

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/321323918>

# Identifying Risk Factors for Drug Use in an Iranian Treatment Sample: A Prediction Approach Using Decision Trees

Article in Substance Use & Misuse · November 2017

DOI: 10.1080/10826084.2017.1392981

CITATIONS

9

READS

116

6 authors, including:



**Alireza Amirabadizadeh**

24 PUBLICATIONS 57 CITATIONS

[SEE PROFILE](#)



**Hossein Nezami**

Mashhad University of Medical Sciences

3 PUBLICATIONS 23 CITATIONS

[SEE PROFILE](#)



**Michael G Vaughn**

Saint Louis University

533 PUBLICATIONS 9,330 CITATIONS

[SEE PROFILE](#)



**Samaneh Nakhaee**

Birjand University of Medical Sciences

43 PUBLICATIONS 86 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Lead poisoning among opium users in Iran: an emerging health hazard [View project](#)



The role of serotonin receptor on long-term potentiation [View project](#)



SUBSTANCE USE & MISUSE  
An International Interdisciplinary Forum

## Substance Use & Misuse

ISSN: 1082-6084 (Print) 1532-2491 (Online) Journal homepage: <http://www.tandfonline.com/loi/isum20>

# Identifying Risk Factors for Drug Use in an Iranian Treatment Sample: A Prediction Approach Using Decision Trees

Alireza Amirabadizadeh, Hossein Nezami, Michael G. Vaughn, Samaneh Nakhaee & Omid Mehrpour

To cite this article: Alireza Amirabadizadeh, Hossein Nezami, Michael G. Vaughn, Samaneh Nakhaee & Omid Mehrpour (2017): Identifying Risk Factors for Drug Use in an Iranian Treatment Sample: A Prediction Approach Using Decision Trees, Substance Use & Misuse, DOI: [10.1080/10826084.2017.1392981](https://doi.org/10.1080/10826084.2017.1392981)

To link to this article: <https://doi.org/10.1080/10826084.2017.1392981>



Published online: 27 Nov 2017.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

ORIGINAL ARTICLE



# Identifying Risk Factors for Drug Use in an Iranian Treatment Sample: A Prediction Approach Using Decision Trees

Alireza Amirabadizadeh<sup>a</sup>, Hossein Nezami<sup>b</sup>, Michael G. Vaughn<sup>c</sup>, Samaneh Nakhaee<sup>a,d</sup>, and Omid Mehrpour<sup>a</sup>

<sup>a</sup>Medical Toxicology and Drug Abuse Research Center (MTDRC), Birjand University of Medical Sciences, Moallem Avenue, Birjand, Iran;

<sup>b</sup>Department of Basic Sciences, Faculty of Medicine, Gonabad University of Medical Sciences, Gonabad, Iran; <sup>c</sup>School of Social Work, Saint Louis University, St. Louis, Missouri, USA; <sup>d</sup>Cardiovascular Diseases Research Center, Birjand University of Medical Sciences, Birjand, Iran

## ABSTRACT

**Introduction and aim:** Substance abuse exacts considerable social and health care burdens throughout the world. The aim of this study was to create a prediction model to better identify risk factors for drug use. **Design and Methods:** A prospective cross-sectional study was conducted in South Khorasan Province, Iran. Of the total of 678 eligible subjects, 70% (n: 474) were randomly selected to provide a training set for constructing decision tree and multiple logistic regression (MLR) models. The remaining 30% (n: 204) were employed in a holdout sample to test the performance of the decision tree and MLR models. Predictive performance of different models was analyzed by the receiver operating characteristic (ROC) curve using the testing set. Independent variables were selected from demographic characteristics and history of drug use. **Results:** For the decision tree model, the sensitivity and specificity for identifying people at risk for drug abuse were 66% and 75%, respectively, while the MLR model was somewhat less effective at 60% and 73%. Key independent variables in the analyses included first substance experience, age at first drug use, age, place of residence, history of cigarette use, and occupational and marital status. **Discussion and Conclusion:** While study findings are exploratory and lack generalizability they do suggest that the decision tree model holds promise as an effective classification approach for identifying risk factors for drug use. Convergent with prior research in Western contexts is that age of drug use initiation was a critical factor predicting a substance use disorder.

## KEYWORDS

Decision tree; substance use; addiction; drug abuse

## 1. Introduction



Substance abuse is one of the most pressing concerns of the present century (Rush, & Wild, 2003) that is closely related to cultural, psychological, economic, religious, social, and historical aspects of a Community (Alavi, Mehrdad, & Makarem, 2016; Jalilian et al., 2016).

Many countries such as Iran are faced with a high rate of substance use (Mehrpour, Karrari, & Sheikhazadi, 2013). Nonetheless, the prevalence of drug use has not yet been directly studied in Iran (Momtazi, & Rawson, 2010).

Official statistics and indirect projections have reported that approximately 700,000–4,000,000 people in Iran are substance-dependent people (Alavi, Mehrdad, & Makarem, 2016; Nassr et al., 2006). However, these estimations are not sufficiently reliable. In other words, it is very difficult to accurately determine the prevalence of substance and alcohol abuse and dependence (because of legal and religious limitations and social stigmatization) in Iran (Alavi, Mehrdad, & Makarem, 2016). In this regard, in 2011 it was reported that 2.3% of the Iranian population aged 15–64 years each year were using opioids

(Rahimi-Movaghar et al., 2015, United Nations Office on Drugs and Crime 2011). In a household survey including 3840 people aged 15 and older, 17.9% of the respondents reported using opium at least once during lifetime (Iran Drug Control Headquarters (IDCH) Rapid Situational Assessment 2007).

Potential causes of the high incidence of recreational drug abuse include neighboring Afghanistan, contributing to 90% of the global opium production, being on the main transit route for drugs (Mehrpour et al., 2016) as well as conventionally using opium for medicinal and recreational purposes in Iran (Metcalf, Olufajo, & Salim, 2015). Because of certain conditions and proximity to major drug producing centers, Iran has the highest per capita opioid use (Rahimi-Movaghar et al., 2015, Mehrpour et al., 2016, Day, Nassirimanesh, Shakeshaft, & Dolan, 2006). Overall Opioids include natural (e.g. Opium, morphine); semisynthetic (e.g. Heroin) and synthetic forms (e.g. Methadone and fentanyl) (Metcalf, Olufajo, & Salim, 2015, Williams, & Padmanabhan, 2009). Among them opium is the most widely used substance in

**CONTACT** Omid Mehrpour  [omid.mehrpour@yahoo.com.au](mailto:omid.mehrpour@yahoo.com.au)  Medical Toxicology and Drug Abuse Research Center (MTDRC), Birjand University of Medical Sciences (BUMS), Moallem Avenue, Birjand, Iran.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/ismu](http://www.tandfonline.com/ismu).

