f(x_i)$ in this case. If there are several such training examples, we assign the majority classification among them. We can distance-weight the instances for real-valued target functions in a similar way, replacing the final line of the algorithm in this case by this formula (Mitchell, 1997a):

$$f(x_{new}) = \frac{\sum_{i=1}^{N} w_i f(x_i)}{\sum_{i=1}^{N} w_i}$$

In this technique, there is no need to try to detect the appropriate value of *k* which represents the nearest neighbors in the kNN method. The *k* value can be smaller than half of the N value. In our experiments larger *k* values can not give feasible solutions. Because contaminating some tuples which are very far from the new record gives inefficient results. Each of the *N* tuples in the *k* area has a contribution in the calculation. It ought to be noticed that the closer tuples according to their distances to the new record have higher effects, the further tuples have lower effects in the calculation.

Determining the appropriate *k* value is usually difficult in kNN technique. But in this technique, the *k* value is omitted and the all tuples in the dataset are contaminated to the calculation. In other words, each record in the training set has either low or high effect on calculations. Thus, this method seems to be the best of all examined so far. To illustrate, an example is presented in the figure below:

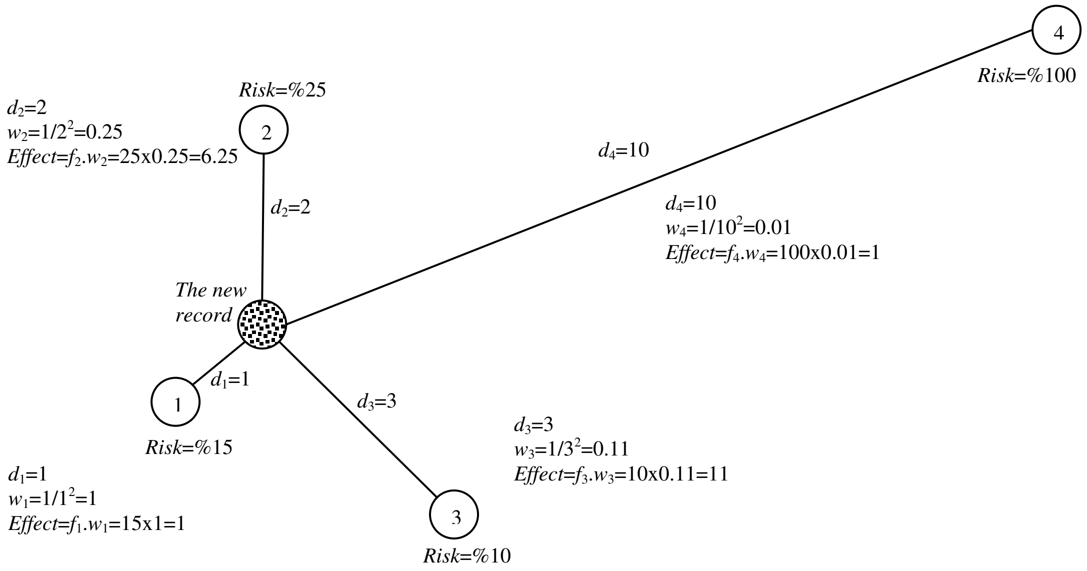


**Figure 4.4** An example to Distance Weighted kNN (k=10).

The 4th tuple has the lowest effect on the new record although its risk is 100%. Its effect is just 1 because of the long distance between itself and the new record. The 1st, 2nd, and 3rd tuples have higher effects on the new record. Their effects are 15, 6.25 and 11 respectively. The risk value of the new record is found by the effects of all the tuples in this formula:

$$f(x_{new}) = \frac{\sum_{i=1}^{N} w_i f(x_i)}{\sum_{i=1}^{N} w_i} = \frac{15+6.25+11+1}{\frac{1}{1}+\frac{1}{4}+\frac{1}{9}+\frac{1}{100}} = 24.27$$

Let's assume that the 4th tuple, whose risk value is 100, is not contaminated to the calculation and try to see the risk value for the new record.

$$f(x_{new}) = \frac{\sum_{i=1}^{N} w_i f(x_i)}{\sum_{i=1}^{N} w_i} = \frac{15+6.25+11}{\frac{1}{1}+\frac{1}{4}+\frac{1}{9}} = 23.71$$

As it is obviously seen above the 4th tuple's effect is very tiny, 0.56. The 4th tuple has an 0.56 increment on the new record due to its long distance.

$$f_2(x_{new}) - f_1(x_q) = 24.27 \text{ - } 23.71 = 0.56$$

There is another example below that illustrates the higher effects of drug users who are very close to the new record and the lower effects of non drug users who are very far from the new record. The 1st tuple's effect as it is seen in the figure is just 0.1.

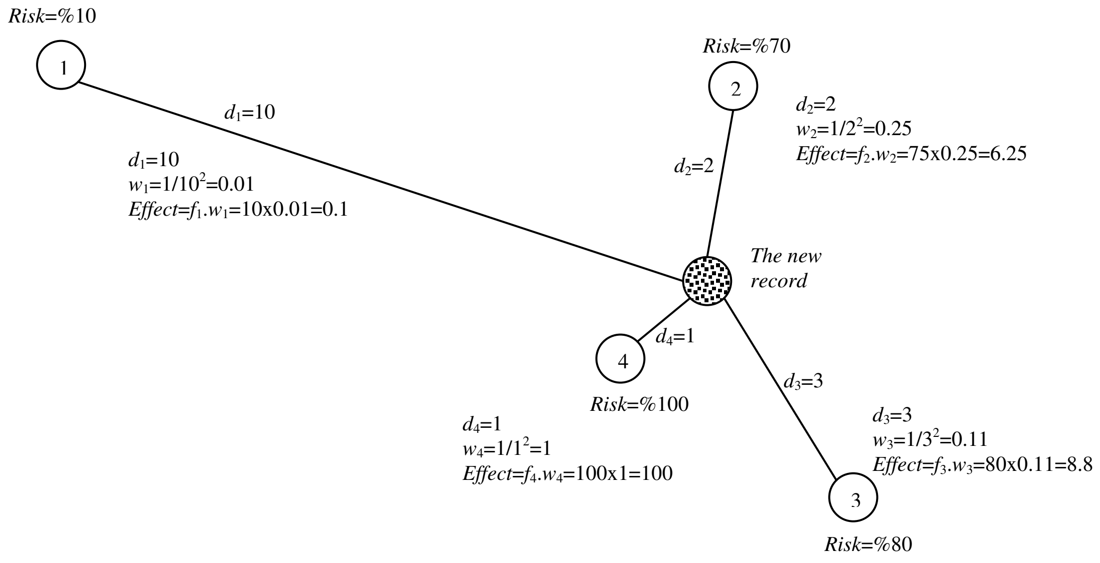


**Figure 4.5** Another example to Distance Weighted kNN (k=10).

$$f(x_{new}) = \frac{\sum_{i=1}^{N} w_i f(x_i)}{\sum_{i=1}^{N} w_i} = \frac{0.1 + 6.25 + 8.8 + 100}{\frac{1}{100} + \frac{1}{4} + \frac{1}{9} + \frac{1}{1}} = 84.05$$

The risk factor of the new record is calculated as 84.05. As it is seen the closer tuples has higher effects on the new record.

Let's assume the 1st tuple is not contaminated to the calculation and try to see the risk value for the new record.

$$f(x_q) = \frac{\sum_{i=1}^{N} w_i f(x_i)}{\sum_{i=1}^{N} w_i} = \frac{6.25 + 8.8 + 100}{\frac{1}{4} + \frac{1}{9} + \frac{1}{1}} = 84.59$$

As it is obviously seen above the 1st tuple's effect on the new record is very tiny, 0.54.

$$f_2(x_{new}) - f_1(x_{new}) = 84.59 - 84.05 = 0.54$$

In this technique, each tuple, whose distances less then or equal to *k*, has an effect on the new record, which makes it better. It is robust and quite effective when there is a large training dataset.

### 4.4.6. Distance and Attribute Weighted kNN

This kNN version is the most advanced and efficient one of all. In the previous versions of kNN methods, one problem is the equal effect of all attributes in calculating the distance between the new tuple and the available tuples in training set. Some of the attributes of a tuple should be less important to the classification and some of them should be more important. This results in misleading of classification process and decreasing the accuracy of classification algorithm. A major approach to deal with this problem is to weight each of the attributes differently when calculating the distance between two records. In this technique a combined method is used to improve the accuracy of kNN (Moradian et al, 2009).

For example, smoking cigarette, being a heavy drinker, being a member of a problematic family are the main factors leading the youth to addiction. That is why; these attributes are more effective than others for classifying upcoming records.

*y* is in the training set and *x* is the new entry waiting to be placed into a class. *x* and *y* are composed of *N* features (attributes). *C* identifies the weights (cost) of *N* attributes of each record in the training set. *dist* has a formula that figures out the distance between the tuples *x* and *y*.

$$dist(x, y) = \sum_{i=1}^{N} \sqrt{((c_i * (x_i - y_i))^2} \text{ where}$$

$x = \{x_1, x_2, x_3, \dots x_N\}$

$y = \{y_1, y_2, y_3, \dots y_N\}$

$C=\{c_1, c_2, c_3, \dots, c_N\}$

In the previous versions of kNN, the all weights of N attributes were equal to one another ($c_1 = c_2 = c_3 = \ldots. = c_N$ ). But in this version, some attributes has higher effects on the other hand some has lower effects.

The *Euclidian Distance* and *Attribute Weighted* formula that finds the effect of *y* on *x* is:

$$f(x) = \underset{v \in V}{\arg\max} \sum_{i=1}^{N} w_i \delta(v, f(y_i)) \text{ where}$$

$$w_i = \frac{1}{dist(x_i, y_i)^2} = \frac{1}{\left( \sum_{i=1}^{N} \sqrt{(c_i * (x_i - y_i))^2} \right)^2}$$

*f* function finds the class of the new entry which is same as in the previous title (Mitchell, 1997a):

$$f(y_i) = \frac{\sum_{i=1}^{N} w_i f(x_i)}{\sum_{i=1}^{N} w_i}$$

To illustrate the Distance and Attribute Weighted kNN technique, an example is presented in the figure below and the risk value is found for the new record.

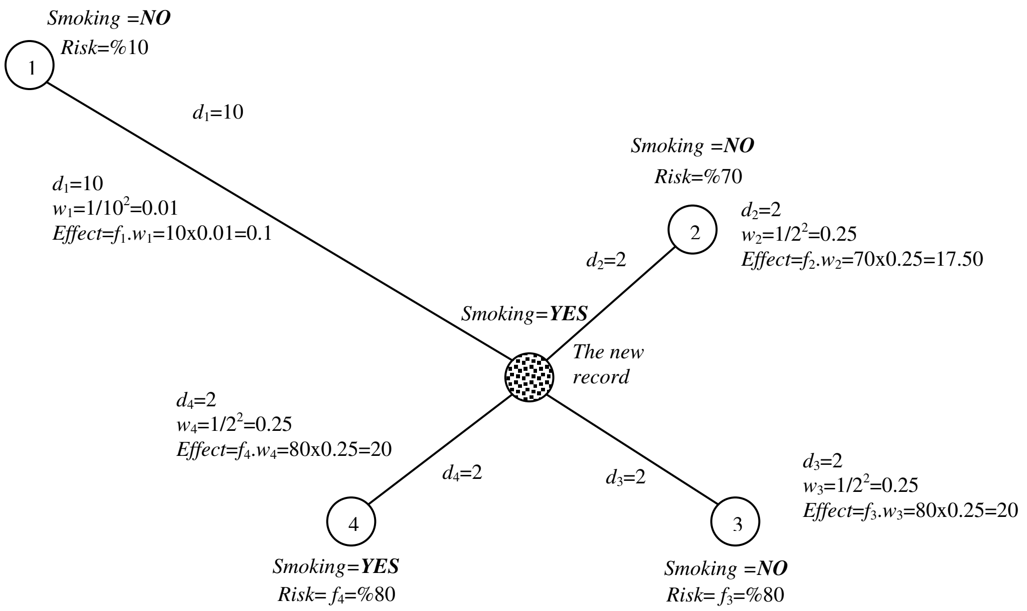


**Figure 4.6** An example to the Distance and Attribute Weighted kNN technique

As it is seen in the figure, only the 4th tuple smokes cigarette. The distances from the 3rd and the 4th tuples to the new record are the same. And the risk values of the 3rd and the 4th tuples are the same. The 4th tuple's effect on the new record will be higher than the 3rd tuple's because the 4th tuple has a smoking habit. The effect of smoking habit is assumed as a little bit greater than the other attributes in the applications. Therefore the effect of 4th tuple will increase to 25.

$$f(y) = \frac{\sum_{i=1}^{N} w_i f(x_i)}{\sum_{i=1}^{N} w_i} = \frac{0.1 + 17.5 + 20 + 25}{\frac{1}{100} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}} = 82.36$$

Without the extra effect of smoking attribute, the risk value for the new record will be found by the distance weighted kNN formula:

$$f(y) = \frac{\sum_{i=1}^{N} w_i f(x_i)}{\sum_{i=1}^{N} w_i} = \frac{0.1 + 17.5 + 20 + 20}{\frac{1}{100} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}} = 75.78$$

In this section, we considered about a new classification version of kNN so as to increase the classification accuracy. In this algorithm, each of attributes is given different weight to the calculation. Each attributes that is more weighted, has more effect on figuring out the distance between two records and each attribute is less weighted has less effect on figuring out the distance. The practical test showed that the suggested algorithm has the more accuracy than previous the algorithms above.

The results of the other methods above have been compared with the results of this combined method. This version of kNN seems to be the best of all. It yields the feasible solutions. The implementation of these types of classification method of kNN is in a file at the attachment.

### 4.4.6.1 Attribute Weights

The weights of the attributes in the training dataset are detected by the help of their information gains. Their information gain ratios are found by Weka (detailed information is under the title, 5.3. Most Relevant Attributes in the Training Set):

```
Information Gain Attribute Evaluation
Instances:    110
Attributes:   23
Ranked attributes:

 0.4635      S22.Disiplin
 0.4337      S24.Sigara
 0.3379      S14.Evebeyn_ilgisi
 0.3346      S12.Internet
 0.2957      S4.KiminleKaliyorsunuz
 0.2904      S9.Kitap
 0.2768      S8.Spor
 0.2524      S19.Kendine_guven
 0.2498      S17.Bos_zaman
 0.2396      S23.Alkolik
 0.2022      S10.Muzik
 0.1826      S1.Yas
 0.1749      S18.Asmak_zorunda
 0.1317      S6.AkrabaZiyareti
 0.118       S13.ArkadasIkram
 0.1135      S3.Evebeyn
 0.095       S7.Enstruman
 0.0897      S5.Gelir
 0.0724      S11.Sinema
 0.0721      S16.Sorun_paylasimi
 0           S15.Gelecekte_hayat
```

As it is seen above, attributes are listed above from the most to the least efficient. The most effective attributes which are the main factors leading adolescents to addiction are at the top of the list. The less effective ones are at the bottom. Because of that the attributes at the top have higher weights.

## 4.5. NAÏVE BAYES CLASSFIER

Naive Bayes Classifier is another classification technique known as 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.

### 4.5.1. 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:

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

To define and examine Bayes Theorem, it is necessary to introduce the notations:

*P*(*h*) is the denote 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 training data *D* will be observed (i.e., the probability of *D* given no knowledge about which hypothesis holds).

Next, *P*(*D* | *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 $P(h \mid D)$ (Mitchell, 1997b).

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.

$P(S \mid D) = 0.944$
$P(D \mid S) = 0.830$

### 4.5.2. 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 { $x_{i1}$ , $x_{i2}$ , $x_{i3}$ , …. $x_{ip}$ }. 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 $(t_i \mid C_j)$ (Dunham, 2005c).

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

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.

