I J C T A, 9(19) 2016, pp. 9141-9149

Comparative Analysis of Data Mining Techniques for Optimal Attribute Selection among Drug Abusers using Weka 3-6-10

Shubpreet Kaur[*] and R. K. Bawa[*]

ABSTRACT

A variety of psychoactive substances has a greater effect on the person's daily life and work competence which is associated with many adverse consequences to medical diseases. Hospitals across the nation are swamped with countless patients suffering from hazardous diseases due to drug abuse. This paper explores the epidemiology of number of problems focusing the treatment for drug abusers. This is a disease that may have particular implication for selecting substance abuse treatment forms. We can apply this to develop a system to estimate risk factors for drug abusers in the real world. Implementation is done using WEKA. Results are computed manually, firstly confusion matrix is compared and then accuracy is calculated that illustrates the major factors that need to be controlled. Accuracy of Age of initiation of drugs is 100% i.e. upto25yrs of age factor is critical, who prompted them to take drugs is 98.03% i.e. outside environment is the major factor towards increasing the abuse and who provided you the drug 96.07% is also the outside parameters that leads to the initiation of drug abuse. Navie bayes, J48, and ID3 data mining techniques are used but among them ID3 algorithm gives the more accurate result and proves the strong inter-relation between the parameters.

Keywords: Drug addicion, data mining techniques, WEKA, Medical diseases.

I. INTRODUCTION

Drug abuse has become a major social problem over the years. Drug addiction is not only a disease that lands the individual in distress but also it badly affects the family [4]. Drugs has become an integral part of one's physical, emotional and social life, the victim encounter with hell lot of problems such as Physical, Psychological, Spiritual and Social [8]. Drug addiction not only a disease that lands the individual in distress but also it badly affects the family. More than 2.6% of people suffer from drug addiction at some time in their life [9].

Drug addiction in Punjab is becoming next to impossible to cure from its deadly state of addiction [2]. According to Ministry of youth Affairs and Sports in Punjab, 15 to 25 years have fallen prey to drugs to about 40 percent by Punjabi youth. Statistics by 2011 report on drug abuse and alcoholism suggest roughly 1.5 to 2 million Punjabi youth is addicted to drugs [5].

Data mining is an analytical process which discovers most important information from the data warehouse of the organizations. Data mining does pre processing. The advantage of using data pre-processing is it reduces memory. The existence of data mining came in the middle of 1990's. Data mining is the most vital and motivating area of research as data ware house contains very vast and complex data and using data mining tools and methods patterns are extracted. Predictive data mining is to exploit and identify patterns from historical datasets and predict the outcome of a disease.

[*] Department of Computer Science, Punjabi University, Patiala, Punjab, *E-mails: shubpreetkaur@gmail.com; rajesh.k.bawa@gmail.com*

II. RELATED WORK

HEART DISEASE is the 1st leading cause of death in the world in the past 10 years and logistic regression, regression tree and neural networks are the data mining techniques are used [7]. By 2030, it is estimated that 23.6 million people will die from CVD (heart disease and stroke). Akhil Jabbar [13] proposed a system heart disease prediction using data mining techniques in Andhra Pradesh.

BREAST CANCER is 2nd leading cause of death for women that ages 40 – 59. Sahan et al [6] used hybrid method based on fuzzy and k-nn for diagnosis of breast cancer.

DRUG ADDICTION, the government of Jammu and Kashmir is contemplating to introduce a special subject on "hazards of smoking and drug addiction" in the curriculum from primary to higher secondary level in all schools of the state. McBride, R. & Mosher, J. F [1] have studied the public health focusing on usage of alcohol especially influenced from European style.

III. DATA MINING (CLASSIFICATION) METHODS

Classification is a supervised learning technique which is used to predict the data from unknown sample size and class categories. Here we list some predictive data mining techniques which are used frequently by various data mininers from relevant KDNuggets (http://www.kdnuggets.com/polls/2006/data mining methods.htm, April 2006).