© 2017 Taylor & Francis Group, LLC

Iran (Mehrpour, Karrari, & Sheikhezadi, 2013, Karrari, Mehrpour, Afshari, & Keyler, 2013, Nikfarjam et al., 2016, Jafari et al., 2010). Also Iran is a traditional opium user in the south-west of Asia. The first reports of opium use date back to the 17th century (Alam-mehrjerdi, Abdollahi, Higgs, & Dolan, 2015). The United Nations Office on Drugs and Crime (UNODC) has estimated that each year 40 tons of opium are consumed in Iran, and statistics show that out of 70 million Iranians, four million are substance abusers (Metcalf, Olufajo, & Salim, 2015). However, in the recent decades in Iran, significant changes in the pattern of substance use have occurred (Rahimi-Movaghar et al., 2015). In the past, opium was used mostly as an analgesic (Karrari, Mehrpour, Afshari, & Keyler, 2013); however, today's youths are more interested in other substances such as heroin, hashish, cannabis and tramadol (Rahimi-Movaghar et al., 2015, Jafari et al., 2010, Goodarzi, Mehrpour, & Eizadi-Mood, 2011, Mehrpour, 2013, Mehrpour, Karrari, & Afshari, 2012, Karrari, Mehrpour, & Balali-Mood, 2012). Abuses of the new drugs in Iran are now a major health problem, particularly for the youth and young adults (Karrari, Mehrpour, & Balali-Mood, 2012). Since 2000, heroin has emerged as a new health concern among some drug users. Low opium availability and low price of heroin facilitated heroin abuse in the country for the first time (Rahimi-Movaghar et al., 2015). Heroin is an easily accessible with high addiction potential that has encompassed the population. After an average 5–7 time regular use of it leads to addiction and it is very difficult to abstain from this drug (Singh et al., 2016). After a short period of time, because of low heroin supply, heroin crack production was initiated in Iran. The reports of the Persian Drug Control Headquarters show that, smoking a new synthetic form of heroin which is colloquially named “crack” became common among opium and heroin users (Alam-mehrjerdi, Abdollahi, Higgs, & Dolan, 2015). And also, nowadays, Iranian crystal is one of the most common abused drugs in Iran (Karrari, Mehrpour, Afshari, & Keyler, 2013). It is one of the newest drugs between the Iranian addicts that are spreading widely among youngsters (Jafari et al., 2010, Karrari, Mehrpour, & Balali-Mood, 2012, Alam-mehrjerdi, 2013). As we know Iranian Crack and Iranian crystal both are heroin-based narcotic which are basically different from ones used in Western countries (Karrari, Mehrpour, & Balali-Mood, 2012, Farhoudian et al., 2014, Sarraimi, Ghorbami, & Taghavi, 2013).

In a previous study in the Khorasan Province, it was found that the types of abused drugs are related to place of residence, age, marital status, occupation, and education (Karrari, Mehrpour, Afshari, & Keyler, 2013). Besides that, socio-demographic factors such as age, gender and occupation contribute greatly to opium

consumption (Chaturvedi, Mahanta, Bajpai, & Pandey, 2013).

Jafari et al., (2009) reported that opiate use is more prevalent in men, individuals under 30 years of age, the married, the employed, and populations with lower education levels in Iran (Jafari et al., 2009). An epidemiologic study of opium use, a part of Pars Cohort Study, on 9000 adults showed young age, male gender, being non-married, and having history of cigarette smoking and alcohol consumption were associated with using opium (Fallahzadeh et al., 2017).

Prior research investigations have relied upon the normal regression model to identify significant factors associated with narcotics. Although regression is useful with regard to interactions, it does not model non-linear complex relationships composed of a high degree of interactions (Meng et al., 2013). One method for doing so is the use of a decision tree, which is a powerful tool for classification. Using a simple technique, this method offers a comprehensible pattern for available observations. The model introduced by the structure is easy and intelligible (Ramezankhani et al., 2014). Factors contributed in type of drug used (Traditional opium or newer drugs; heroin and its related drugs) is remain unclear still. In this study, we attempt to identify people at risk of drug abuse by comparing the decision tree and logistic regression. The aim of the present study is to diagnose individuals exposed to traditional opium and heroin-based substance abuse based on a decision tree model.

## 2. Methods

Having obtained the approval of the Ethics Committee of Birjand University of Medical Sciences (Number 552), we conducted this cross-sectional study from March 21, 2009 to March 21, 2010 in the South Khorasan Province in eastern Iran. There are seven drug rehabilitation centers in Birjand, of which three are run by the government and the others are private.

Stratified sampling performed by using lists of addiction centers, the number of addicts identified in each of the centers and a portion was determined according to the population of our overall sample. Then, selected samples were chosen randomly based on record numbers of people.

The data were collected through checklists and interviews with patients by medical students and the centers psychologists. The checklist included demographic characteristics (e.g., age, gender, marital status, occupation, and education level), as well as the type of drug consumed, and the date on which drug use started. Drugs were divided into two groups: natural opiates (opium, Shire or both), Heroin based drug (Iranian crystal, Iranian

crack, pure heroin) which all of them are heroin based drugs (semi-synthetic).

After data cleaning and preparation, the final data included 678 records and 14 variables. Thirteen independent variables and one dependent (or target) variable were specified. Input variables included age, age of first substance use, age of first cigarette use, age of first alcohol use, the number of drug abstinence, gender (2 items), place of residence (2 items), marital status (3 items), education (3 items), occupation (6 items), history of cigarette use (2 items), history of alcohol use (2 items), and the first substance experience (3 items). Based on the drug type variable and the use of expert medical opinion, the dependent variable is divided into two groups: traditional opium and heroin based drugs.

It is current in method data mining to divide the data set into two parts. A total of 678 people fulfilled the inclusion criteria, of whom 70% [n: 474 (244 people with consumption heroin based substance and 230 ones consuming natural opiates)] were randomly enrolled for developing the training dataset. The rest of the participants [n: 204 cases (105 people consuming heroin based substance and 99 ones consuming natural opiates)] were considered the testing dataset to investigate the performance of the decision tree.

### Decision tree

The decision tree is of the most popular classification algorithms in medicine and several other fields (Bellazzi, & Zupan, 2008, Delen, Walker, & Kadam, 2005). The purpose of a decision tree is to create a model to independent the value of the dependent variable based on a number of independent variables. CART is the most well-known decision tree algorithms (Ceri et al., 2003, Li et al., 2012). A decision tree consists of three parts: root node, internal node, and target node. The decision tree supports both continuous and categorical variables. In categorical variables, each level is considered as a sub-branch, whereas in the case of continuous variables, we select a point (e.g.,  $a$ ) in order to refer to a cut-off point. For the variable  $x$ , we regard members of ( $x \leq a$ ) as one group and those of ( $x > a$ ) as another group (Han, Pei, & Kamber, 2011).

All decision tree algorithms require a decomposition criterion in order to decompose a tree into a node. Criteria analysis, an exploratory process to choose the best criterion, is that the best form of action is divided into D data sets to separate classes. The decomposition criterion is an exploratory process for choosing an appropriate tool that best divides D data sets into separate classes. Information Gain, Gain Ratio, and Gini Index are three decomposition criteria. The decomposition criterion of a rank is determined for each independent variable, and

the variable that has the highest value is selected as the division criterion (Ceri et al., 2003, Loh, 2011).