There is a table below as an example. This table is a predefined training dataset including 14 tuples. The new coming tuples will be put in a class according to the 14 tuples in the set by using the algorithms. In this dataset, there are four attributes as input. These are age, gender, family income and smoking habit.

**Table 4.2** A training data example for Bayesian Classifier

| **ID** | **Age** | **Gender** (1=Male, 2=Female) | **family income** (1=poor, 2=normal, 3=good, 4=rich) | **Sport activity** (1=yes, 0=no) | **Smoking habit** (1=yes, 0=no) | **Output** | |
|---|---|---|---|---|---|---|---|
| | | | | | | **Risk value** | **Class ( *Ci* )** |
| 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:

$P$ (C1) = 4/14 = 0.286 (C$i$ means Class $i$, e.g. C1 means Class1)
$P$ (C2) = 3/14 = 0.214
$P$ (C3) = 2/14 = 0.143
$P$ (C4) = 3/14 = 0.214
$P$ (C5) = 2/14 = 0.143

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

| | | | | |
|---|---|---|---|---|
| $P\ (1\vert C1)=2/4$ | $P(1\vert C2)=2/3$ | $P(1\vert C3)=0$ | $P(1\vert C4)=2/3$ | $P(1\vert C5)=1/2$ |
| $P\ (0\vert C1)=2/4$ | $P(0\vert C2)=1/3$ | $P(0\vert C3)=2/2$ | $P(0\vert C4)=1/3$ | $P(0\vert C5)=1/2$ |

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

| | | | | |
|---|---|---|---|---|
| $P\ (1\vert C1)=2/4$ | $P(1\vert C2)=3/3$ | $P(1\vert C3)=2/2$ | $P(1\vert C4)=2/3$ | $P(1\vert C5)=1/2$ |
| $P\ (0\vert C1)=2/4$ | $P(0\vert C2)=0$ | $P\ (0\vert C3)=0$ | $P(0\vert C4)=1/3$ | $P(0\vert C5)=1/2$ |

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

| | | | | |
|---|---|---|---|---|
| $P(1\vert C1)=2/4$ | $P(1\vert C2)=1/3$ | $P(1\vert C3)=0$ | $P(1\vert C4)=1/3$ | $P(1\vert C5)=1/2$ |
| $P(2\vert C1)=1/4$ | $P(2\vert C2)=1/3$ | $P(2\vert C3)=1/2$ | $P(2\vert C4)=0$ | $P(2\vert C5)=0$ |
| $P(3\vert C1)=1/4$ | $P(3\vert C2)=1/3$ | $P(3\vert C3)=0$ | $P(3\vert C4)=0$ | $P(3\vert C5)=0$ |
| $P(4\vert C1)=0$ | $P(4\vert C2)=0$ | $P(4\vert C3)=1/2$ | $P(4\vert C4)=2/3$ | $P(4\vert C5)=1/2$ |

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

| | | | | |
|---|---|---|---|---|
| $P(1\vert C1)=2/4$ | $P(1\vert C2)=1/3$ | $P(1\vert C3)=1/2$ | $P(1\vert C4)=1/3$ | $P(1\vert C5)=0$ |
| $P(0\vert C1)=2/4$ | $P(0\vert C2)=2/3$ | $P(0\vert C3)=1/2$ | $P(0\vert C4)=2/3$ | $P(0\vert C5)=2/2$ |

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

| | | | | |
|---|---|---|---|---|
| $P(1\vert C1)=0$ | $P(1\vert C2)=0$ | $P(1\vert C3)=0$ | $P(1\vert C4)=3/3$ | $P(1\vert C5)=2/2$ |
| $P(0\vert C1)=4/4$ | $P(0\vert C2)=3/3$ | $P(0\vert C3)=2/2$ | $P(0\vert C4)=0$ | $P(0\vert C5)=0$ |

**Table 4.3 P**robabilities of the class attributes in a table

| | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| **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* we will use this formula: $P\,(t_i \mid C_j) = \prod_{k=1}^{p} P(x_{ik} \mid C_j)$

Calculations for the tuple *x* :

*P*(*x*|C1)=*P*(Age=17|C1).*P*(Male|C1).*P*(Poor|C1).*P*(Sport=No|C1).*P*(Smoke=Yes|C1)
*P*(*x*|C1)=(2/4).(2/4).(2/4).(2/4).(0)=0
*P*(*x*|C2)=*P*(Age=17|C2).*P*(Male|C2).*P*(Poor|C2).*P*(Sport= No|C2).*P*(Smoke=Yes|C2)
*P*(*x*|C2)=(1/3).(3/3).(1/3).(2/3).(0)=0
*P*(*x*|C3)=*P*(Age=17|C3).*P*(Male|C3).*P*(Poor|C3).*P*(Sport= No|C3).*P*(Smoke=Yes|C3)
*P*(*x*|C3)=(2/2).(2/2).(0).(1/2).(0)=0
*P*(*x*|C4)=*P*(Age=17C4).*P*(Male|C4).*P*(Poor|C4).*P*(Sport= No|C4).*P*(Smoke=Yes|C4)
*P*(*x*|C4)=(1/3).(2/3).(1/3).(2/3).(3/3)=0.049
*P*(*x*|C5)=*P*(Age=17|C5).*P*(Male|C5).*P*(Poor|C5).*P*(Sport= No|C5).*P*(Smoke=Yes|C5)
*P*(*x*|C5)=(1/2).(1/2).(1/2).(2/2).(2/2)=0.125

The result of the probability *P*(*x*|C5) is the biggest one of all. Therefore the *x* record will be put into the class 5.

Calculations for the tuple *y* :

$P(y|C1)=P(\text{Age}=13|C1).P(\text{Female}|C1).P(\text{Rich}|C1).P(\text{Sport}=\text{No}|C1).P(\text{Smoke}=\text{No}|C1)$
$P(y|C1)=(2/4).(2/4).(0).(2/4).(4/4)=0$
$P(y|C2)=P(\text{Age}=13|C2).P(\text{Female}|C2).P(\text{Rich}|C2).P(\text{Sport}=\text{No}|C2).P(\text{Smoke}=\text{No}|C2)$
$P(y|C2)=(2/3).(0).(0).(2/3).(3/3)=0$
$P(y|C3)=P(\text{Age}=13|C3).P(\text{Female}|C3).P(\text{Rich}|C3).P(\text{Sport}=\text{No}|C3).P(\text{Smoke}=\text{No}|C3)$
$P(y|C3)=(0).(0).(1/2).(1/2).(2/2)=0$
$P(y|C4)=P(\text{Age}=13C4).P(\text{Female}|C4).P(\text{Rich}|C4).P(\text{Sport}=\text{No}|C4).P(\text{Smoke}=\text{No}|C4)$
$P(y|C4)=(2/3).(1/3).(2/3).(2/3).(0)=0$
$P(y|C5)=P(\text{Age}=13|C5).P(\text{Female}|C5).P(\text{Rich}|C5).P(\text{Sport}=\text{No}|C5).P(\text{Smoke}=\text{No}|C5)$
$P(y|C5)=(1/2).(1/2).(1/2).(2/2).(0)=0$

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 seen above, this classification method gives only the class name as an output. Sometimes the probabilities of some classes are computed as 0. Zero probabilities are always painful for the Naïve Bayes Classifier. Thus, this makes a barrier to predict the class of the new records.

In our experiments, the Weka output of the Bayesian Classifier gives lower accuracies than others. Correctly Classified Instances are 77.27% in this method. This algorithm's detailed outputs are at the attachment as a text file.

## 4.6. DECISION TREE ALGORITHM

In this section C4.5 and ID3 algorithms are implemented to construct a decision tree so as to classify upcoming tuples. Before examining the algorithms and their details, decision tree will be defined first.

### 4.6.1. What is Decision Tree?

Decision tree is used to classify new instances by sorting them down the tree from the root to some leaf nodes. In other words, a decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node at the bottom represents a predefined class as a final decision.

Decision tree are commonly used for classifying the new tuples according to the predefined training dataset. Beside this, it is used for gaining information for the purpose of decision making. Decision tree starts with a root node on which it is for users to take first actions. From this node, users split each node recursively according to the answer of the question in the current node. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. When reaching to a leaf node at the end this will be the class as a final decision for the new tuple (Peng et al, 2010).

### 4.6.2. ID3 Algorithm

ID3 is a useful and common decision tree constructer algorithm developed by Ross Quinlan in 1983. The basic idea of ID3 algorithm is to construct a decision tree by applying a top-down and greedy search technique through the given sets to test each attribute at every tree node. The built tree models the classification problem according to the training dataset. After building the tree, each tuple in database is applied to the tree and put into an output class. At that point, the depth of the tree ought to be shorter in order to get efficient results in a shorter period of time.

To construct a decision tree on vertically partitioned data requires some calculations. The main problem is computing which site has the best attribute to split on – each can compute the gain of their own attributes without reference to the other site. For this reason *entropy* and *information gain* will be defied below.

#### 4.6.2.1. Entropy

In information theory, entropy is used to measure the amount of uncertainty or surprise or randomness in a set of data. In other words, it is a measure which characterizes the impurity of an arbitrary collection of examples. Certainly, when all the data in a dataset belong to a single class, there is no uncertainty. In this case entropy is 0. The value for entropy is always between 0 and 1. And it reaches to a maximum value when the probabilities are all the same. Obviously this maximum value is 1.

If the target attribute takes on $c$ different values, then the entropy $S$ is defined as

$$Entropy(S) = \sum_{i=1}^{c} (-p_i \log_2 p_i)$$

where $p_i$ is the proportion/probability of S belonging to class $i$. Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits. For example if training dataset has 14 instances with 6 drug users and 8 non drug users tuples, the entropy is calculated as

$$\text{Entropy }([6+, 8-]) = -(6/14)\log_2(6/14) - (8/14)\log_2(8/14) = 0.985$$

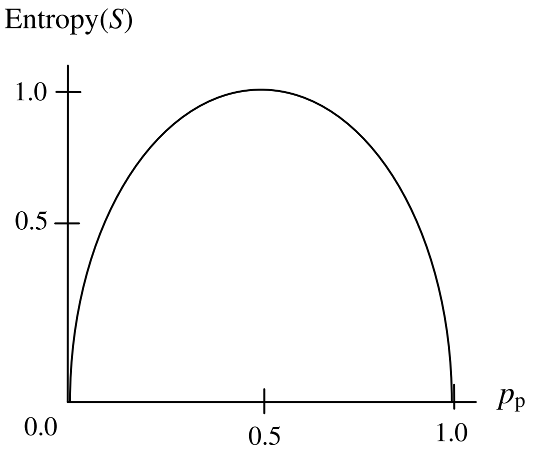


**Figure 4.7** Definition of entropy with a figure (Mitchell, 1997c):

S is a sample of training examples. $p_p$ is the proportion of positive examples in the set S. $p_n$ is the proportion of positive examples in the set S. Entropy measures the impurity of S as

$$Entropy(S) = -p_p \log_2 p_p - p_n \log_2 p_n$$

#### 4.6.2.2. Information gain

Information Gain measures the expected reduction in entropy by partitioning the examples according to this attribute. The formula of information gain, Gain($S$, $A$) of an attribute $A$, relative to the collection of examples $S$, is defined as

$$\text{Gain(S, A)} = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where *Values*(*A*) is the set of all possible values for attribute *A*, and $S_v$ is the subset of *S* for which the attribute *A* has value *v*. This measure is used to rank attributes and build the decision tree where at each node is located the attribute with the highest information gain among the attributes (Bahety, 2010).

### 4.6.2.3. Building the tree

ID3 in this study is used as a builder of the decision tree so as to classify the new records. ID3 is a recursive partitioning algorithm. At first, all the training examples are at the root. Examples are then partitioned recursively based on selected attributes. ID3 has a greedy technique – in each case the attribute with the highest information gain is selected as the partitioning attribute. Partitioning ends either when all samples for a given node belong to the same class, or there are no remaining attributes, or there are no samples left (Aggarwal et al, 2008b).

The gain is used to rank attributes and build the tree where at each node is located the attributes with greatest gain among the attributes not yet considered in the path from the root. The aim of the ordering is to create small trees.

```
ID3 (Learning Sets S, Attributes Sets A, Attributes values V)
Return Decision Tree.

Begin
Load learning sets first, create decision tree root node 'rootNode', add learning set S into
root node as its subset.
For rootNode, we compute Entropy (rootNode.subset) first
        If Entropy(rootNode.subset)==0, then
        rootNode.subset consists of records all with the same value for the categorical
        attribute, return a leaf node with decision attribute:attribute value;

        If Entropy(rootNode.subset)!=0, then
        Compute information gain for each attribute left (have not been used in
        splitting), find attribute A with

        Maximum(Gain(S,A)). Create child nodes of this rootNode and add to
        rootNode in the decision tree.

        For each child of the rootNode, apply ID3(S,A,V) recursively until reach node
        that has entropy=0 or reach leaf node.
End ID3.
```