2.1. Decision Tree

A decision tree (DTree) is a non-parametric supervised learning technique with a schematic tree-shaped structure to take a problem with multiple probable keys and exhibit it in a simple, easy-to-understand layout that illustrates the association between different events or decisions. ID3 and J48 are the common decision tree algorithms. ID3 is used to create tree from dataset. J48 is the implementation of ID3 algorithm. ID3 is used to create tree from dataset [3].

2.2. Navie bayes

A bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from bayesian statistics) with strong (naive) independence assumptions. The term naive in the name naïve bayes draws from the fact that the algorithm uses bayesian procedures but does not take into account dependencies that may exist.

2.3. K-Nearest Neighbor (KNN)

KNN is considered to be simple, easy to understand, versatile and non parametric lazy learning algorithm. The principle of the KNN algorithm is to make use of the database in which the data points are split into several disconnect classes to forecast the classification of a new sample point. KNN classification partitions the data into a test set and a training set.

IV. DATA MINING SUITE WEKA 3-6-10

WEKA (Waikato Environment for Knowledge Analysis) is most commonly used java based data mining tool. It was developed by department of Computer Science, University of Waikato, New Zealand in 1993 and supports several data mining tasks ranging from preprocessing, classification and clustering to visualization and feature selection. WEKA supports own ARFF (attribute relation file format), CSV (comma separated values), Lib SVM and C4.5's format. The particular potency of WEKA is the capacity to acquire data from both SQL databases and from actual web pages by supplying the URL of the webpage containing the information. Apart from providing a toolbox of learning algorithms, WEKA also offers framework

through which researchers can implement new algorithms without bothering about the underneath infrastructure for data manipulation and scheme evaluation.

V. OBJECTIVES

1. To study various classification techniques for attribute selection
2. Implementation of various classification techniques using weka
3. Comparative analysis of data mining techniques?

VI. RISK FACTORS

There are enormous people suffering from variety of diseases these days. Cancer is becoming common like flu, cold. Drugs play a major role behind such incurable diseases [16]. Below is the table 1 showing the foremost reason for these fatal diseases.

Table 1
Risk factors involved for the prediction of various medical diseases [18]

S.No	Disease	Risk factors
1.	Lung cancer	Smoking (beedi, hookah or cigarette) or second hand smoke, High dose of ionizing radiation, Air pollution, Insufficient consumption of fruits & vegetables
2	Skin cancer	UV light, smoking, large no. of moles on the skin, family history of skin cancer, work outdoors
3	Breast cancer	Drinking alcohol, tobacco, passive smoking, obesity and lack of exercise, night shift work.
4	Heart disease	Smoking, Diabetes, high blood pressure, high cholesterol, low cholesterol, not getting enough physical activity and obesity.
5	Kidney failure	Diabetes, High blood pressure, Heart disease, Smoking, Obesity, High cholesterol, Family history of kidney disease
6	Liver Disorder	Excessive alcohol consumption, Obesity, Diabetes, Tobacco use, Cirrhosis, Hereditary, Exposure to aflatoxins, viruses(primarily hepatitis A [HAV], hepatitis B [HBV], or hepatitis C [HCV])

VII. ATTRIBUTE SELECTION

A survey of 200(approx) patients admitted at various drug addiction centres of Punjab has been interviewed. Parameters and values are shown below that contributes towards the drug addicted people.

Table 2
Parameters for drug addicted people [18]

Parameters	Values
Sex	Male, Female
Age	upto25,25_40,above40
Residence	urban, rural
Type of locality	slum, private, govt. approved
Marital Status	Married, Unmarried, Divorced
Type of family	Joint, Nuclear
Education	Illiterate, Primary, Less than primary, Secondary, graduation, above graduation
Occupation	unemployed, student, job, business
Age of Initiation	upto25, above25
Duration of substance abuse	1_10 yrs,11_20 yrs ,above 20yrs
Family Income	Rs(upto5000, 6_10000, 11_25000, 26_40000, above40000)
House	Own, Rented
Take more drugs at a time	Yes, No
Can you go through a week without taking drugs	Yes, No
Who prompted you for drugs	internal (personal) factors, external factors
In which company it was taken	internal (house, shop, etc) place, external place (marriage , schools, colleges, etc)
Family history of drugs	Yes, No
Relation with family	Aggressive, Polite
Daily expenditure on drugs	Rs (1_200, 201_500, 501_1000, above1000)
In case no money for drugs	Yes(stealing, theft, borrow, etc), No
Did you sell any house hold articles if you don't have money for buying drugs	Yes, No
Do you suffer from medical problems	Yes, No
Any accident while having drugs	Yes (1time or more), No
Conflict with law	Yes (1time or more), No
Anyone helped you in getting rid of this drug taking problem	Yes, No