### Developing decision tree

In a decision tree, the entire data set randomly is divided into two sections in different ways. The commonly used method is to consider two-third (70%) for model development and the remaining one-third (30%) for assessing model (Han, Pei, & Kamber, 2011, Rokach, & Maimon, 2014, Tayefi et al., 2017, Ramezankhani et al., 2016). CART decision tree algorithms are used to create the model. This CART algorithm takes advantage of the Gini index to choose the best variable as the decomposition criterion. This criterion for D data set is calculated as follows:

$$Gini(D) = \sum_{j=1}^n p_j^2$$

In this formula:  $n$  is the number of available classes in data,  $p_j$  stands for the probability of data belonging to  $j$  class. In the decision tree, the first variable (root node) is the most important factor. Each path in the decision tree can be converted into a decision rule. In this method, a set of logical conditions (IF-THEN) are used as a tree algorithm to make classifications or predictions (Ceri et al., 2003).

### Logistic regression

Assuming that errors are uncorrelated and normally distributed with zero mean and constant variance, in regression analysis, one examines the relationship between dependent and independent variables. In the case of problems having a binary dependent variable with two values of 0 and 1, it is possible to use generalized linear models. We will have a logistic model when the dependent variable has a Bernoulli distribution with probability success  $p$  and a link function. Logistic regression modelling is a standard method for qualitative variables. In logistic regression analysis, we use the concept of the odds ratio  $\frac{p}{1-p}$ , i.e. occurrence/non-occurrence ratio. When we have several independent variables and just one dependent variable, the logistic regression model is defined as (Agresti, 1996):

$$\text{logit}(p) = \ln \left( \frac{p}{1-p} \right) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} \cdots + \beta_k x_{ki} \quad 1, \dots, n$$

In the Hosmer-Lemeshow test, a Pearson test is used to compare the observed and fitted counts for this section.

This statistic does not incorporate precisely a limiting chi-square distribution. Cox and Snell's R-squared is based on the log likelihood for the model in comparison to the log likelihood for a baseline model. However, it offers a theoretical maximum value of  $< 1$  for categorical outcomes, even in case of a "perfect" model (Agresti, 1996).

The analysis was conducted in three stages. In the first phase, a comparison was made between the two groups, using an independent t-test for continuous variables and a chi-square test for categorical variables. In the second stage, using a decision tree and logistic regression, we dealt with the task of identifying people at risk of substance abuse. Eventually, we obtained sensitivity, specificity, accuracy, and area under the ROC curve for the two models, using a set of testing data. Statistical analysis was conducted via R3.3.1 software and pROC as well as rpart packages. In our study, all the variables whose significant levels in Mann-Whitney U-test and independent t-test were  $< 0.1$  were concurrently included in multiple logistic regression. The hypotheses for logistic regression model were taken into account. Each

observation was independent, there was at least 10 outcomes for every independent variable and the collinearity of the independent variables was investigated by analyzing the collinearity of age, age of smoking onset, age of alcohol consumption onset, and other quantitative variables, the value of the variance inflation factor (VIF) was under 2, the tolerance index was over 0.7, and the correlation coefficient between all of these variables was under 0.5, suggesting that there is no collinearity between these variables.

The level of statistical significance was considered  $< 0.05$ .

### 3. Results

According to Table 1, among the total of 678 participants, 329 people (48.5%) had used different types of natural opiates, and the remaining ones had specifically abused heroin. Out of this population, 611 (90.1%) were male. The average age of natural opiate abusers was  $36.26 \pm 10.66$  years and for heroin based drug users was

**Table 1.** Baseline demographic categorical characteristics of study subjects.

	Total	Natural opiate (n = 329)	Heroin based drugs (n = 349)	p-value
Sex				
Male <sup>†</sup>	611(90.1%)	296(90.0%)	315(90.3%)	p = 0.90
Female	67(9.9%)	33(10.0%)	34(9.7%)	
Residence				
Urban <sup>†</sup>	618(91.2%)	290(88.1%)	328(94%)	p = 0.007
Rural	60(8.8%)	39(11.9%)	21(6.0%)	
Marital status				
Married <sup>†</sup>	472(69.6%)	264(80.2%)	208(59.6%)	p < 0.001
Single	168(24.8%)	51(15.5%)	117(33.5%)	
Widow/ Divorced	38(5.6%)	14(4.3%)	24(6.9%)	
Occupation				
Student	163(24.0%)	51(15.5%)	112(32.1%)	p < 0.001
Housewife	66(9.7%)	32(9.7%)	34(9.7%)	
Self-employed <sup>†</sup>	357(52.7%)	181(55.0%)	176(50.4%)	
Employed	33(4.9%)	23(7.0%)	10(2.9%)	
Retired	18(2.7%)	17(5.2%)	1(0.3%)	
Driver	41(6.0%)	25(7.6%)	16(4.6%)	
Education status				
Primary school	202(29.8%)	108(32.8%)	94(26.9%)	p = 0.03
Secondary school <sup>†</sup>	227(33.5%)	92(28.0%)	135(38.7%)	
High school diploma	183(27.0%)	94(28.6%)	89(25.5%)	
University	66(9.7%)	35(10.6%)	31(8.9%)	
History of Cigarette use				
Yes	493(72.7%)	221(67.2%)	272(77.9%)	p = 0.002
History of alcohol use				
Yes	95(14.0%)	36(10.9%)	59(16.9%)	p = 0.02
First substance experience				
Hashish	22(3.2%)	3(0.9%)	19(5.4%)	p < 0.001
Natural opiate <sup>†</sup>	600(88.5%)	320(97.3%)	280(80.2%)	
Heroin based drugs	56(8.3%)	6(1.8%)	50(14.3%)	
Age <sup>*</sup>	33.85 $\pm$ 10.09	36.26 $\pm$ 10.66	31.57 $\pm$ 8.95	p < 0.001
Age of first substance <sup>*</sup>	21.86 $\pm$ 7.11	23.47 $\pm$ 7.51	20.36 $\pm$ 6.38	p < 0.001
Number of abstinence <sup>*</sup>	1.90 $\pm$ 2.38	1.68 $\pm$ 2.08	2.11 $\pm$ 2.62	p = 0.02
Smoking age of onset <sup>*</sup>	20.32 $\pm$ 5.72	21.77 $\pm$ 6.17	19.17 $\pm$ 5.05	p < 0.001
Alcohol age of onset <sup>*</sup>	19.33 $\pm$ 3.91 $\pm$	19.51 $\pm$ 4.82	19.23 $\pm$ 3.36	p = 0.77

\*Mean  $\pm$  SD.

<sup>†</sup>group with the highest frequency.



31.57  $\pm$  8.95 years. Descriptively, 33.5% had secondary school education level, 52.7% were self-employed, 72.7% had a history of smoking, and 88.5% had first experience of substance abuse with natural opiates.

The average number of drug abusers was higher among the subjects who used heroin compared to natural opiate abusers (2.11  $\pm$  2.62 vs. 1.68  $\pm$  2.08) (Table 1).