### 4.6.3. Weka Application of ID3 Decision Tree

The decision trees are built by Weka, the popular and effective machine learning toolkit program. All the attributes in the training set has a part in the decision tree calculations. Some of the attributes according to their less importance are extracted from the calculations by this program (Ian, 2005). ID3 and C4.5 as decision tree algorithms are applied. The Weka output of ID3 according to the training set is as follows:

```
=== Run information ===
Scheme:       weka.classifiers.trees.Id3
Relation:     Madde_Bagimliligi
Instances:    110
Attributes:   23
               S1.Yas,                S2.Cinsiyet
               S3.Evebeyn,            S4.KiminleKaliyorsunuz
               S5.Gelir,              S6.AkrabaZiyareti
               S7.Enstruman,          S8.Spor
               S9.Kitap,              S10.Muzik
               S11.Sinema,            S12.Internet
               S13.ArkadasIkram,      S14.Evebeyn_ilgisi
               S15.Gelecekte_hayat,   S16.Sorun_paylasimi
               S17.Bos_zaman,         S18.Asmak_zorunda
               S19.Kendine_guven,     S22.Disiplin
               S23.Alkolik,           S24.Sigara
               class

Test mode:     evaluate on training data

=== Classifier model (full training set) ===

S22.Disiplin = 1
|  S14.Evebeyn_ilgisi = 1
|  |  S12.Internet = 1
|  |  |  S1.Yas = 15: Class3
|  |  |  S1.Yas = 16: Class4
|  |  |  S1.Yas = 20: Class3
|  |  S12.Internet = 2
|  |  |  S6.AkrabaZiyareti = 1: Class5
|  |  |  S6.AkrabaZiyareti = 2: Class5
|  |  |  S6.AkrabaZiyareti = 3: Class4
|  |  S12.Internet = 3: Class5
|  S14.Evebeyn_ilgisi = 2: Class5
|  S14.Evebeyn_ilgisi = 3
|  |  S8.Spor = 1
|  |  |  S1.Yas = 14: Class5
|  |  |  S1.Yas = 15: Class3
|  |  |  S1.Yas = 16: Class4
|  |  |  S1.Yas = 17: Class4
|  |  |  S1.Yas = 18
|  |  |  |  S6.AkrabaZiyareti = 1: Class4
|   |   |   |   S6.AkrabaZiyareti = 2: Class3
|   |   |   |   S6.AkrabaZiyareti = 3: Class3
|   |   S8.Spor = 0
|   |   |   S15.Gelecekte_hayat = 2: Class5
|   |   |   S15.Gelecekte_hayat = 3
|   |   |   |   S5.Gelir = 1: Class5
|   |   |   |   S5.Gelir = 2
|   |   |   |   |   S9.Kitap = 2: Class5
|   |   |   |   |   S9.Kitap = 3: Class3
|   |   |   |   |   S9.Kitap = 4: Class5
|   |   |   |   S5.Gelir = 4: Class3
|   |   |   S15.Gelecekte_hayat = 4
|   |   |   |   S1.Yas = 17: Class3
|   |   |   |   S1.Yas = 18: Class3
|   S14.Evebeyn_ilgisi = 4
|   |   S13.ArkadasIkram = 1: Class5
|   |   S13.ArkadasIkram = 2
|   |   |   S1.Yas = 15: Class4
|   |   |   S1.Yas = 18
|   |   |   |   S6.AkrabaZiyareti = 1: Class4
|   |   |   |   S6.AkrabaZiyareti = 2: Class5
|   |   |   S1.Yas = 20: Class4
|   |   S13.ArkadasIkram = 3: Class5
S22.Disiplin = 0
|   S1.Yas = 11: Class1
|   S1.Yas = 12
|   |   S5.Gelir = 1: Class2
|   |   S5.Gelir = 2: Class1
|   S1.Yas = 13
|   |   S16.Sorun_paylasimi = 1: Class1
|   |   S16.Sorun_paylasimi = 2: Class2
|   |   S16.Sorun_paylasimi = 5: Class1
|   |   S16.Sorun_paylasimi = 6: Class5
|   S1.Yas = 14
|   |   S10.Muzik = 1: Class3
|   |   S10.Muzik = 2: Class2
|   |   S10.Muzik = 3: Class3
|   S1.Yas = 15
|   |   S17.Bos_zaman = 1: Class1
|   |   S17.Bos_zaman = 2: Class2
|   |   S17.Bos_zaman = 3
|   |   |   S13.ArkadasIkram = 1: Class4
|   |   |   S13.ArkadasIkram = 2: Class1
|   |   |   S13.ArkadasIkram = 3: Class3
|   |   S17.Bos_zaman = 4: Class2
|   S1.Yas = 16
|   |   S16.Sorun_paylasimi = 1: Class1
|   |   S16.Sorun_paylasimi = 2: Class3
|   |   S16.Sorun_paylasimi = 3: Class1
|   |   S16.Sorun_paylasimi = 5
|   |   |   S5.Gelir = 1: Class4
|   |   |   S5.Gelir = 2: Class1
|   S1.Yas = 17
|   |   S16.Sorun_paylasimi = 1
|  |  |  S6.AkrabaZiyareti = 1: Class4
|  |  |  S6.AkrabaZiyareti = 4: Class3
|  |  S16.Sorun_paylasimi = 2: Class1
|  |  S16.Sorun_paylasimi = 3: Class1
|  |  S16.Sorun_paylasimi = 4: Class3
|  |  S16.Sorun_paylasimi = 5
|  |  |  S11.Sinema = 1: Class4
|  |  |  S11.Sinema = 2: Class1
|  |  |  S11.Sinema = 4: Class3
|  |  S16.Sorun_paylasimi = 6: Class3
|  S1.Yas = 18
|  |  S5.Gelir = 1: Class3
|  |  S5.Gelir = 3: Class5
|  S1.Yas = 19
|  |  S9.Kitap = 1: Class3
|  |  S9.Kitap = 2: Class5
|  |  S9.Kitap = 3: Class4
|  S1.Yas = 20: null

=== Evaluation on training set ===

Correctly Classified Instances         107         97.3 %
Incorrectly Classified Instances         3          2.7 %

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
1         0          1          1         1         1        Class1
1         0          1          1         1         1        Class2
1         0.034      0.875      1         0.933     0.998    Class3
0.885     0          1          0.885     0.939     0.998    Class4
1         0          1          1         1         1        Class5
Weighted Avg.
0.973     0.006      0.976      0.973     0.973     0.999

=== Confusion Matrix ===
      a  b  c  d  e   <-- classified as
     17  0  0  0  0 |  a = Class1
      0  5  0  0  0 |  b = Class2
      0  0 21  0  0 |  c = Class3
      0  0  3 23  0 |  d = Class4
      0  0  0  0 41 |  e = Class5
```

### 4.6.4. Weka Application of C4.5 Decision Tree

There is a pruned tree constructed by C4.5 Algorithm as an extension of ID3. In this algorithm all the attributes in the questionnaire are included in the calculation. The Weka output of pruned C4.5 decision tree is as follows:

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 (Application of C4.5)
Instances:    110
Attributes:   23

C4.5 (J48) pruned tree
----------------------

    S22.Disiplin = 1
    |   S23.Alkolik = 1: Class5
    |   S23.Alkolik = 0
    |   |   S24.Sigara = 1
    |   |   |   S8.Spor = 1
    |   |   |   |   S2.Cinsiyet = 1
    |   |   |   |   |   S15.Gelecekte_hayat <= 3
    |   |   |   |   |   |   S7.Enstruman = 1: Class3
    |   |   |   |   |   |   S7.Enstruman = 0: Class4
    |   |   |   |   |   S15.Gelecekte_hayat > 3
    |   |   |   |   |   |   S1.Yas <= 17: Class4
    |   |   |   |   |   |   S1.Yas > 17: Class5
    |   |   |   |   S2.Cinsiyet = 2: Class5
    |   |   |   S8.Spor = 0
    |   |   |   |   S2.Cinsiyet = 1
    |   |   |   |   |   S15.Gelecekte_hayat <= 3: Class5
    |   |   |   |   |   S15.Gelecekte_hayat > 3: Class3
    |   |   |   |   S2.Cinsiyet = 2: Class3
    |   |   S24.Sigara = 0: Class5
    S22.Disiplin = 0
    |   S24.Sigara = 1
    |   |   S13.ArkadasIkram = 1: Class3
    |   |   S13.ArkadasIkram = 2: Class3
    |   |   S13.ArkadasIkram = 3: Class4
    |   S24.Sigara = 0
    |   |   S16.Sorun_paylasimi <= 5
    |   |   |   S13.ArkadasIkram = 1: Class1
    |   |   |   S13.ArkadasIkram = 2: Class1
    |   |   |   S13.ArkadasIkram = 3
    |   |   |   |   S2.Cinsiyet = 1: Class3
    |   |   |   |   S2.Cinsiyet = 2: Class2
    |   |   S16.Sorun_paylasimi > 5: Class5

Number of Leaves  : 18
Size of the tree  : 33

=== Evaluation on training set ===
Correctly Classified Instances        95            86.4 %
Incorrectly Classified Instances      15            13.6 %
```

Detailed outputs of the C4.5 technique is in a text file at the attachment. In the experiments, the accuracy of ID3 is found surprisingly greater than C4.5's. In fact, it is expected that the fact that the accuracy of C4.5 should be greater.

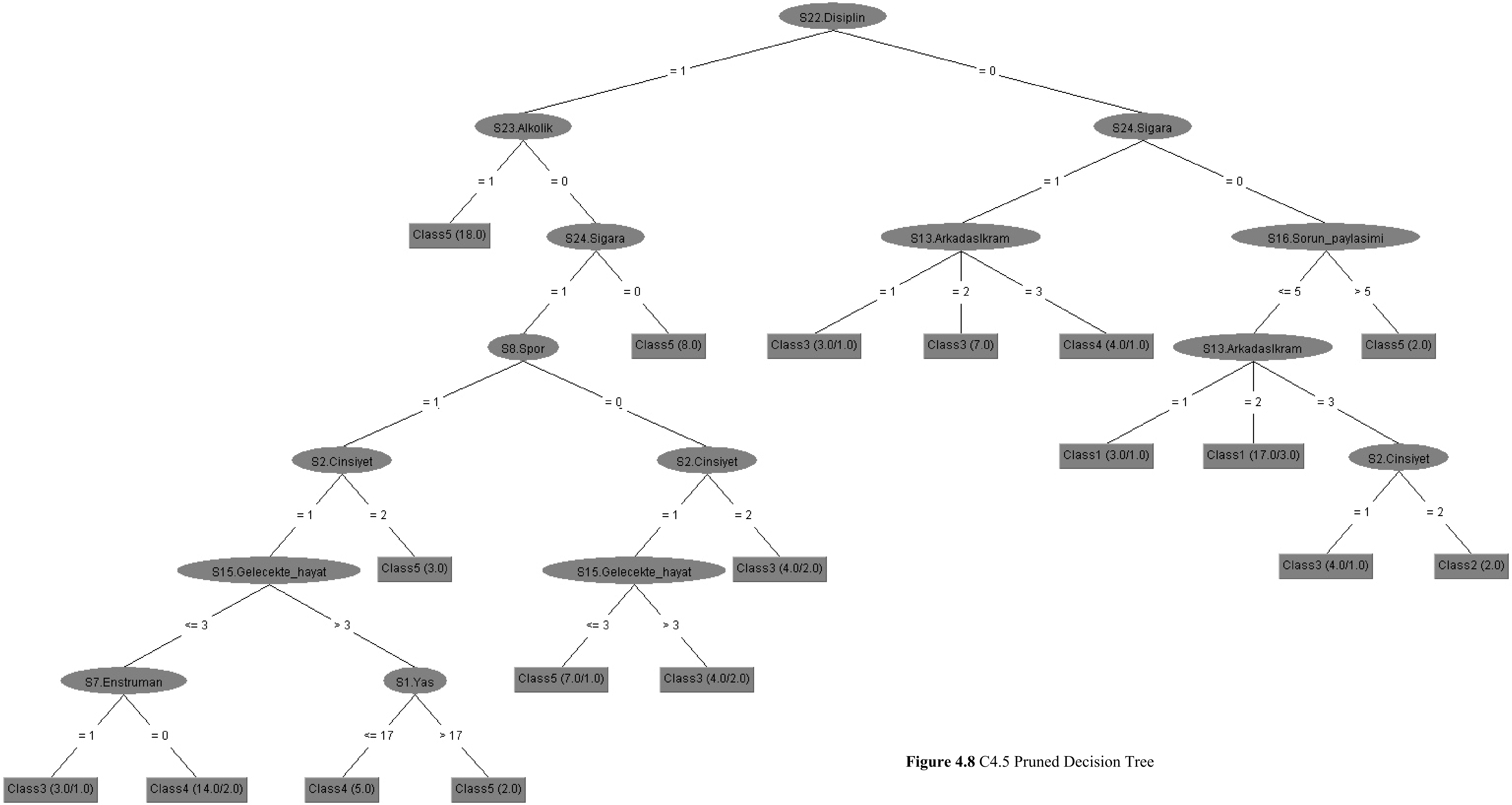


**Figure 4.8** C4.5 Pruned Decision Tree

As it is seen above, a decision tree classifies new instances quite fast and is evaluated even faster than the other algorithms. For example an instance *x*, has given to the answers to the questions as:

```
x = <S1.Yas=15, S2.Cinsiyet=1, S3.Evebeyn=1,
S4.KiminleKaliyorsunuz=1, S5.Gelir=2, S6.AkrabaZiyareti=1,
S7.Enstruman=0, S8.Spor=0, S9.Kitap=1, S10.Muzik=1,
S11.Sinema=1, S12.Internet=1, S13.ArkadasIkram=2,
S14.Evebeyn_ilgisi=1, S15.Gelecekte_hayat=2,
S16.Sorun_paylasimi=3, S17.Bos_zaman=1, S18.Asmak_zorunda=1,
S19.Kendine_guven=1, S22.Disiplin=0, S23.Alkolik=0,
S24.Sigara=0>
```

First of all, the first question at the root of the decision tree will be answered. As the reply of the tuple *x* to that question is `S22.Disiplin=0,` we will go to the right side. Since the answer to the next question at that node is `S24.Sigara=0` we will go to the right side again. The answer to `S16.Sorun_paylasimi` is 3, then we will go to the left side to answer the question, `S13.ArkadasIkram.` Since the answer is 2, the output class for the tuple *x* is Class1.

## 4.7. RULE BASED ALGORITHMS

### 4.7.1. One-Attribute-Rule Algorithm

One simple approach in classification algorithms is OneR (one-attribute-rule, or shortly 1R) as an attribute based algorithm proposed by Holte (Holte, 1993). The basic idea in this technique is to choose the best attribute to perform the classification based on the training data. Only one attribute (because of that it is called 1R) is assumed sufficient to classify new coming records. The idea of the OneR algorithm is to find the one attribute to use that makes fewest prediction errors. In this method, some calculations should be done to find the "best" attribute. "Best" is defined by counting the number of *errors* for each attribute. The one which has less total *error* of all values will be the "best" attribute for the OneR algorithm. In other words, OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen randomly.

The OneR algorithm creates one rule for each attribute in the training dataset and then selects the rule with the smallest error rate as its “one rule”. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class; one such binding for each attribute value of the attribute the rule is based on (Buddhinath et al, 2010). The pseudo-code for the OneR algorithm is as follows:

```
Begin

    For each attribute $A$,
        For each value $V_A$ of the attribute, make a rule as follows:
            count how often each class appears
            find the most frequent class $C_f$
            create a rule when $A = V_A$; class attribute value = $C_f$
        End For-Each.
        Calculate the error rate of all rules
    End For-Each.

Chose the rule with the smallest error rate

End.
```

The Weka output for OneR algorithm is in a text file at the attachment. Its summary is as follows:

```
=== Run information ===

Relation:    Madde_Bagimliligi
Instances:   110
Attributes:  23

=== Classifier model : OneR ===

S22.Disiplin:
    1 -> Class5
    0 -> Class1

(55/110 instances correct)

Correctly Classified Instances          55    50%
Incorrectly Classified Instances        55    50%
```

As it is seen above, the 22nd question in the questionnaire is the best attribute (minimum-error attribute) that is used for prediction.

### 4.7.2. PART Algorithm

PART is a separate-and-conquer rule classifier proposed by Eibe and Witten. This algorithm producing sets of rules called '*decision lists*' which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule. PART builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning (Eibe, 1998).

The Weka output of PART algorithm is in a text file at the attachment. Its summary is as follows:

```
=== Run information ===
Relation:     Madde_Bagimliligi
Instances:    110
Attributes:   23
=== Classifier model : PART decision list ===

S22.Disiplin = 1 AND S23.Alkolik = 1                  : Class5
S24.Sigara = 1 AND S19.Kendine_guven = 1              : Class5
S24.Sigara = 1 AND S8.Spor = 1 AND
S22.Disiplin = 0 AND S13.ArkadasIkram = 2             : Class3
S24.Sigara = 1 AND S8.Spor = 1 AND
S14.Evebeyn_ilgisi = 3 AND S1.Yas <= 17               : Class4
S22.Disiplin = 1 AND S24.Sigara = 0                   : Class5
S24.Sigara = 1 AND S5.Gelir = 2 AND S11.Sinema = 2    : Class3
S24.Sigara = 1 AND S5.Gelir = 2 AND S11.Sinema = 3    : Class5
S24.Sigara = 1 AND S5.Gelir = 1 AND S22.Disiplin = 0  : Class3
S24.Sigara = 1 AND S5.Gelir = 1 AND S18 = 2           : Class4
S24.Sigara = 1 AND S5.Gelir = 1 AND
S11.Sinema = 4 AND S12.Internet = 2 AND S1.Yas <= 19  : Class5
S24.Sigara = 1 AND S11.Sinema = 4 AND
S14.Evebeyn_ilgisi = 3                                : Class3
S24.Sigara = 1 AND S5.Gelir = 1 AND S8.Spor = 1       : Class4
S24.Sigara = 1 AND S5.Gelir = 1                       : Class5
S24.Sigara = 0 AND S16.Sorun_paylasimi <= 5 AND
S13.ArkadasIkram = 2                                  : Class1
S17.Bos_zaman = 3 AND S1.Yas > 13                     : Class3
S12.Internet = 3                                      : Class5
S15.Gelecekte_hayat > 3                               : Class2