VIII. PERFORMANCE EVALUATION CRITERIA [KRZYSZTOF J. CIOS, G. WILLIAM MOORE]

To measure the performance of classifiers, we compute accuracy. Accuracy is defined as percentage of rows that are correctly classified instances [11].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- Where TP, TN stands for True Positive and True Negative and FP, FN stands for False Positive and False Negative. Where TP signifies true A's that are correctly classified as A
- TN signifies all other classes correctly classified as not A
- FP signifies other classes incorrectly classified as A
- FN signifies A's that were incorrectly classified as not A

IX. PROPOSED METHODOLOGY USING WEKA TOOL

We have applied various data mining techniques in weka, but first is the data that is being interviewed at various drug de addiction centres of Patiala, Ludhiana and Mohali district of Punjab. Then dataset is made from the data. Data is then loaded into the WEKA tool. Naïve Bayes, Decision tress (J48), ID3, KNN classification techniques are applied as shown in figure1. Results obtained are then cross validated and their performance is analyzed in terms of Accuracy.

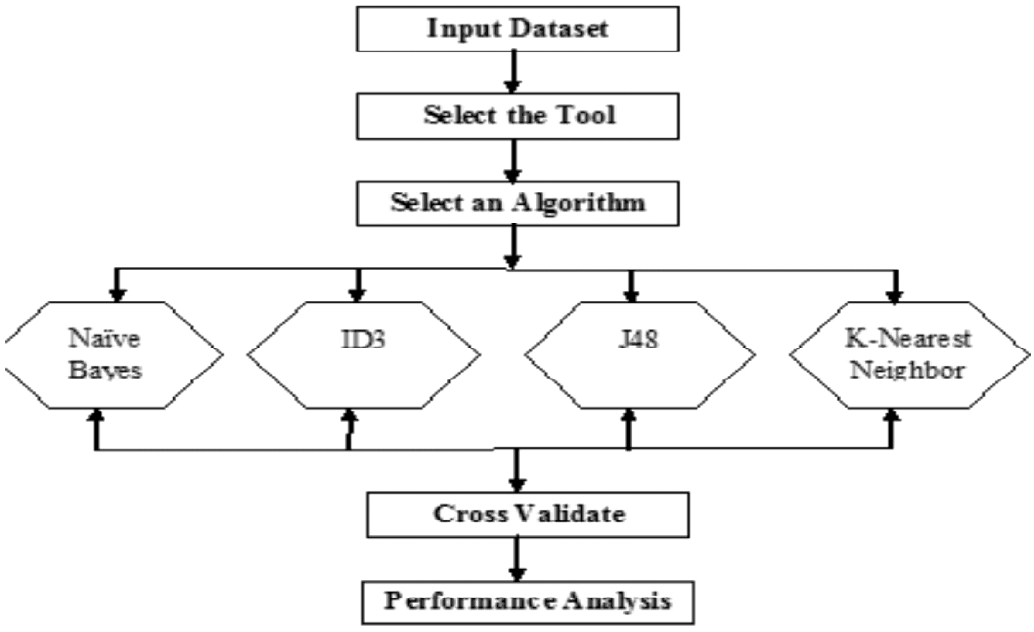

Figure 1: Working of WEKA tool

X. RESULTS

1. Age of initiation of drugs

Age of initiation of drugs play an important role where two parameter values upto25 and above25 yrs of initiation age is taken. Confusion matrix is a layout which visualize the performance. Here, column classifies instances of predicted class and row signifies instances of actual class. Confusion matrix in figure 2. depicts upto25 is the alarming age that needs to be controlled.