Moreover, the chi-square test results showed a significant relationship between the type of consumables and place of residence ( $p = 0.007$ ), marital status ( $p < 0.001$ ), occupational status ( $p < 0.001$ ), education level ( $p = 0.03$ ), previous smoking ( $p = 0.002$ ) and alcohol use ( $p = 0.02$ ), and first experience with substance ( $p < 0.001$ ).

According to Bonferroni test, using natural opiates and heroin-based drugs was significantly higher in married people than the single and divorced ( $p < 0.001$ ); and using natural opiates was significantly higher in self-employed people, retirees, and drivers than people with other occupations ( $p < 0.001$ ), while the use of heroin-based drugs was more frequent among university students ( $p < 0.001$ ). Using natural opiates was also significantly higher in the people with elementary education than the people with other education levels ( $p = 0.03$ ). Heroin-based drugs use was more frequent in the people with academic education ( $p = 0.02$ ). Individuals with

first substance experience of natural opiate were more likely to consume natural opiates while those with first substance experience of heroin-based drug were more likely to use heroin-based drugs ( $p < 0.001$ ).

This study was conducted with the aim of developing a decision tree and a model of logistic regression using the training set (474 records) and then examining their accuracy according to the testing set (204 records).

### Logistic regression

Table 2 presents the variables at risk of substance abuse using logistic regression analysis based on data sets.

The final model comprised 11 variables. The results of logistic regression reveal that place of residence, history of cigarette use, history of alcohol use, first substance experience, age and smoking age of onset have significant effects on the use of natural opiates. It is projected that natural opiate consumption's odds ratio increases by 6.42 times in non-smokers. In addition, with increasing age, the odds ratio of natural opiate consumption exceeds that of heroin based drugs by 3%.

The Hosmer-Lemeshow goodness-of-fit was derived insignificant ( $\chi^2 = 7.35$ ,  $df = 8$ ,  $p$  value = 0.49). Therefore, the null hypothesis of the model's fitting the data

**Table 2.** Multiple logistic regression analysis on the influential factors of opium in training test.

variables	OR (95% CI)	P-value
Residence		
Urban	Reference	
Rural	1.53(1.08–3.14)	0.03
Marital status		
Single	Reference	
Married	1.65(0.85–3.25)	0.14
Widow/ Divorced	1.28(0.69–4.43)	0.69
Occupation		
Student	Reference	
Housewife	0.79(0.25–1.79)	0.77
Self-employed	2.03(1.09–3.83)	0.03
Employed	6.61(1.26–8.98)	0.03
Retired	7.37(0.91–10.61)	0.09
Driver	3.63(1.25–8.11)	0.02
Education status		
University	Reference	
Primary school	0.81(0.34–1.98)	0.79
Secondary school	0.45(0.23–1.23)	0.22
High school diploma	1.12(0.58–2.95)	0.51
History of Cigarette use		
Yes	Reference	
No	6.24(2.18–28.65)	0.01
History of alcohol use		
Yes	Reference	
No	0.44(0.21–0.94)	0.03
First substance experience		
Natural opiate	Reference	
Hashish	0.05 (0.01–0.52)	0.01
Heroin based drugs	0.12(0.05–0.45)	<0.001
Age	1.03(1.01–1.08)	0.002
Substance age of onset	0.99(0.93–1.06)	0.88
Number of abstinence	0.95(0.85–1.05)	0.34
Smoking age of onset	1.03(1.001–1.10)	0.04

The reference category is heroin based drugs

well is not rejected. Cox and Snell R-squared was derived 26%, indicating that independent variables were effective in predicting substance abuse but not representing the variance in the linear regressions.

### Decision tree

In the CART decision tree model on calculated based Gini index by software R, the variables of the first substance experience, age, history of cigarette use, age of first substance, smoking age of onset, education status, and history of alcohol use are the most important factors in identifying individuals at risk. Figure 1 shows the decision tree developed in this study (size 13, 5 layers, and 7 leaves). The first experience with substance was derived to be the most effective variable that was selected to represent the tree's root node. In the next step, the variable of age and ultimately the age at first experience with substance and occupation were included in the model as important factors. The decision tree showed that if the first experience is with heroin based substance and hashish, the probability of consuming heroin based substance is 91.23%. In addition, if the first experience is with natural opiates, age is  $> 28.5$  years, the age at

smoking onset is  $> 14.5$  years, and one is employed or retired, the probability of consuming natural opiates is 89.65%. The 7 rules produced by tracing the path from the root node to each leaf node are presented in Table 3.

Sensitivity of logistic regression and the decision tree model are, respectively, 60% and 66%, and their specificity rates are 73% and 75%, respectively (Table 4). We used the area under the ROC curve ( $\pm$  standard error) to compare the two methods. The related value in the case of logistic regression amounted to  $70 \pm 0.006$  and  $72 \pm 0.007$  for the decision tree model (Figure 2). The Z-score was 1.72 ( $p = 0.08$ ), indicating no statistically significant difference between decision tree and logistic regression model.

### 4. Discussion

In the past two decades, researchers have employed predictive models to identify people at risk of contracting various diseases (Noble et al., 2011). By using the decision tree and logistic regression in the present study, we proposed a specific and accurate predictive model for identifying individuals at risk of drug abuse. Then, we assessed and compared the two models.