Number of Rules :   18
Correctly Classified Instances:     98    89 %
Incorrectly Classified Instances:   12    11 %
```

As it is seen above, as its accuracy is 89%, it is higher than OneR.

## 4.8. USING GRAPH THEORIES

Peer pressure plays a great role in leading a person to addiction (Kırcan, 2006), (Kimmel, 1995), (Windle, 2003). In the questionnaires, their best friends are asked to the school students for the purpose of indentifying the friendships among them.

To illustrate that, the relations among students are shown in a graph as an example in the figure below. It is an undirected and unweighted graph. In the graph each node which has a number in it represents the ID number of a record in the database. The edge between two nodes represents the adjacency, in other words it indicates the friendship between these two friends.

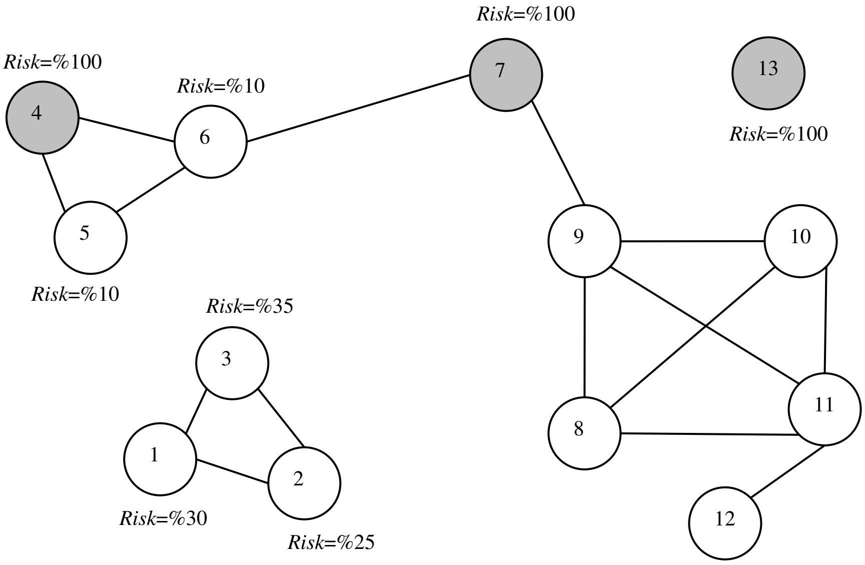


**Figure 4.9** Graph representation of friendships.

As it is seen in this figure, the 4th, 7th, and 13th nodes are drug abusers. Although the 13th person is absolutely a drug user, he/she has no bad effect on the others because he has no relationship with other students.

1st, 2nd, and 3rd students have a friend group. Their risks are 30, 25, and 30 respectively. None of them is a drug user. Thus, there is not a big risk for these students.

5th, 6th, and 9th friends are highly at risk because they have friendship with the addicted students.

8th, 9th, 10th, and 11th have an adjacency among them. Only 9th student is at risk because of the 7th student. The 9th student has higher risk of addiction than the others.

The relationships among students represented in the figure above are put into a two dimensional array. In order to keep the graph with its nodes and edges, two dimensional array is used to store the nodes and adjacencies so as to implement the BFS algorithm. To illustrate, there is an array below which keeps the graph in the memory.

**Table 4.4** Matrix graph representation using two dimensional array

| Nodes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Breadth First Search (BFS) as a graph search algorithm is used to find the relationship among friends. It begins at a node and explores all the neighboring nodes. Then for each of those nearest nodes, it explores their unexplored neighbor nodes. BFS is an uninformed search technique that aims to expand and examine all nodes of a graph by systematically searching through every solution. In other words, it exhaustively searches the entire graph or sequence without considering the goal. All child nodes obtained by expanding a node are added to a FIFO (First In First Out) queue data structure.

The algorithm of BFS is as follows (Wikipedia, 2010b):

1. Enqueue a node in the graph (put it into the queue).
2. Dequeue a node and examine it (take a node from the queue).
    i. If the element is found in this node, quit the search and return a result. (Or continue to search)
    ii. Otherwise enqueue any successors (the direct child nodes) that have not yet been discovered.
3. If the queue is empty AND every node on the graph has been examined – quit the search and return "not found".
4. Repeat from Step 2.

# CHAPTER 5

# EXPERIMENTAL RESULTS, EVALUATION, COMPARISONS

## 5.1. Implementation of the Algorithms

The C programming language is used on the Windows environment in order to implement the classifier algorithms. The compiler is Devcpp v4.9 as an ANSI C Compiler (It can be downloaded from www.bloodshed.net). After recording all the questionnaires to the database `db0.txt`, the output files as text formats are created by the codes in the files a the attachment. The Weka v3.6.0. program is also used so as to check the results and measure the accuracies of the algorithms.

Even though implementing the algorithms in C programming language is very difficult, it provides us many benefits. First, the different versions of kNN algorithms can only be possible by this way. Especially applying the last two version, *Distance Weighted kNN* and *Attribute and Distance Weighted kNN*, can be possible in C. Programming gives us a flexible opportunity and scalability. This could be very difficult in Weka or in another program.

## 5.2. Input and Output Files

Our implemented codes need some input files to run. Each algorithm gives some outputs. Input and generated output files are listed below.

**DB0.txt** : The initial (raw) database file. The new coming records take place at the end of file. All the attributes are numeric ranging from 0 to 20. -1 represents missing value.

**TrainingSet.txt** : In this file there are 110 tuples. 37 of them are definitely drug users. The others are in Class2, Class3, Class3 and Class4. The classifier algorithms uses this file to classify upcoming queries.

**DB1.txt** : Some attributes is omitted in the DB0.txt and this file is converted to the DB1.txt file. In this section, some attributes(columns) are deleted. If 33% of the attribute is missing, these attribute are deleted by changing the values with -2

**DB2.txt** : Some missing data in the DB1.txt file is filled with appropriate values. Then it is converted to the DB2.txt

**DB3.txt** : Some noisy data in the DB2.txt file is removed and converted to DB3.txt.

**DB4.txt** : Some outliers and liars in the DB3.txt are cleaned and this file is converted the DB.txt file for the use of algorithms. -2 represents omitted attributes.

**Frequency Table.txt** : Values in this file is the purpose of checking if there is an error or not. In addition, this integer values are used in the Naïve Bayes Technique.

**Missing Data Info.txt** : This file shows the number of missing data of each attribute in the database. The attribute number in the 1st line, number of missing values of the attribute in the 2nd line, the missing percentage of each attribute in the 3rd line are showed in this file. In our experiments the 26th attribute, it is seen that there are too many missing values. This C program deletes this 26th attribute in the database.

**kNN1.txt** : In this file there are outputs of the records in the database by the help of Simple kNN algorithm.

**kNN2.txt** : In this file there are outputs from the algorithm, Absolute Distance kNN.

**kNN3.txt** : In this file there are outputs from the algorithm, Euclidian Distance kNN.

**kNN4.txt** : In this file there are outputs from the algorithm, Distance Weighted kNN.

**kNN5.txt** : In this file there are output classes for the records in the database by using the attribute and distance weighted kNN algorithms.

**Bayesian.txt** : In this file there are outputs of the records in the database by the help of Naïve Bayes Classifier.

**Missing Data Info.txt** : In this file there is a list that represents the extracted attributes from the dataset and training set. In our experiments the 26th has been deleted.

**Liars ID List.txt** : This file shows the liars who have filled out the questionnaires randomly or intentionally.

**Weak TestSet Outputs** : All the output files of the implemented algorithms (Id3, C4.5, kNN, OneR, PART, Naïve Bayes and so on ) in text formats is at the attachment. All the files contains the results.

**Weka Outputs of Algorithms Accuracies** : These files contain the text files of the accuracies of the algorithms.

## 5.3. Most Relevant Attributes in the Training Set

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.

```
Information Gain Attribute Evaluation
Instances:    110
Attributes:   23
Ranked attributes:

 0.4635      S22.Disiplin
 0.4337      S24.Sigara
 0.3379      S14.Evebeyn_ilgisi
 0.3346      S12.Internet
 0.2957      S4.KiminleKaliyorsunuz
 0.2904      S9.Kitap
 0.2768      S8.Spor
 0.2524      S19.Kendine_guven
 0.2498      S17.Bos_zaman
 0.2396      S23.Alkolik
 0.2022      S10.Muzik
 0.1826      S1.Yas
 0.1749      S18.Asmak_zorunda
 0.1317      S6.AkrabaZiyareti
 0.118       S13.ArkadasIkram
 0.1135      S3.Evebeyn
 0.095       S7.Enstruman
 0.0897      S5.Gelir
 0.0724      S11.Sinema
 0.0721      S16.Sorun_paylasimi
 0           S15.Gelecekte_hayat
```

As it is seen above, attributes are listed above from the most to the least efficient. The most effective attributes which are the main factors leading adolescents to addiction are at the top of the list. The less effective ones are at the bottom.

### 5.4. Comparisons among Algorithms

Six different classification algorithms have been applied. They have all given similar results. Criticizing these techniques is appeared below by looking at the output files. Comparing the accuracy of algorithms according to the training set is as follows:

**Table 5.1** Comparing the accuracies of implemented algorithms

| | **OneR** | **PART** | **ID3** | **C4.5** | **Naive Bayes** | **kNN** (k=3) | **kNN** (k=10) | **Distance Weighted kNN** | **Distance & Attribute Weighted kNN** |
|---|---|---|---|---|---|---|---|---|---|
| Correctly classified instances | 50% | 89% | 97.30% | 86.40% | 77.20% | 73.60% | 73.60% | 97.30% | 98.20% |
| Incorrectly classified instances | 50% | 11% | 2.70% | 13.60% | 23.80% | 26.40% | 26.40% | 2.70% | 1.80% |

The most efficient algorithms as it is seen in the table above are Distance Weighted kNN and ID3. 97.30% percentage of the data can be classified correctly. This ration is the biggest one of all.

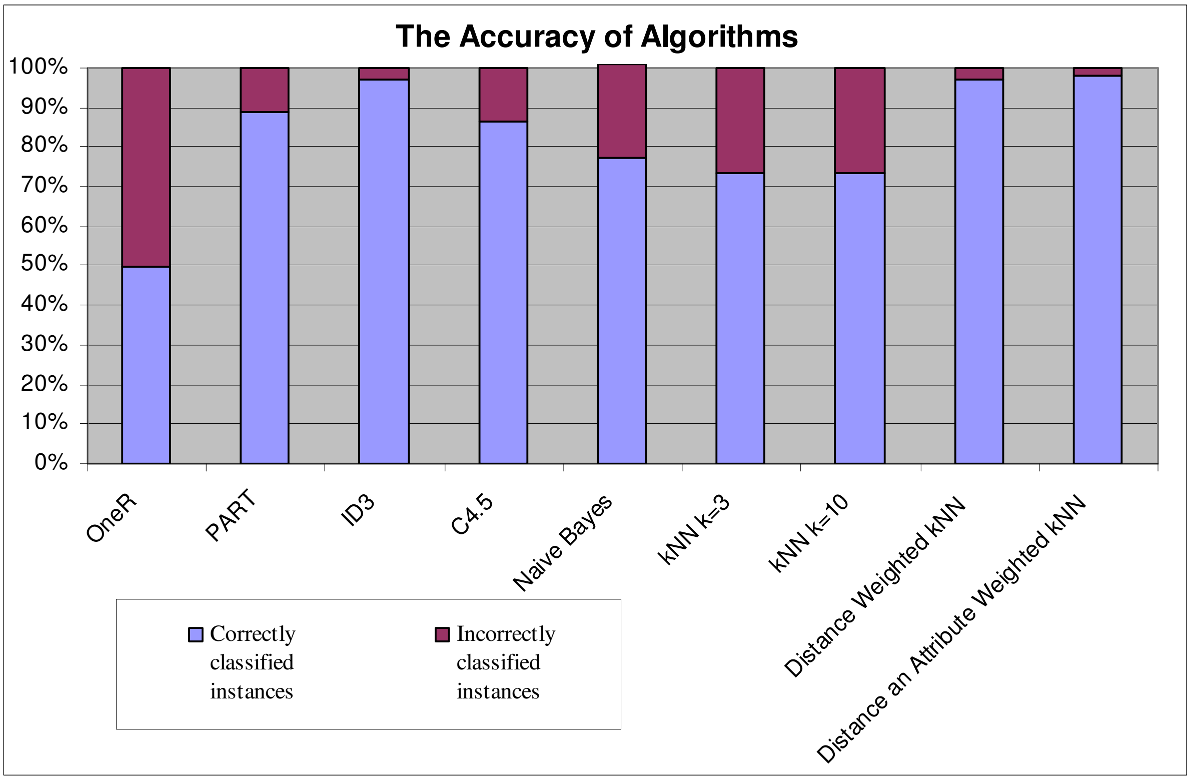


**Figure 5.1** Comparing the accuracies of implemented algorithms

### 5.4.1. Naive Bayes Classifier

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 big barrier to calculations.

But the Naïve Bayes Classifier has given 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 our 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.

In addition, this method requires a well prepared maybe the best prepared training data. All the probabilities and estimations in this technique depend on the training set. There should be more than 2000 tuples in the training set so as to get satisfactory and feasible results.

### 5.4.2. kNN Classifiers

Since kNN is the most flexible and customizable algorithm according to our needs, it can be accepted as the best one. Especially the "*distance and attribute weighted version of kNN*" has given the best results of all. In this method, weighting some particular attributes can be done by only writing the C codes. There has been a chance for us to customize the calculations in C programming. Besides, kNN gives the precise results ranging from 0% to 100%. For example kNN can give some outputs such as 14.56%, 47.59%, and 70.44%. But the other methods give us only the class names such as Class$_1$, Class$_2$, Class$_3$, Class$_4$, and Class$_5$. kNN performances in different *k* values are as follows:

**Table 5.2** Comparing the accuracies of kNN algorithms

| | kNN (k=1) | kNN (k=3) | kNN (k=5) | kNN (k=10) | kNN (k=15) | kNN (k=20) | Distance Weighted kNN | Distance and Attribute Weighted kNN |
|---|---|---|---|---|---|---|---|---|
| Correctly classified instances | 97.30% | 73.60% | 71.80% | 70.00% | 61.80% | 63.60% | 97.30% | 98.20% |
| Incorrectly classified instances | 2.70% | 26.40% | 28.20% | 30.00% | 38.20% | 36.40% | 2.70% | 1.80% |

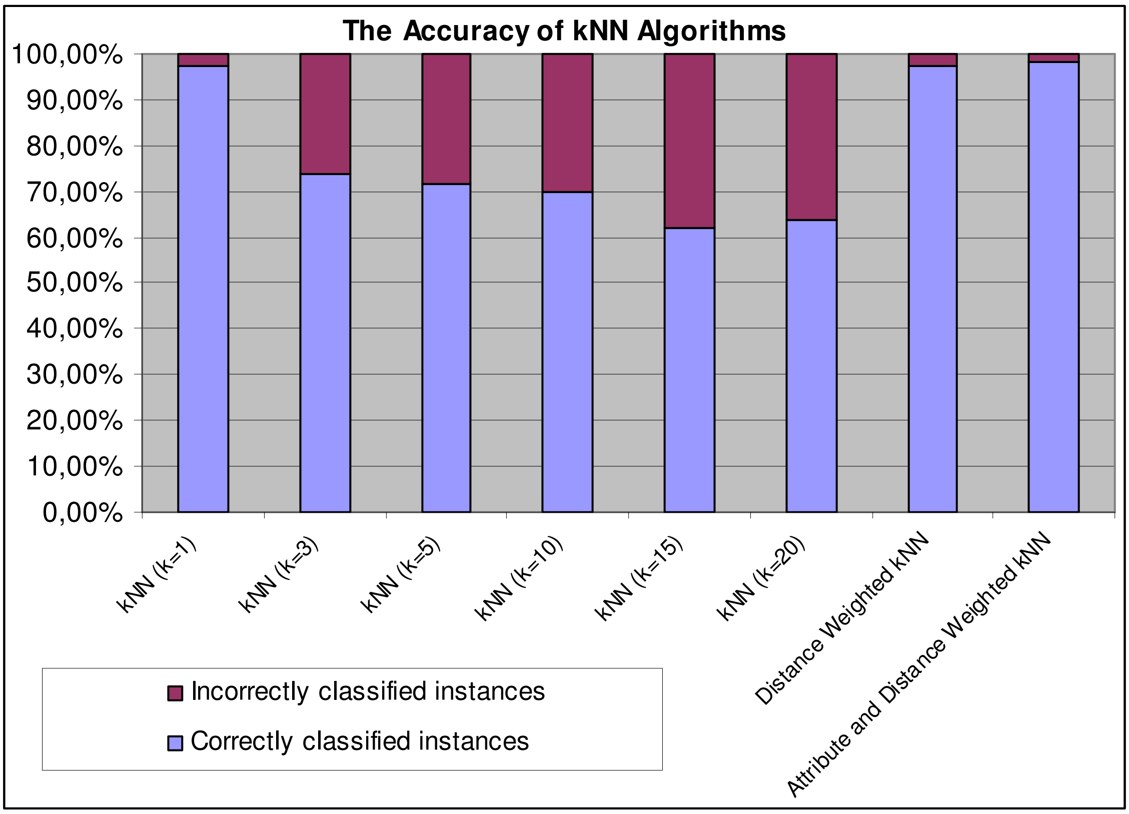


**Figure 5.2** Comparing the accuracies of kNN algorithms

As it is seen above in the table and in the figure, when the *k* value gets higher the accuracy slows down. The best solution can be taken by the Distance and Attribute Weighted kNN algorithm.

### 5.4.3. ID3 and C4.5 Classifiers

Similar to Naïve Bayes Classifier, ID3 depends on the training data. First of all, the decision tree is constructed according to the training set. Then, the decision tree is applied to each tuple in the database. But building the decision tree has required a lot of calculations. The accuracy of DTs is not good as others even though classifying new tuples very easily.

### 5.4.4. OneR and PART Classifiers

Generating a rule according to only one attribute among a lot of attributes should not be accepted sufficient for this study. As it is known, there are many reasons leading a person to addiction. Depending on only one attribute, these algorithms can not be accepted as applicable to this study.

In the experiments, OneR gives us just two possible classes, Class1 and Class5 by checking out the `S.22.Disiplin` attribute. On the other hand OneR ignores the other classes, e.g. Class2, Class3 and Class4.

## 5.5. Outputs of the Implemented Algorithms

In the table, all the tuples in the database are put in five classes by the implemented algorithms. The percentages of the classes for each algorithm are shown below.

**Table 5.3** The percentage of the classes for each algorithm

| | **Class1** | **Class2** | **Class3** | **Class4** | **Class5** |
|---|---|---|---|---|---|
| C4.5 | 66.0% | 11.4% | 9.8% | 3.2% | 9.5% |
| ID3 | 66.2% | 8.9% | 8.8% | 3.4% | 12.8% |
| Naive Bayes Classifier | 56.8% | 14.8% | 16.2% | 4.2% | 8.2% |
| OneR | 94.8% | - | - | - | 5.2% |
| PART | 69.5% | 6.8% | 10.2% | 6.0% | 7.5% |
| kNN (k=5) | 82.6% | 2.8% | 2.5% | 7.7% | 5.2% |
| Distance Weighted kNN (k=5) | 76.3% | 4.6% | 5.7% | 5.8% | 7.5% |
| Distance & Attribute Weighted kNN (k=5) | 73.4% | 16.3% | 8.5% | 0.8% | 1.1% |

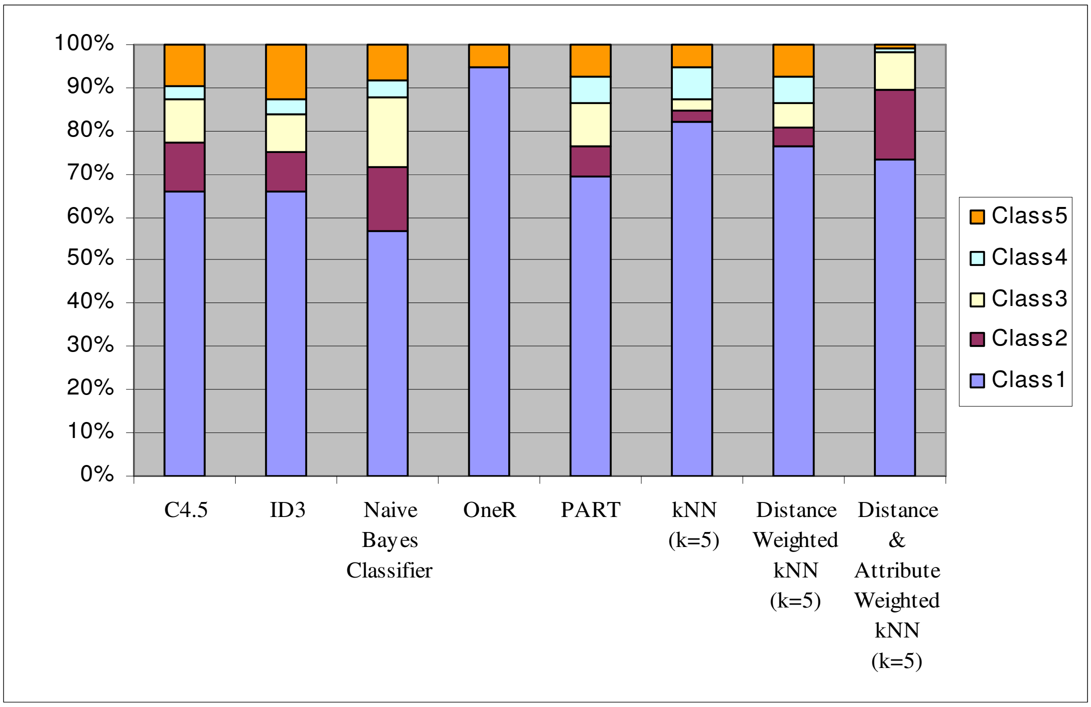


**Figure 5.3** The percentages of the classes generated by the algorithms

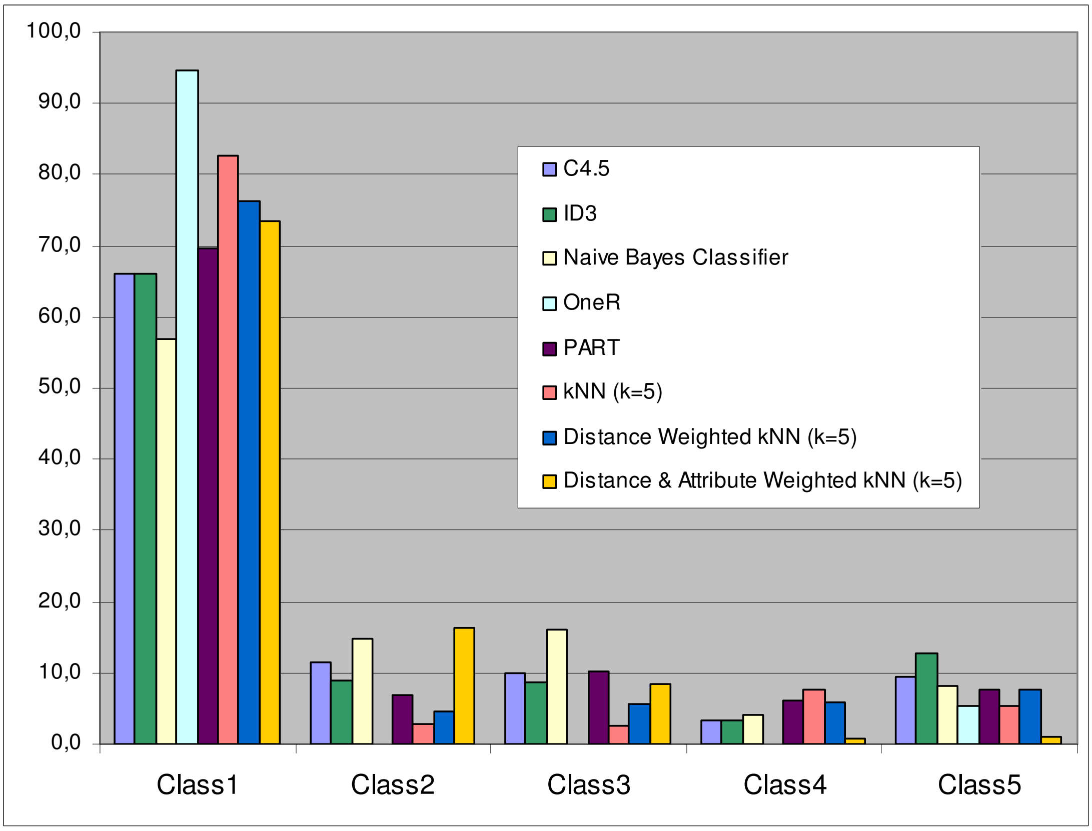


**Figure 5.4** The percentages of the classes in groups

## 5.6. Outputs of the Implemented kNN Versions

In the table below, all the records in the dataset are put into the classes by the versions of kNN algorithms. This table shows only the percentages of the classes. In this table, the reliable version of kNN is obviously the Attribute and Distance Weighted one.

**Table 5.4** Outputs of the kNN algorithms in percentages

| | **Class1** | **Class2** | **Class3** | **Class4** | **Class5** |
|---|---|---|---|---|---|
| **Attribute and Distance Weighted kNN** | 73.4% | 16.3% | 8.5% | 0.8% | 1.1% |
| **Distance Weighted kNN** | 76.3% | 4.6% | 5.7% | 5.8% | 7.5% |
| **kNN (k=1)** | 65.7% | 11.5% | 9.2% | 4.5% | 9.1% |
| **kNN (k=3)** | 69.5% | 3.1% | 4.2% | 5.7% | 3.7% |
| **kNN (k=5)** | 82.6% | 2.8% | 2.5% | 7.7% | 5.2% |
| **kNN (k=10)** | 86.2% | 1.1% | 1.8% | 7.7% | 3.2% |
| **kNN (k=15)** | 86.3% | 0.2% | 1.8% | 8.0% | 3.7% |

# CHAPTER 6

# CONCLUSION

As it is known, it is a reality that to digitalize the whole psychological status of a human being is really hard. Normally converting someone's emotional status, habits, attitudes and some other specifications to numeric values seems impossible. But in this thesis, we have tried to overcome this hardship this on an interesting and hard title, *detecting students at risk of substance abuse by using data mining classification algorithms*. At the end of this study, we have seen that some advanced scientific studies can be done on this subject.

In order to get more feasible and more efficient results from this kind of study there are some suggestions below about a scientific committee, the applied questionnaire, the used training set, the implemented algorithms and the results of the study.

Firstly, there ought to be a scientific committee strictly studying on this project. In the committee there might be some psychologists, sociologists, psychiatrists, academicians, educators, PDR teachers, computer engineers, IT programmers, and some relevant specialists. In the application of the *Distance and Attribute Weighted kNN* method, this committee can give different weights to each attribute in the training set. Only this kind of a committee can have the opportunity to find the feasible and reliable solutions.

In the questionnaire there should be at least 50 questions for a strict and serious application for the youth in schools. Moreover the questions should be selected by the committee. 20 or 25 questions in the questionnaire can be seen sufficient in the initial steps but in further levels, the number of questions aimed to decipher the whole specialties of a human being should be at least 50.

If the training set would be larger than the one used in this project, we could have better outputs. In our study the number of tuples in the training set is 110. There should be at least 2000 different entries in the predefined training set defined by the committee. Each of the entries should be different from the others in the set and has a different risk value.

The training set should contain a lot of different entries. All the probabilities of the attributes in that phase should have a value which is different from zero. If the conditional probability used in the Naïve Bayes Classifier is zero, classifying the new records gets difficult, sometimes impossible. For this reason, there should be a variety of tuples in the training set.

For the training dataset we have prepared 110 tuples in different risk ratios raging from 0% to 100%. In our applications, if a person as a tuple in the training set smokes cigarette his/her risk ratio is established as higher than 50%. Additionally if a person both smokes and has misbehaviors, his/her risk ratio established as higher than 70%. This is the way that we have chosen so as to create the predefined training dataset.

On the other hand, there is another approach in the style of defining the training dataset. It is to put only drug users and non drug users into the training set. Therefore the risk values for the tuples can be either 0% or 100%. In this style there is no need to give different risk ratios raging from 0 and 100 to the tuples in the training set. By applying the classifiers to some particular new records some different risk ratios can be calculated. These particular records as intervals, whose risk values are ranging from 0% to 100%, then can be added to the training set for the purpose of enriching the training set.

As it is seen in the experiments, *The Attribute Weighted kNN* and *The Distance and Attribute Weighted kNN* methods have given the best results of all. Beside these applied methods in this thesis, more advanced and effective classifier algorithms can be implemented to this study. Especially, in *The Distance and Attribute Weighted kNN* method, different weights to each question (attribute) in the training set can be given by the scientific committee. By applying the other classifiers techniques similar types of results can be taken. Therefore, an opportunity occurs in order to compare these algorithms for the purpose of figuring out the most efficient one. At the end weaknesses and strengths of the implemented algorithms can be compared at the same time.

This thesis deals with an interesting and beneficial study, drug dependency of the youth. It focuses on a common problem of the new generation in the world by using a new and original method. Besides, it tries to find a plausible and feasible solution as an *urgent precaution system* for those students who are highly at risk of drug dependence. As a conclusion, it can be said that this thesis is a blueprint of further and advanced steps.

# REFERENCES

Aggarwal C.A., Yu P.S., *Privacy-Preserving Data Mining, Models and Algorithms,* Springer, ISBN:9780387709918, p.166, p.325, 2008a.

Aggarwal C.A., Yu P.S., *Privacy-Preserving Data Mining, Models and Algorithms,* Springer, ISBN:9780387709918, pp. 344-346, 2008b.

Aydın C., *A Socio demographic Evaluation of Cases Applying to a Child and Adolescent Dependency Centre During a Period of Two Years Attending*, Ege University Faculty of Science**,** Journal of Dependence, Vol: 7, N.: 1 , pp. 31-37, 2006.

Bahety, A., *Extension and Evaluation of ID3 Decision Tree Algorithm*, Department of Computer Science, University of Maryland, College Park, pp.1-8. retrieved in 06/15/2010. http://cs.umd.edu

Buddhinath, G., Derry D., *A Simple Enhancement to One Rule Classification,* Department of Computer Science, University of Melbourne, Melbourne, Australia, retrieved in 05/08/2010. http://www.buddhinath.net/OtherLinks/Documents/Improved%20OneR%20Algorithm.pdf

Bukstein. O.G., *Adolescent substance abuse: Assessment, prevention and treatment.* New York: John Wiley & Sons, New York, 1995.

Can, M.Ş., *İlköğretim II. Kademe Öğrencilerinde Görülen "Madde Bağımlılığı" Alışkanlığı.* M.S. Thesis, Sosyal Bilimleri Enstitüsü, Sakarya Üniversitesi, Kocaeli., p.122, p.146, p.154, p.162, p.173, pp. 197-208, 2007.

Drugfree World Web Site, *The Title : Drug Types and Effects*, retrieved in 06/22/2010. www.drugfreeworld.org

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

Dunham, M. H., *Data Mining Introductory and Advanced Topics*, Prentice Hall Publishing, ISBN: 0130888923, p76, 2005b.

Dunham, M. H., *Data Mining Introductory and Advanced Topics*, Prentice Hall Publishing, ISBN: 0130888923, p.87, 2005c.

Eibe F., Ian H., *Witten: Generating Accurate Rule Sets Without Global Optimization,* In: Fifteenth International Conference on Machine Learning, pp. 144-151, 1998.

Erdem G., Eke C.Y., Ögel K., Tanver S., *Peer Characteristics and Substance Use Among High School Students,* Journal of Dependence, Vol: 7, N.: 3, pp. 111-116, 2006.

Gök, Dr.Ali, *Madde Bağımlılığın Nedenleri,* retrieved in 07/12/2010. http://www.tavsiyeediyorum.com

Gülkan, B., *Eroin Bağımlılarının Kişilik ve Sosyodemografik Özellikleri,* M.S. Thesis, İstanbul Üniversitesi Adli Tıp Enstitüsü Sosyal Bilimler Anabilim Dalı, İstanbul, 1994.

Gürol D.T., *Uyuşturucu/Uyarıcı Maddeler ve Çocuğunuz,Anne Babalar için Kitapçık*, ÇEMATEM Yayınları, Bakıköy Ruh ve Sinir Hastalıkları Hastanesi, İstanbul, p.14, pp. 16-32, 2007a.

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

Holte, R.C., *Very simple classification rules perform well on most commonly used datasets*, Machine Learning, pp. 11:63-91, 1993.

Hürriyet Gazetesi, Gündem Sayfası, retrieved in 03/24/2010. www.hurriyet.com.tr

Ian H. Witten and Eibe Frank (2005) "*Data Mining: Practical machine learning tools and techniques*", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Jiang, L., Zhang H., Cai Z., *Dynamic K-Nearest-Neighbor Naive Bayes with Attribute Weighted*, Springer-Verlag Berlin Heidelberg, pp. 365–368, 2006.

Kimmel. D.C. & Weiner. I.B., *Adolescence: A developmental transition.* 2nd ed. New York: John Wiley & Sons. Inc., New York, 1995.

Kırcan S., *The Relationship Between Peer Pressure, Internal versus External Locus of Control and Adolescent Substance Use*, M.S. Thesis, Boğaziçi University, İstanbul, pp. 14-17, 2006.

Köknel Ö., *Bağımlılık Alkol ve Madde Bağımlılığı*, ISBN: 9789754057959, pp. 20-319, İstanbul, 1998.

Listverse Web Site, *The Title : Drug Types and Effects*, retrieved in 06/22/2010. http://listverse.com/2007/09/27/top-10-drugs-and-their-effects/

Mitchell, T.M., *Machine Learning*, *Title: 8.2.1 Distance-Weighted kNN*, McGrow-Hill, ISBN: 0070428077, pp. 233-236, 1997a.

Mitchell, T.M., *Machine Learning*, McGrow-Hill, ISBN: 0070428077, pp. 156-159, 1997b.

Mitchell, T.M., *Machine Learning Lecture Slides for text book*, ISBN: 0070428077, McGrow-Hill, 1997c.

Moradian M., Barani A., *KNNBA: K-Nearest-Neighbor-Based Association Algorithm*, JATIT, Journal of Theoretical and Applied Information Technology, pp. 123-130, www.jatit.org, 2009.

Nigsch F., Bender A., Buuren B., Tissen J., Nigsch E., Mitchell J.B.O., "*Melting Point Prediction Employing k-nearest Neighbor Algorithms and Genetic Parameter Optimization*". Journal of Chemical Information and Modeling 46 (6): 2412-2422. 2006.

Ögel, K., *Ergen Uçucu Madde Bağımlılarının Özellikleri - Umatem Verileri*, Bakıköy Ruh ve Sinir Hastalıkları Hastanesi, İstanbul, p.11, 2004.

Peng, W., Chen J., Zhou H., *An Implementation of ID3 - Decision Tree Learning Algorithm*, University of New South Wales, Australia, pp. 1-6, 2010.

Roberts, P., *Drug Types and Effects*, retrieved in 06/22/2010. http://www.darvsmith.com/dox/lecturenotes.html,

Saatçioğlu, Ö., *Evaluation of Inpatient Cases with Alcohol and Drug Use Between Years of 1998 and 2002,* Journal of Dependence, Vol: 4, N.: 3, pp. 109-117, 2003.

Seyman, İ., *Uyuşturucu Sorununun Türkiye' deki Boyutları,* M.S. Thesis,, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü Disiplinlerarası Adli Tıp Ana Bilim Dalı Fizik İnceleme ve Kriminalistik Bilim Dalı, Ankara, 2000.

Tosun M., *Substance Use Patterns of Hospitalized Adolescent Inhalant Users: A Comparison According to Gender and Place of Residence,* Journal of Dependence, Vol: 6, N.: 2, pp.-76-83, 2005.

Tuncer, L., *Cumhuriyet Döneminden Bugüne Madde Bağımlılığı İle Mücadelede İç Güvenlik Ve Milli Ahlak Faktörlerinin Yeri Ve Önemi Üzerine Bir Deneme,* M.S. Thesis, Fırat Üniversitesi, Elazığ, pp. 73-75, 2007.

Tuncer, L., *Cumhuriyet Döneminden Bugüne Madde Bağımlılığı İle Mücadelede İç Güvenlik Ve Milli Ahlak Faktörlerinin Yeri Ve Önemi Üzerine Bir Deneme,* M.S. Thesis, Fırat Üniversitesi, Elazığ, pp. 70-71, 2007.

US Government, *Safe and Drug Free Schools, Specific Drugs and Their Effects,* Washington, 2010a, retrieved in 07/12/2010.
http://www.yic.gov/drugfree/drugeffects.html

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

Windle. M. & Windle. 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, 2003.

Wikipedia, *The free web encyclopedia*, *k-nearest neighbor algorithm,* retrieved in 07/12/2010, 2010a.
http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm

Wikipedia, *Breadth First Search,* retrieved in 08/16/2010, 2010b.
http://en.wikipedia.org/wiki/Breadth-first_search

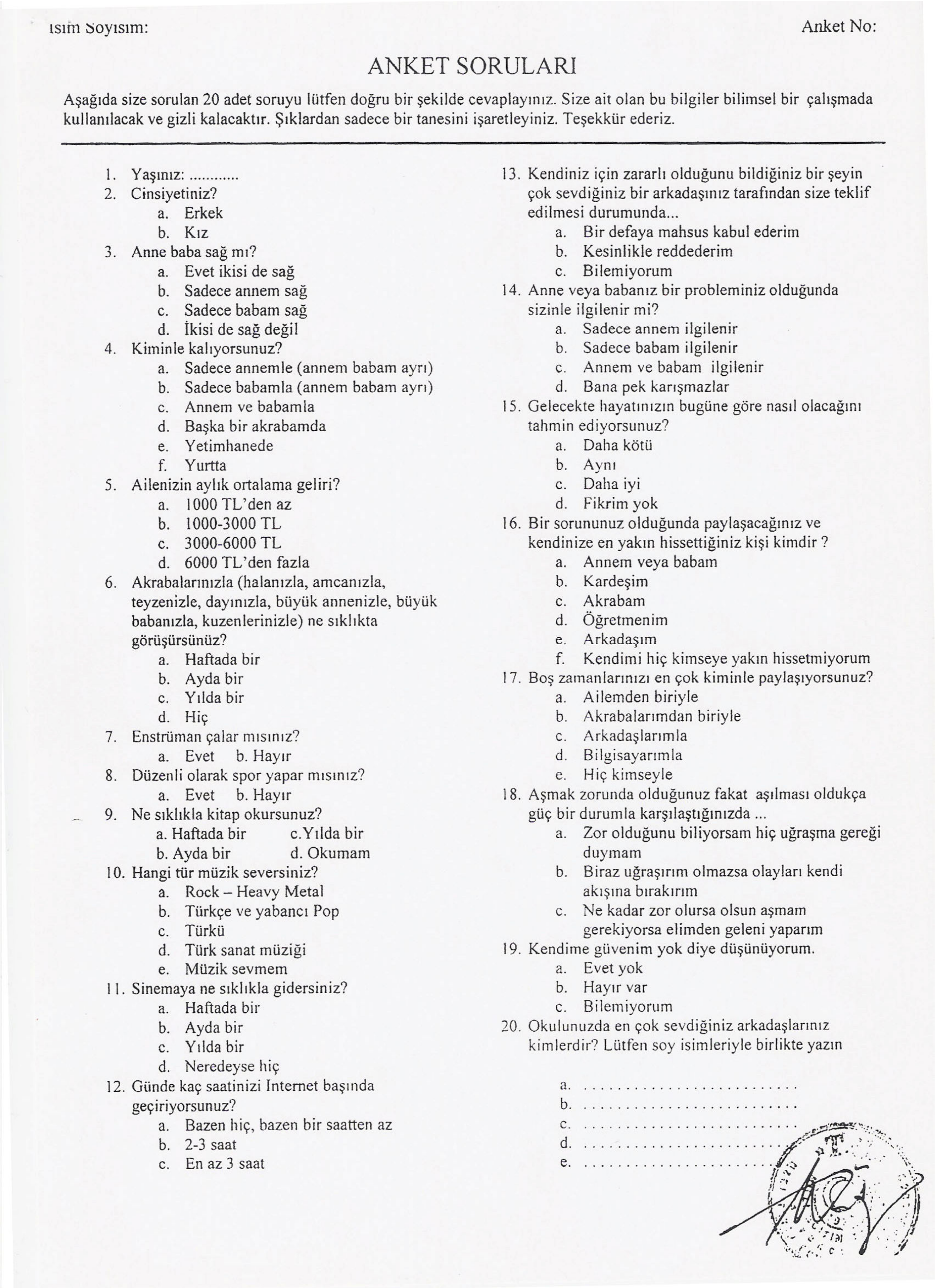
Zor, C., *Ortaöğretim Öğrenci Ailelerinin Uyuşturucu Madde Kullanımının Zararları İle Korunma Yolları Hakkındaki Görüşleri*, M.S. Thesis, Ankara Üniversitesi, Ankara, p.40, pp. 100-107, 2005.

# APPENDIX A – DATA PREPARATION PROCEDURE CODES