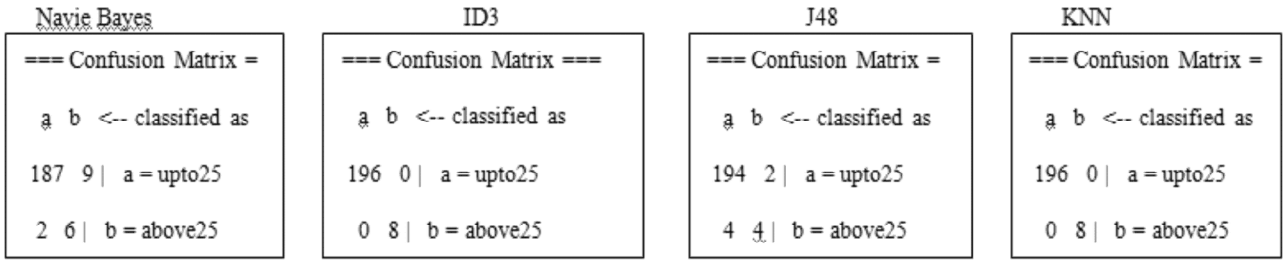

Figure 2: Confusion matrix of age of initiation of drugs

Below is the table 3 that shows the accuracy computed from the confusion matrix.

Table 3
Comparison of different classification techniques on Age of Initiation

Age of Initiation (Upto25/ Above25)

Algorithm	Correctly classified Instances	Incorrectly classified Instances
Navie Bayes	94.60%	5.39%
ID3	100%	0%
J48	97.05	2.94%
KNN	100%	0%

2. Who prompted you to take drugs?

It is broadly categorized into inside/outside where inside covers internal factors like family problems, academics pressure etc and outside covers external factors like peer pressure, trend, society, work/job conditions etc. as shown in figure 3.

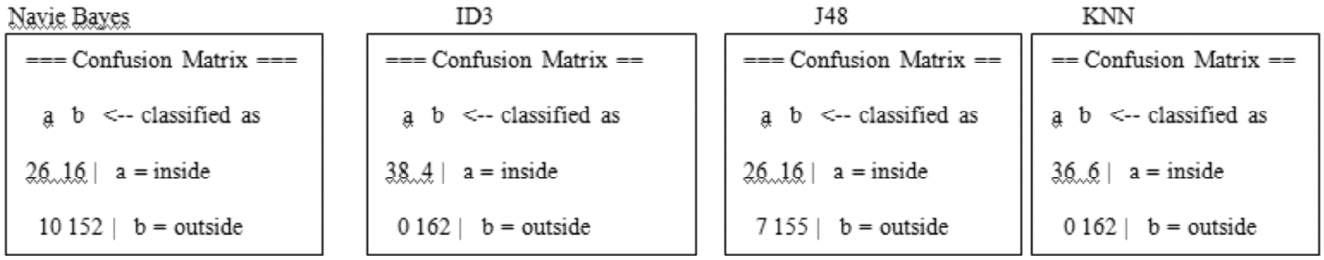

Figure 3: Confusion matrix of who prompted you to take drugs

The confusion matrix shows the external factors is affecting the early life of the person which destroy the later life, either it is an illness or death which affect only the family to whom we less prioritize than taking drugs.

Table 4
Comparison of different classification techniques on who prompted you to take drugs

Who prompted you (Outside/ Inside)

Algorithm	Correctly classified Instances	Incorrectly classified Instances
Navie Bayes	87.25%	12.74%
ID3	98.03%	1.96%
J48	88.72%	11.27%
KNN	97.05%	2.94%

Above is the accuracy shown in table 4 depicting that how much the initiator is easily trapped into the hands of drugs.

3. Who provided you drugs at initiation time?

The confusion matrix of who provided you drugs for the first time in figure 4 reveals the external factors like friends, marriages and parties etc to be hilarious.