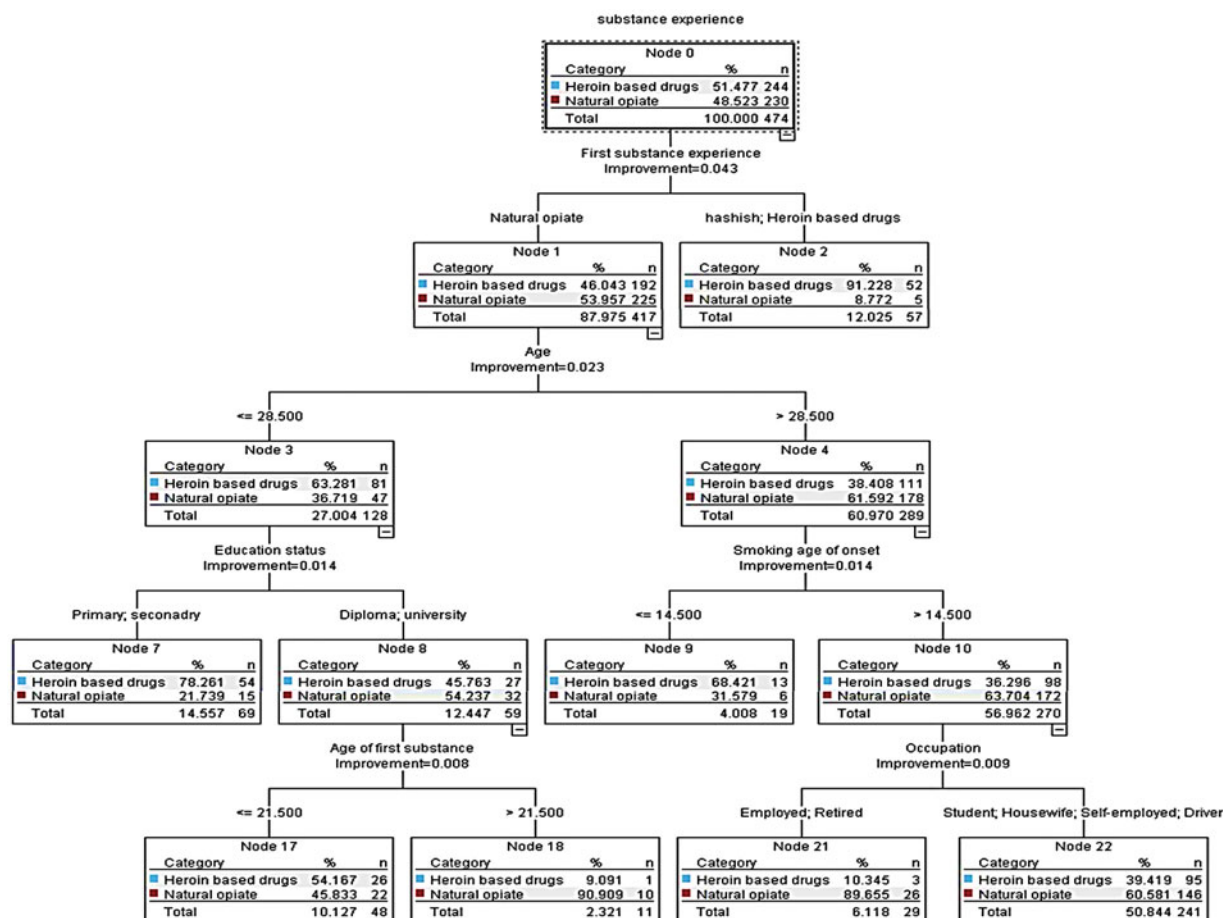


Figure 1. The decision tree with training data.



**Table 3.** The 11 rules extracted through decision tree.

R1: IF First substance experience = heroin based drugs& hash, THEN class: heroin based drugs (52/57 or 91.23%)
R2: IF First substance experience = natural opiate, age < = 28.5, Education status = primary and secondary, THEN class: heroin based drugs (54/69 or 78.26%)
R3: IF First substance experience = natural opiate, age < = 28.5, Education status = Diploma or university, Age of First substance < = 21.5, THEN class: heroin based drugs (26/48 or 54.16%)
R4: IF First substance experience = natural opiate, age < = 28.5, Education status = Diploma or university, Age of First substance > 21.5, THEN class: natural opiate (10/11 or 90.91%)
R5: IF First substance experience = natural opiate, age > 28.5, smoking age of onset < = 14.5, THEN class: heroin based drugs (13/19 or 68.42%)
R6: IF First substance experience = natural opiate, age > 28.5, smoking age of onset > 14.5, occupation = employ or retired, THEN class: natural opiate (26/29 or 89.65%)
R7: IF First substance experience = natural opiate, age > 28.5, smoking age of onset > 14.5, occupation = student or self-employed or driver, THEN class natural opiate (146/241 or 60.58)

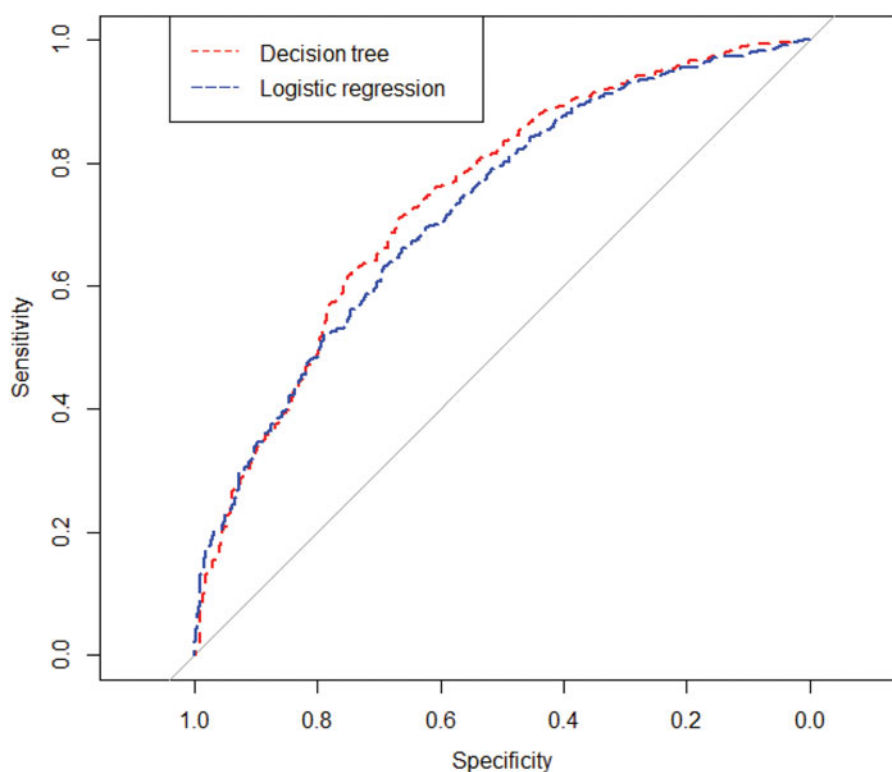
**Table 4.** Comparing the results of the logistic regression model and decision tree on training and testing data.

	Model	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC (95%CI)
Training data	Logistic regression (%)	72 (60.1–84.2)	70 (59.8–81.7)	71 (65.6–78.4)	74 (71.4–76.8)
	Decision tree (%)	82 (67.3–95.6)	78 (66.2–87.7)	81 (74.1–90.9)	86 (82.6–90.1)
Testing data	Logistic regression (%)	60 (52.6–72.6)	73 (54.1–82.6)	65 (55.7–77.6)	70 (67.7–73.0)
	Decision tree (%)	66 (59.9–77.8)	75 (63.4–79.8)	69 (63.7–74.5)	72 (69.4–75.8)

AUC: Area Under Curve

A decision tree is a popular technique with several advantages, including the capacity to model a non-linear relationship and create conceivable and understandable rules (Kammerer et al., 2005). A decision tree can be used easily and effectively in public health programmers (Yoo et al., 2012). The results of this research, emerging from the model of the decision tree used, show that, if the first drug experience is with the natural opiates, and if one is younger than 28.5 years old

and has elementary or secondary education level, then there is a 78.26% probability that he is a heroin based drug abuser. Besides that, the findings regarding the other group, showed that if the first experience is with a natural opiate, one is older than 28.5 years, his/her age at first smoking is >14.5 years, he/she is a student, self-employed or driver, then we can say that he/she is 60.58% probable to be a natural based opiate abuser (Table 3).

**Figure 2.** Receiver operating characteristics graphs of the Decision tree and Logistic regression model in testing data.