```
/*  TITLE :
 *  Detecting Students at Risk of Substance Abuse
 *  by Using Data Mining Classification Algorithms
 *  Author   : Faruk BULUT
 *  Language : C
 */

#include <stdio.h>
#include <math.h>        // Library for mathematical calculations
#include <stdlib.h>
#define MAX 10000        // The maximum array size for the tuples
#define AGE 21           // Limit of the human age in the survey
#define TRAINING_SET 110 // Training Set (Predefined Class) Member 
Limit

// Global Variable used in all functions
int DB0[MAX][30];
// The Initial (raw) DataBase (DB) taken from DB0.txt
int DB1[MAX][30];
// The DB1, filled the missing data with appropriate values. This 
array recorded in DB1.txt
int DB2[MAX][30];
// The DB2, removed outliers. This array recorded in DB2.txt
int DB3[MAX][30];
// The DB3, Preparing The Training Set for 100 tuples.
int DB4[MAX][30];
// The newest DB with some omitted attributes. This array recorded 
in DB4.txt
int DBx[2][30];
// Number of missing values in each attribute
int F[AGE][30]={0};
// The table that shows the frequencies of each attribute and each 
tupple(entry)
int Pmax[20];
// The most signed answers in the questionnaire
int N=0;
// The number of records in the Data Base, DB0[][]

FILE *oku,*yaz,*yaz2;
// The reading and the writing pointers for the text files


// Reading the Data from DB0.txt to the array DB0[MAX]
void readData()
{
     int i=0,j;
     oku=fopen("DB0.txt","r");
     // Reading all the records in DB0.txt
     while(!feof(oku))
     {
             i++;         // i:record number
             for (j=0; j<30; j++)
                 fscanf(oku,"%d",&DB0[i][j]);
             N++;         // N:The total record number
     }
     N--;
     fclose(oku);
}

/* Omitting some Attributes in the DB0[][] and converting it to
DB1[][] and DB1.txt. In this function, some attributes(columns) will
be deleted. If 33% of the attribute is missing, this attribute will
deleted by changing the values with -2 */

void omitSomeAttributes()
{
      int i,j;
      // Checking all the datas to compute how many missing value
there are
      for (i=1; i<=N; i++)
         for (j=1; j<30; j++)
             if (DB0[i][j]==-1)
                DBx[j][0]++;
      yaz=fopen("Missing Data Info.txt","w");
      fprintf(yaz,"1st line is the attribute number \n2nd line is
number of missing values of the attribute \n3rd line is the missing
persantage\n\n");

      for (j=1; j<=30; j++)
          fprintf(yaz,"%-4d",j);
      fprintf(yaz,"\n");

      for (j=0; j<30; j++)
          fprintf(yaz,"%-4d",DBx[j][0]);
      fprintf(yaz,"\n");

      // Finding the percentage of the missing value
      for (j=0; j<30; j++)
      {
          DBx[j][1]=(DBx[j][0]*100)/N;
          fprintf(yaz,"%c%-3d",37,DBx[j][1]);
      }
      fprintf(yaz,"\n\n");

      for (j=0; j<30; j++)
          if (DBx[j][1]>33)
          {
             fprintf(yaz,"In the %d. attribute, there are too many
missing values. It will be deleted.\n",j+1);
             for (i=1; i<N; i++)
                 DB0[i][j]=-2;
          }

      yaz2=fopen("DB1.txt","w");
      // The new data base will be written to DB1.txt and DB1[][]
      for (i=1; i<=N; i++)
      {
         for (j=0; j<30; j++)
         {
             DB1[i][j]=DB0[i][j];
             fprintf(yaz2,"%d ",DB1[i][j]);
         }
         fprintf(yaz2,"\n");
      }
      fclose(yaz);
      fclose(yaz2);
}

void findFrequencies()
{
     yaz=fopen("Frequency Table.txt","w");
     int i,j,max;
     int A[]={2,1,3,1,1,0,1,1,3,2,1,2,3,3,1,3,3}; // Test, omit

     // This loop finds the frequencies of each attribute in the
array DB1[][]
     for(i=1; i<=N-1; i++)
     {
         for (j=1; j<=19; j++)
         {
              if (DB4[i][j]<0)
                 continue;
              F[DB4[i][j]][j]++;
         }
         for (j=26; j<=28; j++)
         {
             if (DB4[i][j]<0)
                 continue;
             F[DB4[i][j]][j]++;
         }
     }
     // The frequencies are written to the "Frequency Table.txt"
file
     for (i=0; i<AGE; i++)
     {
         fprintf(yaz,"Value %3d: ",i);
         for (j=1; j<=19; j++)
             fprintf(yaz,"%-5d",F[i][j]);

         for (j=26; j<=28; j++)
             fprintf(yaz,"%-5d",F[i][j]);

         fprintf(yaz,"\n");
     }
     fclose(yaz);

     for (i=1; i<18; i++)
     {
         max=0;
         for(j=0; j<6; j++)
         {
             if (F[i][j]>F[i][j+1])
                max=j;
         }
         Pmax[i]=A[i-1];
     }
}

void findPercentages()
{
     int i,j;
     float Bayes;
     yaz=fopen("Frequency Percentage Table.txt","w");

     // The Probabilities are written to the "Bayseian Probability
Table.txt" file
     for (i=0; i<AGE; i++)
     {
         fprintf(yaz,"Value %3d: ",i);
         for (j=1; j<20; j++)
             fprintf(yaz,"%c%-5.1f",37,(float)100*F[i][j]/N);
             /*37 is for the % sign in ASCII table*/
         fprintf(yaz,"\n");
     }
     fclose(yaz);
}

// Repairing some missing data in DB1[][] and converting it to
DB2[][] and DB2.txt
void repairMissingData()
{
     int i,j,k,max=0;
     for(i=0; i<N; i++)
         for (j=1; j<20; j++)
              if (DB1[i][j]==-1)
                  DB1[i][j]=Pmax[j-1];

      yaz=fopen("DB2.txt","w");
      // The new data base will be written to DB2.txt
      for (i=1; i<=N; i++)
      {
         for (j=0; j<30; j++)
         {
             DB2[i][j]=DB1[i][j];
             fprintf(yaz,"%d ",DB2[i][j]);
         }
         fprintf(yaz,"\n");
      }
      fclose(yaz);
}
```