Most of the variables (i.e. age, smoking, place of residence, education, employment status) arrived at in this study were important risk factors a finding consistent with other studies (Goodarzi, Mehrpour, & Eizadi-Mood, 2011, Ahmadi et al., 2006). The logistic regression model in our study identified the variables of the first substance experience, age, history of cigarette use, place of residence, and smoking age of onset to be the most important factors (Figure 3). In another study using a logistic regression model, smoking, age, employment status, gender, and alcohol consumption were reported to be important risk factors (Meysamie et al., 2009).

Karrari et al. (2013) conducted research using the data of the current study and found that most addicts were between 30 and 40 years of age (Karrari, Mehrpour, Afshari, & Keyler, 2013). Consistent with our research, the variables of place of residence, education level, marital status, smoking, smoking and drug history were found to be the most consequential risk factors among substance abusers in previous studies (Mehrpour, Karrari, & Sheikhezadi, 2013, Noohi et al., 2011).

In the present study, most addicts had primary school and high school education. This is in line with the results of previous studies on users of natural opiates and heroin based drugs (Meysamie et al., 2009, Goodarzi et al., 2011, Kadri, Bhagylaxmi, & Kedia, 2003). This pattern suggests a high level of addiction among people with low education. Further, about 70% of individuals consuming heroin based substance and natural opiates were married; a finding consistent with earlier studies (Mehrpour, Karrari, and Sheikhezadi, 2013, Meysamie et al., 2009, Goodarzi et al., 2011).

Our findings illustrate that the decision tree model offers greater predictive accuracy than the logistic

regression model. This shows that the model of decision tree can be more helpful to physicians in identifying and screening addicts. Sensitivity and specificity are crucial when testing whether a model can accurately detect positive and negative results (Ho et al., 2012). An ideal model is one that has high sensitivity and specificity (Iran Drug Control Headquarters (IDCH) Rapid Situational Assessment 1990). In case the final purpose is screening, the model should have a higher sensitivity than specificity (Hillis et al., 2012). In this study, a comparison of performances shows that the decision tree model has a higher sensitivity and specificity than the logistic regression model; yet, in both models, sensitivity is lower than specificity. Our results correspond to previous studies (Samanta et al., 2009, Lei et al., 2015, Rastegari, Haghdooost, & Baneshi, 2013).

The ROC curve is a technique to visualize, organize, and choose classifications based on the performance of the classifications. The area under the curve (AUC) is an index of which model performs better and has a high level of accuracy. This index, which compares the performance of true positive and false positive of two different decision extremes, is often used to evaluate the predictive accuracy of classification models (Fawcett, 2006, Ke, Hwang, & Lin, 2010). In the case of testing a data set, the area under the ROC curve obtained by the decision tree model for identifying the factors proved to be higher than that in the logistic regression model. Hence, the decision tree model offers a higher accuracy than the logistic regression model.

One of the constraints of the present study is the lack of attention to biochemical and other variables, which are taken into account in internal and external studies. However, one of the strengths of our study is that few researchers have used the decision tree model

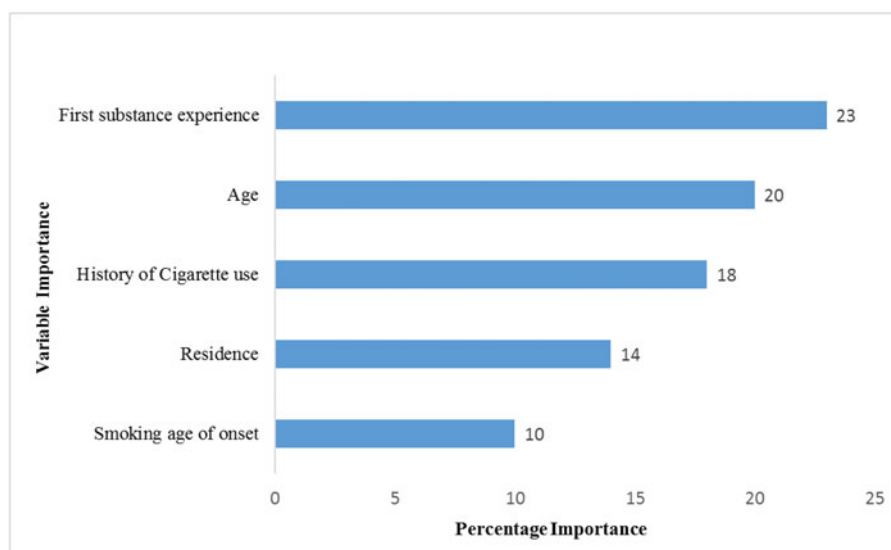


Figure 3. The importance of input variable in logistic regression.

to investigate addiction risk factors. Other researchers, however, have employed this model to identify the risk factors of other diseases because of its simple interpretation and high level of prediction (Samanta et al., 2009, Pamer, Serpi, & Finkelstein, 2008, Tayefi et al., 2017). Substance abuse represents a social problem with associated adverse health-related consequences. In countries where substance abuse is common, early intervention, particularly for the populations at risk, is essential. With regards to the remarkable variations in the pattern of substance use in Iran, identifying the populations at high risk is not only important but also useful for policy makers in order to design more efficient intervention and prevention programs. Our findings are also helpful in identifying demographic characteristics, prioritizing the most salient of predictor variables, and classifying the people who are at risk of addiction. The current study can offer clearer directions toward identifying the individuals at high risk of abusing traditional opium or newer drugs. This study can also contribute to detecting the people with multiple risk factors who are in need of planning for prevention.

## 5. Limitation

Although the present investigation possesses several assets, current study results should be interpreted in light of some limitations. First, and perhaps foremost, although our study samples were large, they were collected from treatment clinic centers and this limits the generalizability of our findings to the general population. Different sample size, i.e. number of women vs. that of men limits the interpretation of these study findings as it has not been possible to analyze the data on both genders. Epidemiological studies on drug use in Iran have indicated that overall, all types of drugs are used more frequently by men (Momtazi & Rawson, 2010, Mohammad Poorasl et al., 2007). For example RSA 2007 has reported the male to female ratio to be nearly 9:1 (Iran Drug Control Headquarters (IDCH) Rapid Situational Assessment 2007).

## 6. Conclusion

Our findings illustrate that the decision tree model offers a greater predictive accuracy than logistic regression. This shows that the model of decision tree can be more helpful to physicians in identifying and screening addicts. Using decision tree to identify factors associated with drug use in the literature is recommended. In the present study, the most important factors to identifying natural opiates and heroin based drug use were the first substance and age.

We have developed in this study a precise classification approach to identify and screen risk factors in individuals consuming natural opiates and heroin based narcotics. Given its simple and intelligible interpretation and great accuracy compared to ordinary methods, we recommend that researchers and physicians take advantage of this model.

## Conflict of interest

None

## Acknowledgment

This study was supported by Medical Toxicology and Drug Abuse Research Centre (MTDRC) of Birjand University of Medical Sciences.

## References

- Agresti, A. (1996). An introduction to categorical data analysis: Wiley New York.
- Ahmadi, J., Fallahzadeh, H., Salimi, A., Rahimian, M., Salehi, V., Khaghani, M., et al. (2006). Analysis of opium use by students of medical sciences. *Journal of clinical nursing*, 15(4), 379–86. doi:10.1111/j.1365-2702.2006.01157.x.
- AlamMehrerjedi, Z. (2013). Crystal in Iran: Methamphetamine or heroin kerack. *Daru: Journal of Faculty of Pharmacy, Tehran University of Medical Sciences*, 21(1), 22. doi:10.1186/2008-2231-21-22.
- Alam-mehrerjedi, Z., Abdollahi, M., Higgs, P., Dolan, K. (2015). Drug use treatment and harm reduction programs in Iran: A unique model of health in the most populated Persian Gulf country. *Asian journal of psychiatry*, 16, 78–83. doi:10.1016/j.ajp.2015.06.002.
- Alavi, S. S., Mehrdad, R., Makarem, J. (2016). *Prevalence of Substance Abuse/Alcohol Consumption and their Predictors among Patients Admitted in Operating Rooms of a General Educational Hospital*. Tehran, Iran: Asian Journal of Pharmaceutical Research and Health Care, 8.
- Bellazzi, R., Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *Int J Med Inform*, 77(2), 81–97. doi:10.1016/j.ijmedinf.2006.11.006.
- Ceri, S., Fraternali, P., Bongio, A., Brambilla, M., Comai, S., Matera, M. (2003). Morgan Kaufmann series in data management systems: Designing data-intensive Web applications: Morgan Kaufmann.
- Chaturvedi, H. K., Mahanta, J., Bajpai, R. C., Pandey, A. (2013). Correlates of opium use: Retrospective analysis of a survey of tribal communities in Arunachal Pradesh, India. *BMC public health*, 13(1), 325. doi:10.1186/1471-2458-13-325.
- Day, C., Nassirimanesh, B., Shakeshaft, A., Dolan, K. (2006). Patterns of drug use among a sample of drug users and injecting drug users attending a General Practice in Iran. *Harm reduction journal*, 3(1), 2. doi:10.1186/1477-7517-3-2.

- Delen, D., Walker, G., Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *ArtifIntell Med*, 34(2), 113–27.
- Fallahzadeh, M. A., Salehi, A., Naghshvarian, M., Fallahzadeh, M. H., Poustchi, H., Sepanlou, S. G., et al. (2017). Epidemiologic Study of Opium Use in Pars Cohort Study: A Study of 9000 Adults in a Rural Southern Area of Iran. *Arch Iran Med*, 20(4), 205–10.
- Farhoudian, A., Sadeghi, M., KhoddamiVishteh, H. R., Moazen, B., Fekri, M., RahimiMovaghar, A. (2014). Component analysis of Iranian crack; a newly abused narcotic substance in Iran. *Iranian journal of pharmaceutical research: IJPR*, 13(1), 337–44.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861–74. doi:10.1016/j.patrec.2005.10.010.
- Goodarzi, F., Karrari, P., Eizadi-Mood, N., Mehrpour, O., Misagh, R., Setude, S., et al. (2011). Epidemiology of drug abuse (chronic intoxication) and its related factors in a MMT Clinic in Shiraz, Southern Iran. *Iranian Journal of Toxicology*, 4(4), 377–80.
- Goodarzi, F., Mehrpour, O., Eizadi-Mood, N. (2011). A study to evaluate factors associated with seizure in Tramadol poisoning in Iran. *Indian Journal of Forensic Medicine & Toxicology*, 5(2), 66–9.
- Han, J., Pei, J., Kamber, M. (2011). *Data mining: Concepts and techniques*: Elsevier.
- Hillis, G. S., Woodward, M., Rodgers, A., Chow, C. K., Li, Q., Zoungas, S., et al. (2012). Resting heart rate and the risk of death and cardiovascular complications in patients with type 2 diabetes mellitus. *Diabetologia*, 55(5), 1283–90. doi:10.1007/s00125-012-2471-y.
- Ho, W. H., Lee, K. T., Chen, H. Y., Ho, T. W., Chiu, H. C. (2012). Disease-free survival after hepatic resection in hepatocellular carcinoma patients: A prediction approach using artificial neural network. *PloS one*, 7(1), e29179. doi:10.1371/journal.pone.0029179.
- Iran Drug Control Headquarters (IDCH) Rapid Situational Assessment. (2007).
- Jafari, S., Movaghar, A. R., Craib, K., Baharlou, S., Mathias, R. (2009). Socio-cultural factors associated with the initiation of opium use in Darab, Iran. *International journal of mental health and addiction*, 7(2), 376–88. doi:10.1007/s11469-008-9176-y.
- Jafari, S., Rahimi-Movaghar, A., Craib, K. J., Baharlou, S., Mathias, R. (2010). A follow-up study of drug users in Southern Iran. *Addiction Research & Theory*, 18(1), 59–70. doi:10.3109/16066350902825930.
- Jalilian, F., Matin, B. K., Ahmadpanah, M., Ataee, M., Alavijeh, M. M., Eslami, A. A., et al. (2016). The Personality Factors Predictors in Substance Abuse Among Iranian College Students. *International Journal of High Risk Behaviors and Addiction*, 6(1). doi:10.5812/ijhrba.27551.
- Kadri, A., Bhagylaxmi, A., Kedia, G. (2003). Study of socio-demographic profile of substance users attending a de-addiction centre in Ahmedabad city. *Indian Journal of Community Medicine*, 28(2), 74–6.
- Kammerer, J. S., McNabb, S. J., Becerra, J. E., Rosenblum, L., Shang, N., Iademarco, M. F., et al. (2005). Tuberculosis transmission in nontraditional settings: A decision-tree approach. *American journal of preventive medicine*, 28(2), 201–7. doi:10.1016/j.amepre.2004.10.011.
- Karrari, P., Mehrpour, O., Afshari, R., Keyler, D. (2013). Pattern of illicit drug use in patients referred to addiction treatment centres in Birjand, Eastern Iran. *J Pak Med Assoc*, 63(6), 711–6.
- Karrari, P., Mehrpour, O., Balali-Mood, M. (2012). Iranian Crystal: A misunderstanding of the crystal-meth. *J Res Med Sci*, 17(2), 203–4.
- Ke, W. S., Hwang, Y., Lin, E. (2010). Pharmacogenomics of drug efficacy in the interferon treatment of chronic hepatitis C using classification algorithms. *Advances and applications in bioinformatics and chemistry: AABC*, 3, 39–44.
- Lei, Y., Nollen, N., Ahluwalia, J. S., Yu, Q., Mayo, M. S. (2015). An application in identifying high-risk populations in alternative tobacco product use utilizing logistic regression and CART: A heuristic comparison. *BMC Public Health*, 15, 341. doi:10.1186/s12889-015-1582-z.
- Li, C. P., Zhi, X. Y., Ma, J., Cui, Z., Zhu, Z. L., Zhang, C., et al. (2012). Performance comparison between Logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. *Chinese medical journal*, 125(5), 851–7.
- Loh, W. Y. (2011). *Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 14–23.
- Mehrpour, O. (2013). Addiction and seizure ability of tramadol in high-risk patients. *Indian J Anaesth*, 57(1), 86–7. doi:10.4103/0019-5049.108584.
- Mehrpour, O., Karrari, P., Afshari, R. (2012). Recreational use and overdose of ingested processed cannabis (Majoon Birjandi) in the eastern Iran. *Human & experimental toxicology*, 31(11), 1188–9. doi:10.1177/0960327112446814.
- Mehrpour, O., Karrari, P., Sheikhzadi, A. (2013). Survey of factors related to criminal behavior in a sample of Iranian substance abusers. *Journal of forensic and legal medicine*, 20(8), 1078–81. doi:10.1016/j.jflm.2013.09.022.
- Mehrpour, O., Sheikhzadi, A., Barzegar, A., Husein, A., Malic, C., Sheikhzadi, E., et al. (2016). Comparison of Quantitative and Qualitative Dermatoglyphic Characteristics of Opium Addicts with Healthy Individuals. *Addiction & health*, 8(2), 76.
- Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung J Med Sci*, 29(2), 93–9. doi:10.1016/j.kjms.2012.08.016.
- Metcalfe, D., Olufajo, O. A., Salim, A. (2015). Pre-hospital opioid analgesia for traumatic injuries. *Status and date: New, published in*, (9), Art. No.: CD011863. doi:10.1002/14651858.CD011863.
- Meysamie, A., Sedaghat, M., Mahmoodi, M., Ghodsi, S. M., Eftekhari, B. (2009). Opium use in a rural area of the Islamic Republic of Iran. *Eastern Mediterranean health journal = La revue de sante de la Mediterraneeorientale = al-Majallah al-sihhiyah li-sharq al-mutawassit*, 15(2), 425–31.
- Mohammad Poorasl, A., Vahidi, R., Fakhari, A., Rostami, F., Dastghiri, S. (2007). Substance abuse in Iranian high school students. *Addict Behav*, 32(3), 622–7. doi:10.1016/j.addbeh.2006.05.008.
- Momtazi, S., Rawson, R. A. (2010). Substance abuse among Iranian high school students. *Current opinion in psychiatry*, 23(3), 221. doi:10.1097/YCO.0b013e328338630d.