```
void makeLier(int x)
{
     int i,j;
     for (j=0; j<30; j++)
         DB2[x][j]=-2;

     // OutLier'ları sevenlerin listesinden Outlier'lar siliniyor
     // DB2[i][ 20-24 arasi ]
     // Gpaph yapisinda problem oluşmaması için
     for (i=1; i<=N; i++)
         for (j=20; j<=24; j++)
         {
             if(x==DB2[i][j])
                  DB2[i][j]=0;
         }
}

// Cleaning the outliers in the DB2[][] and converting it to DB4[][]
and DB4.txt
void cleanOutLiers()
{
    int i,j,flag=0,LierCounter=0;
    for (i=1; i<=N; i++)
    {
        // Some answers to the questions in the query could be
logically wrong
        // In this case, Liers can be detected easily
        if (DB2[i][1]>20 || DB2[i][1]<10)   // Age Liars
            makeLier(i);
        if (DB2[i][3]==4 && DB2[i][4]==3)   // Parents Liars
            makeLier(i);
        if (DB2[i][3]==2 && DB2[i][4]==2)   // Parents Liars
            makeLier(i);
        if (DB2[i][3]==3 && DB2[i][4]==1)   // Parents Liars
            makeLier(i);
        if (DB2[i][3]==4 && (DB2[i][4]==1 || DB2[i][4]==2 ||
DB2[i][4]==3) )  // Parents Liers
            makeLier(i);
        if (DB2[i][14]==1 && DB2[i][3]==3)  // Parents Liars
            makeLier(i);
        if (DB2[i][14]==2 && DB2[i][3]==2)  // Parents Liars
            makeLier(i);
        if (DB2[i][14]==3 && DB2[i][3]!=1)  // Parents Liars
            makeLier(i);
        if (DB2[i][7]==0 && DB2[i][10]==5)  // Music Liars
            makeLier(i);
        if (DB2[i][16]==1 && DB2[i][3]==4)  // Parents Liars
            makeLier(i);
        if (DB2[i][25]>=1 && DB2[i][25]<=5) // Grade Point
Avarage(GPA) Liars
            makeLier(i);
    }

    yaz=fopen("DB4.txt","w");
    yaz2=fopen("OutLiers ID List.txt","w");

    fprintf(yaz2,"These are Outliars' ID:\n\n");

    // The new data base will be written to DB2.txt
    for (i=1; i<=N; i++)
    {
         flag=0;
         for (j=0; j<30; j++)
         {
             if (DB2[i][0]==-2)
             {
                 fprintf(yaz2,"%d,",i);
                 flag=1;
                 LierCounter++;
                 break;
             }
             DB4[i][j]=DB2[i][j];
             fprintf(yaz,"%d ",DB2[i][j]);
         }
         //Silinen satırın yerine \n koymaması için
         if (flag==1)
            continue;
         fprintf(yaz,"\n");
    }
    fprintf(yaz2,"\n\nThere are %d Liers in the questionaries. They
will be removed from the data base.",LierCounter);

    fclose(yaz);
    fclose(yaz2);
}

int main(void)
{
    // Preparation of the data

    // Reading the Data from DB0.txt to the array DB0[][]
    readData();
    // Omiting some Attributes in the DB0[][] and converting it to
DB1.txt and DB1[][]
    omitSomeAttributes();
    // Repairing some missing data in DB1[][] and converting it to
DB2.txt and DB2[][]
    repairMissingData();
    // Cleaning some outliers in DB2[][] and converting it to
DB4.txt and DB4[][]
    cleanOutLiers();

    getchar();
    return 0;
}
```

# APPENDIX B - OFFICIAL PERMISSIONS

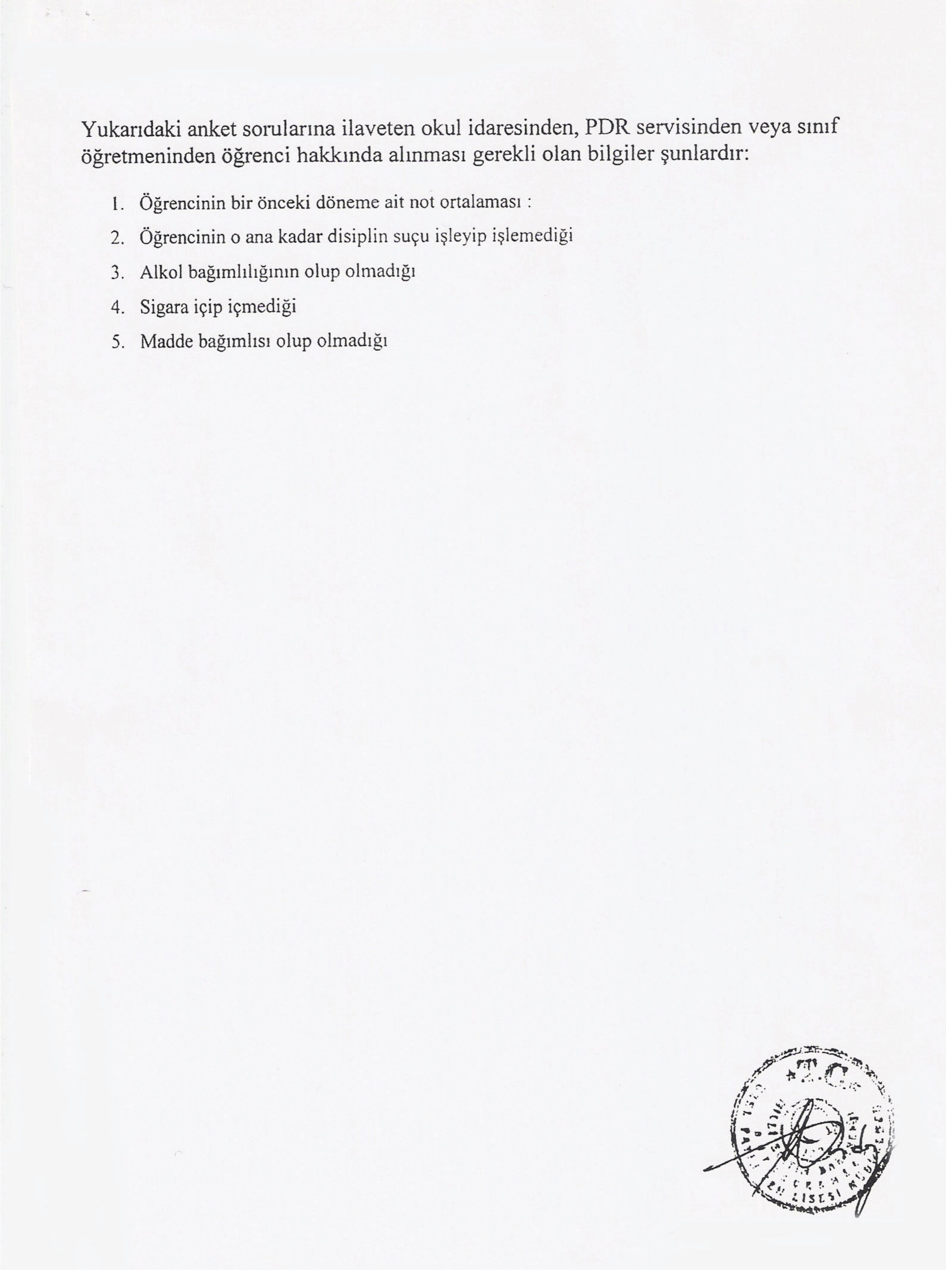


İsim Soyisim: Anket No:

## ANKET SORULARI

Aşağıda size sorulan 20 adet soruyu lütfen doğru bir şekilde cevaplayınız. Size ait olan bu bilgiler bilimsel bir çalışmada kullanılacak ve gizli kalacaktır. Şıklardan sadece bir tanesini işaretleyiniz. Teşekkür ederiz.

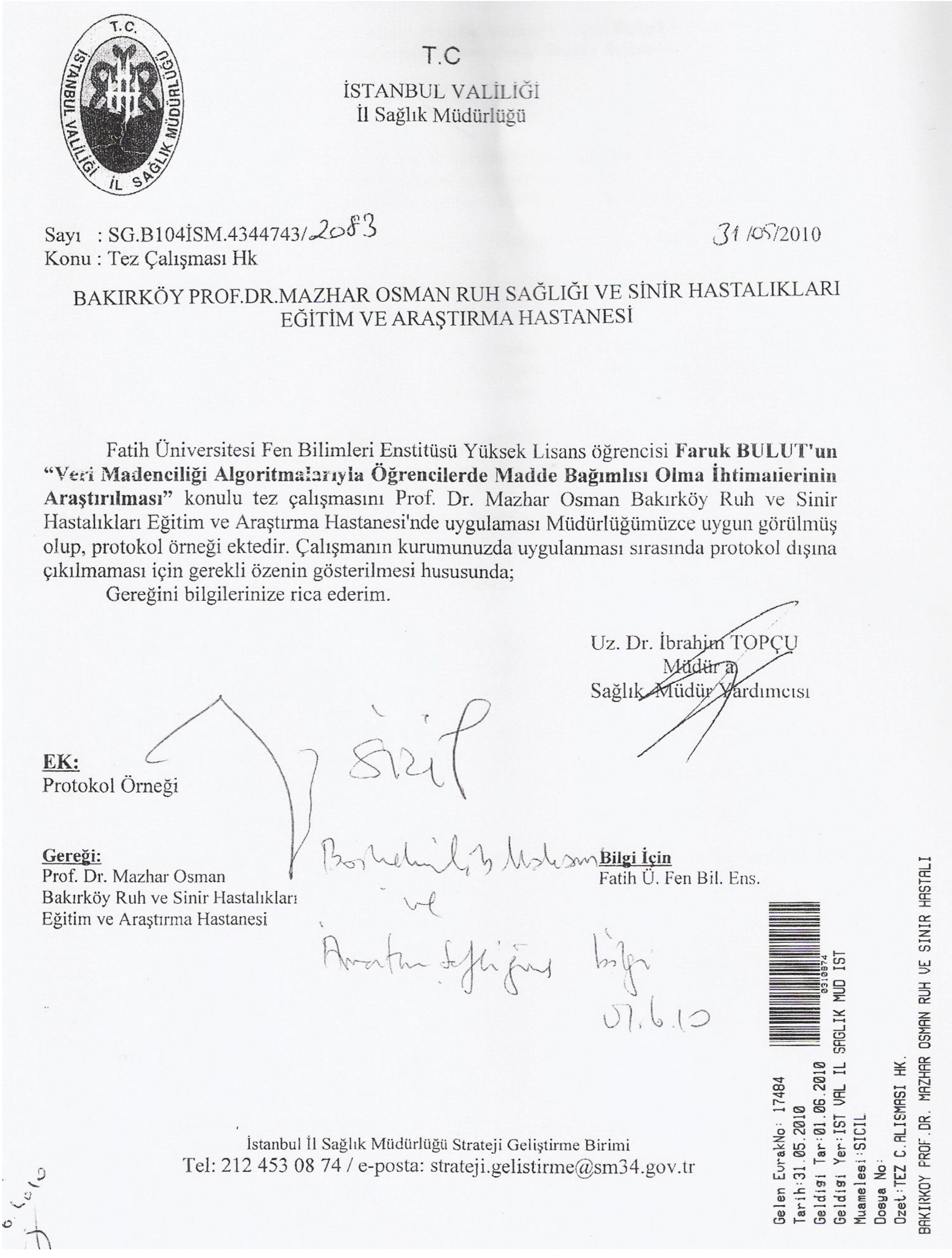
1. Yaşınız: ...........
2. Cinsiyetiniz?
   a. Erkek
   b. Kız
3. Anne baba sağ mı?
   a. Evet ikisi de sağ
   b. Sadece annem sağ
   c. Sadece babam sağ
   d. İkisi de sağ değil
4. Kiminle kalıyorsunuz?
   a. Sadece annemle (annem babam ayrı)
   b. Sadece babamla (annem babam ayrı)
   c. Annem ve babamla
   d. Başka bir akrabamda
   e. Yetimhanede
   f. Yurtta
5. Ailenizin aylık ortalama geliri?
   a. 1000 TL'den az
   b. 1000-3000 TL
   c. 3000-6000 TL
   d. 6000 TL'den fazla
6. Akrabalarınızla (halanızla, amcanızla, teyzenizle, dayınızla, büyük annenizle, büyük babanızla, kuzenlerinizle) ne sıklıkta görüşürsünüz?
   a. Haftada bir
   b. Ayda bir
   c. Yılda bir
   d. Hiç
7. Enstrüman çalar mısınız?
   a. Evet b. Hayır
8. Düzenli olarak spor yapar mısınız?
   a. Evet b. Hayır
9. Ne sıklıkla kitap okursunuz?
   a. Haftada bir c.Yılda bir
   b. Ayda bir d. Okumam
10. Hangi tür müzik seversiniz?
    a. Rock – Heavy Metal
    b. Türkçe ve yabancı Pop
    c. Türkü
    d. Türk sanat müziği
    e. Müzik sevmem
11. Sinemaya ne sıklıkla gidersiniz?
    a. Haftada bir
    b. Ayda bir
    c. Yılda bir
    d. Neredeyse hiç
12. Günde kaç saatinizi Internet başında geçiriyorsunuz?
    a. Bazen hiç, bazen bir saatten az
    b. 2-3 saat
    c. En az 3 saat
13. Kendiniz için zararlı olduğunu bildiğiniz bir şeyin çok sevdiğiniz bir arkadaşınız tarafından size teklif edilmesi durumunda...
    a. Bir defaya mahsus kabul ederim
    b. Kesinlikle reddederim
    c. Bilemiyorum
14. Anne veya babanız bir probleminiz olduğunda sizinle ilgilenir mi?
    a. Sadece annem ilgilenir
    b. Sadece babam ilgilenir
    c. Annem ve babam ilgilenir
    d. Bana pek karışmazlar
15. Gelecekte hayatınızın bugüne göre nasıl olacağını tahmin ediyorsunuz?
    a. Daha kötü
    b. Aynı
    c. Daha iyi
    d. Fikrim yok
16. Bir sorununuz olduğunda paylaşacağınız ve kendinize en yakın hissettiğiniz kişi kimdir ?
    a. Annem veya babam
    b. Kardeşim
    c. Akrabam
    d. Öğretmenim
    e. Arkadaşım
    f. Kendimi hiç kimseye yakın hissetmiyorum
17. Boş zamanlarınızı en çok kiminle paylaşıyorsunuz?
    a. Ailemden biriyle
    b. Akrabalarımdan biriyle
    c. Arkadaşlarımla
    d. Bilgisayarımla
    e. Hiç kimseyle
18. Aşmak zorunda olduğunuz fakat aşılması oldukça güç bir durumla karşılaştığınızda ...
    a. Zor olduğunu biliyorsam hiç uğraşma gereği duymam
    b. Biraz uğraşırım olmazsa olayları kendi akışına bırakırım
    c. Ne kadar zor olursa olsun aşmam gerekiyorsa elimden geleni yaparım
19. Kendime güvenim yok diye düşünüyorum.
    a. Evet yok
    b. Hayır var
    c. Bilemiyorum
20. Okulunuzda en çok sevdiğiniz arkadaşlarınız kimlerdir? Lütfen soy isimleriyle birlikte yazın
    a. ...........................
    b. ...........................
    c. ...........................
    d. ...........................
    e. ...........................



Yukarıdaki anket sorularına ilaveten okul idaresinden, PDR servisinden veya sınıf öğretmeninden öğrenci hakkında alınması gerekli olan bilgiler şunlardır:

1. Öğrencinin bir önceki döneme ait not ortalaması :
2. Öğrencinin o ana kadar disiplin suçu işleyip işlemediği
3. Alkol bağımlılığının olup olmadığı
4. Sigara içip içmediği
5. Madde bağımlısı olup olmadığı

T.C.
LİSESİ



T.C. İSTANBUL VALİLİĞİ İL SAĞLIK MÜDÜRLÜĞÜ

T.C

İSTANBUL VALİLİĞİ
İl Sağlık Müdürlüğü

Sayı : SG.B104İSM.4344743/2083 31/05/2010
Konu : Tez Çalışması Hk

BAKIRKÖY PROF.DR.MAZHAR OSMAN RUH SAĞLIĞI VE SİNİR HASTALIKLARI
EĞİTİM VE ARAŞTIRMA HASTANESİ

Fatih Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans öğrencisi **Faruk BULUT'un "Veri Madenciliği Algoritmalarıyla Öğrencilerde Madde Bağımlısı Olma İhtimallerinin Araştırılması"** konulu tez çalışmasını Prof. Dr. Mazhar Osman Bakırköy Ruh ve Sinir Hastalıkları Eğitim ve Araştırma Hastanesi'nde uygulaması Müdürlüğümüzce uygun görülmüş olup, protokol örneği ektedir. Çalışmanın kurumunuzda uygulanması sırasında protokol dışına çıkılmaması için gerekli özenin gösterilmesi hususunda;

Gereğini bilgilerinize rica ederim.

Uz. Dr. İbrahim TOPÇU
Müdür a.
Sağlık Müdür Yardımcısı

Sicil

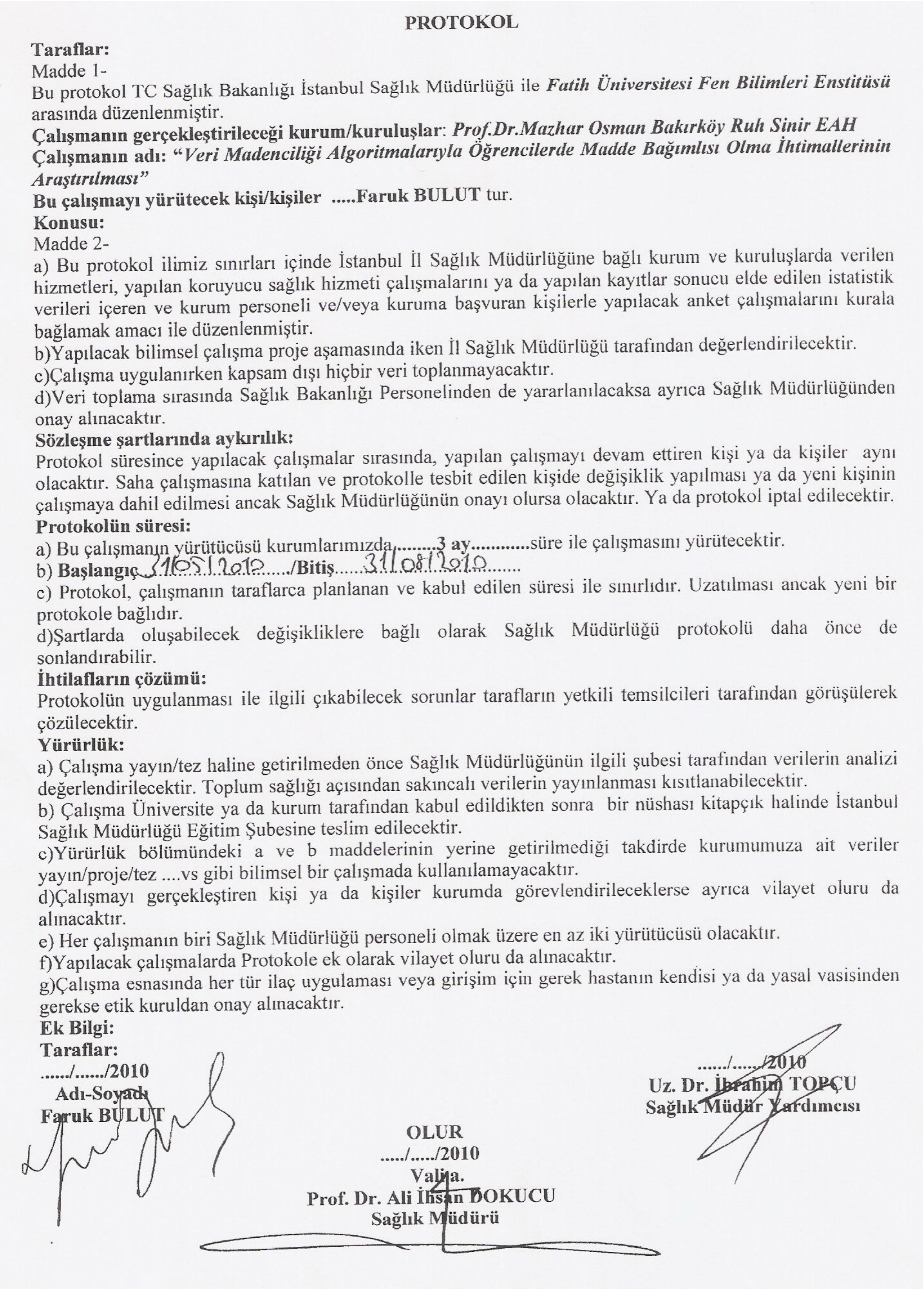
**EK:**
Protokol Örneği

Bağımlılık Merkezi ve Araştırma Şefliğine bilgi
07.6.10

**Gereği:**
Prof. Dr. Mazhar Osman
Bakırköy Ruh ve Sinir Hastalıkları
Eğitim ve Araştırma Hastanesi

**Bilgi İçin**
Fatih Ü. Fen Bil. Ens.

İstanbul İl Sağlık Müdürlüğü Strateji Geliştirme Birimi
Tel: 212 453 08 74 / e-posta: strateji.gelistirme@sm34.gov.tr



**PROTOKOL**

**Taraflar:**

Madde 1-

Bu protokol TC Sağlık Bakanlığı İstanbul Sağlık Müdürlüğü ile ***Fatih Üniversitesi Fen Bilimleri Enstitüsü*** arasında düzenlenmiştir.

**Çalışmanın gerçekleştirileceği kurum/kuruluşlar**: ***Prof.Dr.Mazhar Osman Bakırköy Ruh Sinir EAH***

**Çalışmanın adı:** ***"Veri Madenciliği Algoritmalarıyla Öğrencilerde Madde Bağımlısı Olma İhtimallerinin Araştırılması"***

**Bu çalışmayı yürütecek kişi/kişiler .....Faruk BULUT** tur.

**Konusu:**

Madde 2-

a) Bu protokol ilimiz sınırları içinde İstanbul İl Sağlık Müdürlüğüne bağlı kurum ve kuruluşlarda verilen hizmetleri, yapılan koruyucu sağlık hizmeti çalışmalarını ya da yapılan kayıtlar sonucu elde edilen istatistik verileri içeren ve kurum personeli ve/veya kuruma başvuran kişilerle yapılacak anket çalışmalarını kurala bağlamak amacı ile düzenlenmiştir.

b)Yapılacak bilimsel çalışma proje aşamasında iken İl Sağlık Müdürlüğü tarafından değerlendirilecektir.

c)Çalışma uygulanırken kapsam dışı hiçbir veri toplanmayacaktır.

d)Veri toplama sırasında Sağlık Bakanlığı Personelinden de yararlanılacaksa ayrıca Sağlık Müdürlüğünden onay alınacaktır.

**Sözleşme şartlarında aykırılık:**

Protokol süresince yapılacak çalışmalar sırasında, yapılan çalışmayı devam ettiren kişi ya da kişiler aynı olacaktır. Saha çalışmasına katılan ve protokolle tesbit edilen kişide değişiklik yapılması ya da yeni kişinin çalışmaya dahil edilmesi ancak Sağlık Müdürlüğünün onayı olursa olacaktır. Ya da protokol iptal edilecektir.

**Protokolün süresi:**

a) Bu çalışmanın yürütücüsü kurumlarımızda**........3 ay............**süre ile çalışmasını yürütecektir.

b) **Başlangıç**..31/05/2010.....**/Bitiş**......31/08/2010.......

c) Protokol, çalışmanın taraflarca planlanan ve kabul edilen süresi ile sınırlıdır. Uzatılması ancak yeni bir protokole bağlıdır.

d)Şartlarda oluşabilecek değişikliklere bağlı olarak Sağlık Müdürlüğü protokolü daha önce de sonlandırabilir.

**İhtilafların çözümü:**

Protokolün uygulanması ile ilgili çıkabilecek sorunlar tarafların yetkili temsilcileri tarafından görüşülerek çözülecektir.

**Yürürlük:**

a) Çalışma yayın/tez haline getirilmeden önce Sağlık Müdürlüğünün ilgili şubesi tarafından verilerin analizi değerlendirilecektir. Toplum sağlığı açısından sakıncalı verilerin yayınlanması kısıtlanabilecektir.

b) Çalışma Üniversite ya da kurum tarafından kabul edildikten sonra bir nüshası kitapçık halinde İstanbul Sağlık Müdürlüğü Eğitim Şubesine teslim edilecektir.

c)Yürürlük bölümündeki a ve b maddelerinin yerine getirilmediği takdirde kurumumuza ait veriler yayın/proje/tez ....vs gibi bilimsel bir çalışmada kullanılamayacaktır.

d)Çalışmayı gerçekleştiren kişi ya da kişiler kurumda görevlendirileceklerse ayrıca vilayet oluru da alınacaktır.

e) Her çalışmanın biri Sağlık Müdürlüğü personeli olmak üzere en az iki yürütücüsü olacaktır.

f)Yapılacak çalışmalarda Protokole ek olarak vilayet oluru da alınacaktır.

g)Çalışma esnasında her tür ilaç uygulaması veya girişim için gerek hastanın kendisi ya da yasal vasisinden gerekse etik kuruldan onay alınacaktır.

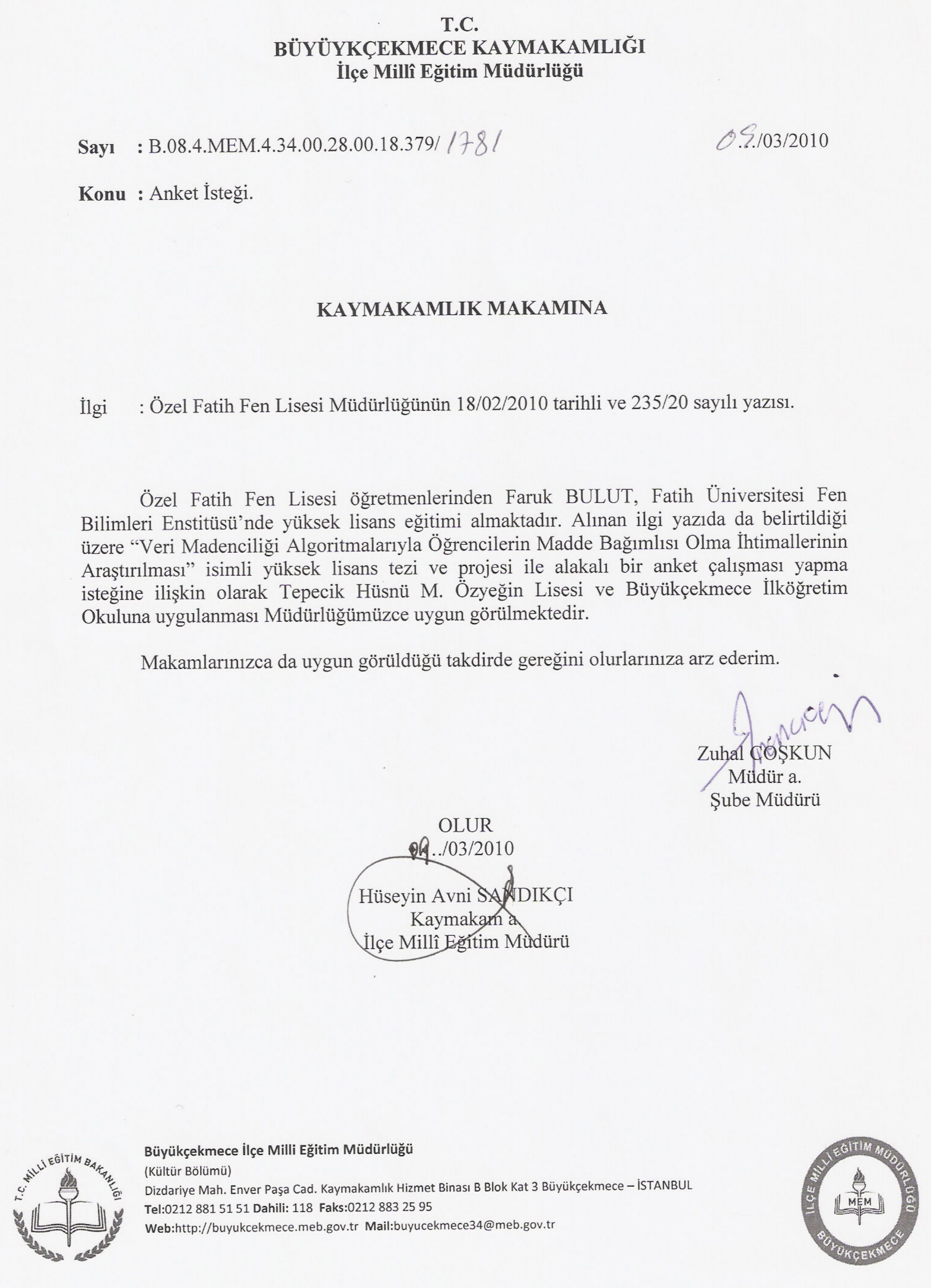
**Ek Bilgi:**

**Taraflar:**

**....../....../2010**
**Adı-Soyadı**
**Faruk BULUT**

**....../....../2010**
**Uz. Dr. İbrahim TOPÇU**
**Sağlık Müdür Yardımcısı**

**OLUR**
**...../...../2010**
**Vali a.**
**Prof. Dr. Ali İhsan DOKUCU**
**Sağlık Müdürü**



T.C.
BÜYÜYKÇEKMECE KAYMAKAMLIĞI
İlçe Millî Eğitim Müdürlüğü

**Sayı** : B.08.4.MEM.4.34.00.28.00.18.379/ 1781 09../03/2010

**Konu** : Anket İsteği.

**KAYMAKAMLIK MAKAMINA**

İlgi : Özel Fatih Fen Lisesi Müdürlüğünün 18/02/2010 tarihli ve 235/20 sayılı yazısı.

Özel Fatih Fen Lisesi öğretmenlerinden Faruk BULUT, Fatih Üniversitesi Fen Bilimleri Enstitüsü'nde yüksek lisans eğitimi almaktadır. Alınan ilgi yazıda da belirtildiği üzere "Veri Madenciliği Algoritmalarıyla Öğrencilerin Madde Bağımlısı Olma İhtimallerinin Araştırılması" isimli yüksek lisans tezi ve projesi ile alakalı bir anket çalışması yapma isteğine ilişkin olarak Tepecik Hüsnü M. Özyeğin Lisesi ve Büyükçekmece İlköğretim Okuluna uygulanması Müdürlüğümüzce uygun görülmektedir.

Makamlarınızca da uygun görüldüğü takdirde gereğini olurlarınıza arz ederim.

Zuhal COŞKUN
Müdür a.
Şube Müdürü

OLUR
09../03/2010

Hüseyin Avni SANDIKÇI
Kaymakam a.
İlçe Millî Eğitim Müdürü

Büyükçekmece İlçe Milli Eğitim Müdürlüğü
(Kültür Bölümü)
Dizdariye Mah. Enver Paşa Cad. Kaymakamlık Hizmet Binası B Blok Kat 3 Büyükçekmece – İSTANBUL
Tel:0212 881 51 51 Dahili: 118 Faks:0212 883 25 95
Web:http://buyukcekmece.meb.gov.tr Mail:buyucekmece34@meb.gov.tr