- Morrison, G (1990). *Clinical Methods: The History, Physical, and Laboratory Examinations*. Boston: Butterworths.
- Nassr, M., Daneshamuz, B., Gharai, B., Salehi, M., Ardebili, M. E., Ghalebandi, M. F. (2006). A Household Study on the Prevalence of Substance Misuse in Tehran: The need for other methods to estimate the prevalence. *Iranian Journal of Psychiatry*, 1(4), 158–61.
- Nikfarjam, A., Shokoohi, M., Shahesmaeili, A., Haghdoost, A. A., Baneshi, M. R., Haji-Maghsoudi, S., et al. (2016). National population size estimation of illicit drug users through the network scale-up method in 2013 in Iran. *International Journal of Drug Policy*, 31, 147–52. doi:10.1016/j.drugpo.2016.01.013.
- Noble, D., Mathur, R., Dent, T., Meads, C., Greenhalgh, T. (2011). Risk models and scores for type 2 diabetes: Systematic review. *BMJ (Clinical research ed)*, 343, d7163. doi:10.1136/bmj.d7163.
- Noohi, S., Azar, M., Behzadi, A. H., Sedaghati, M., Panahi, S. A., Dehghan, N., et al. (2011). A comparative study of characteristics and risky behaviors among the Iranian opium and opium dross addicts. *Journal of addiction medicine*, 5(1), 74–8. doi:10.1097/ADM.0b013e3181db69ef.
- Pamer, C., Serpi, T., Finkelstein, J. (2008). Analysis of Maryland poisoning deaths using classification and regression tree (CART) analysis. *AMIA Annual Symposium proceedings AMIA Symposium*, 550–4.
- Rahimi-Movaghar, A., Amin-Esmaeili, M., Shadloo, B., Noroozi, A., Malekinejad, M. (2015). Transition to injecting drug use in Iran: A systematic review of qualitative and quantitative evidence. *International Journal of Drug Policy*, 26(9), 808–19. doi:10.1016/j.drugpo.2015.04.018.
- Ramezankhani, A., Hadavandi, E., Pournik, O., Shahrabi, J., Azizi, F., Hadaegh, F. (2016). Decision tree-based modelling for identification of potential interactions between type 2 diabetes risk factors: A decade follow-up in a Middle East prospective cohort study. *BMJ open*, 6(12), e013336. doi:10.1136/bmjopen-2016-013336.
- Ramezankhani, A., Pournik, O., Shahrabi, J., Khalili, D., Azizi, F., Hadaegh, F. (2014). Applying decision tree for identification of a low risk population for type 2 diabetes. *Tehran Lipid and Glucose Study. Diabetes Res Clin Pract*, 105(3), 391–8.
- Rastegari, A., Haghdoost, A. A., Baneshi, M. R. (2013). Factors Influencing Drug Injection History among Prisoners: A Comparison between Classification and Regression Trees and Logistic Regression Analysis. *Addict Health*, 5(1–2), 7–15.
- Rokach, L., Maimon, O. (2014). *Data mining with decision trees: Theory and applications*: World scientific.
- Rush, B. R., Wild, T. C. (2003). Substance abuse treatment and pressures from the criminal justice system: Data from a provincial client monitoring system. *Addiction*, 98(8), 1119–28. doi:10.1046/j.1360-0443.2003.00420.x.
- Samanta, B., Bird, G. L., Kuijpers, M., Zimmerman, R. A., Jarvik, G. P., Wernovsky, G., et al. (2009). Prediction of periventricular leukomalacia. *Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms. ArtifIntell Med*, 46(3), 201–15.
- Sarrami, H., Ghorbami, M., Taghavi, M. (2013). The survey two decades of prevalence studies among Iran University students. *Research on addiction*, 7(27), 9–36.
- Singh, G., Garg, P., Singh, M., Arora, R., Garg, A. (2016). Emerging Trends and Prevalence of Drug Abuse: A Study Conducted at Swami Vivekananda Drug De-Addiction and Treatment Centre. *Dual Diagnosis: Open Access*. doi:10.21767/2472-5048.100008.
- Tayefi, M., Esmaeili, H., SaberiKarimian, M., AmirabadiZadeh, A., Ebrahimi, M., Safarian, M., et al. (2017). The application of a decision tree to establish the parameters associated with hypertension. *Computer methods and programs in biomedicine*, 139, 83–91. doi:10.1016/j.cmpb.2016.10.020.
- Tayefi, M., Tajfard, M., Saffar, S., Hanachi, P., Amirabadizadeh, A. R., Esmaeily, H., et al. (2017). Hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. *Computer methods and programs in biomedicine*, 141, 105–9. doi:10.1016/j.cmpb.2017.02.001.
- United Nations Office on Drugs and Crime. (2011). *World Drug Report*.
- Williams, D., Padmanabhan, V. (2009). Substance misuse and intoxication in adolescents. *The Foundation Years*, 5(2), 67–71. doi:10.1016/j.mpfou.2008.12.004.
- Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., et al. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *J Med Syst*, 36(4), 2431–48. doi:10.1007/s10916-011-9710-5.