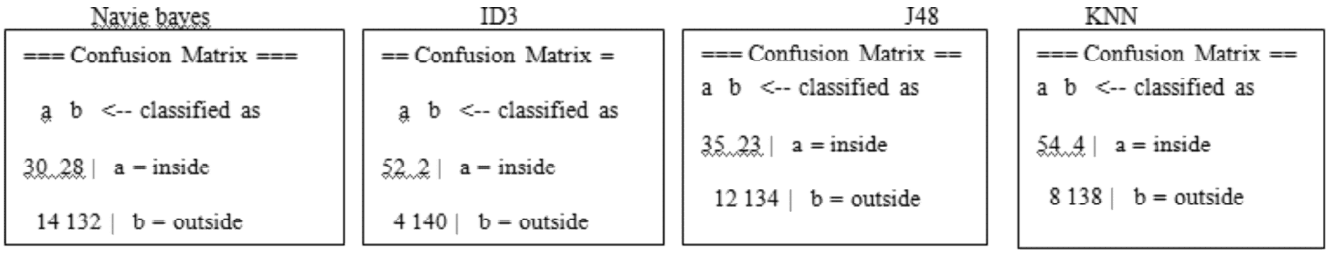

Figure 4: Confusion matrix of who provided you drugs at initiation time

Accuracy of who provided you drugs uncovers greater part of population is affected as shown in table 5. Outside parameters raise the inner level of person to go for drugs.

Table 5
Comparison of different classification techniques on who provided you drugs at initiation time

Who provided you (Outside/ Inside)

Algorithm	Correctly classified Instances	Incorrectly classified Instances
Navie Bayes	79.41%	20.58%
ID3	94.11%	2.94%
J48	82.84%	17.15%
KNN	94.11%	5.88%

4. In which company it was taken for the first time?

Confusion matrix in figure 5 reveals the external factors are at large responsible for their support or insist them to initiate.

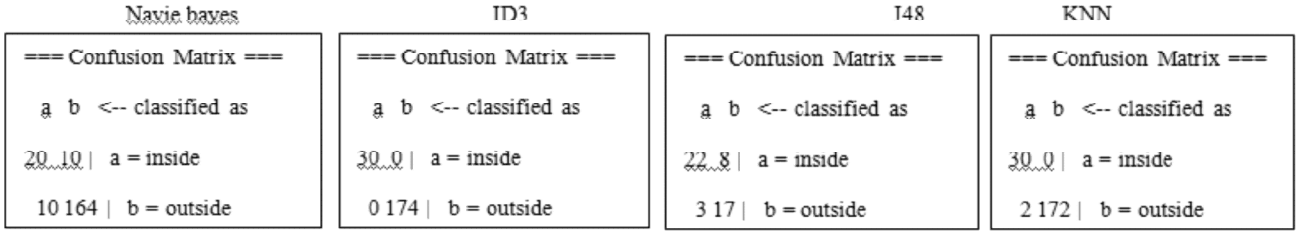

Figure 5: Confusion matrix of in which company it was taken for the first time

Outside sources has flare up the accuracy to 100% that implies linking to other parameters as shown in table 6.

Table 6
Comparison of different classification techniques in which company it was taken

In which company it was taken (Outside/ Inside)

Algorithm	Correctly classified Instances	Incorrectly classified Instances
Navie Bayes	90.19%	9.80%
ID3	100%	0%
J48	94.60%	5.39%
KNN	99.01%	0.98%

Among all data mining techniques, ID3 gives more accurate results than other comparative data mining techniques.

XI. INTERPRETATION OF RESULTS

The best technique among other techniques is ID3 and major the affected factor is Age of initiation which is below25 years of age which is showing at high rates as compared to other attributes. The attribute "In which company it was taken" is also affected at major concern that shows the root cause problem to drugs and its disastrous effects arise at age below25 yrs of age and it is taken outside and the parameter "who prompted you" results in external factors to encourage them to do so as shown in figure 6. There is urgent need to educate youth about the drugs and to make them aware of their ill-effects. For this students should be encouraged to participate in drug awareness camps. Support for drug prevention should come from all sides including families, friends, schools, community groups and health professionals.

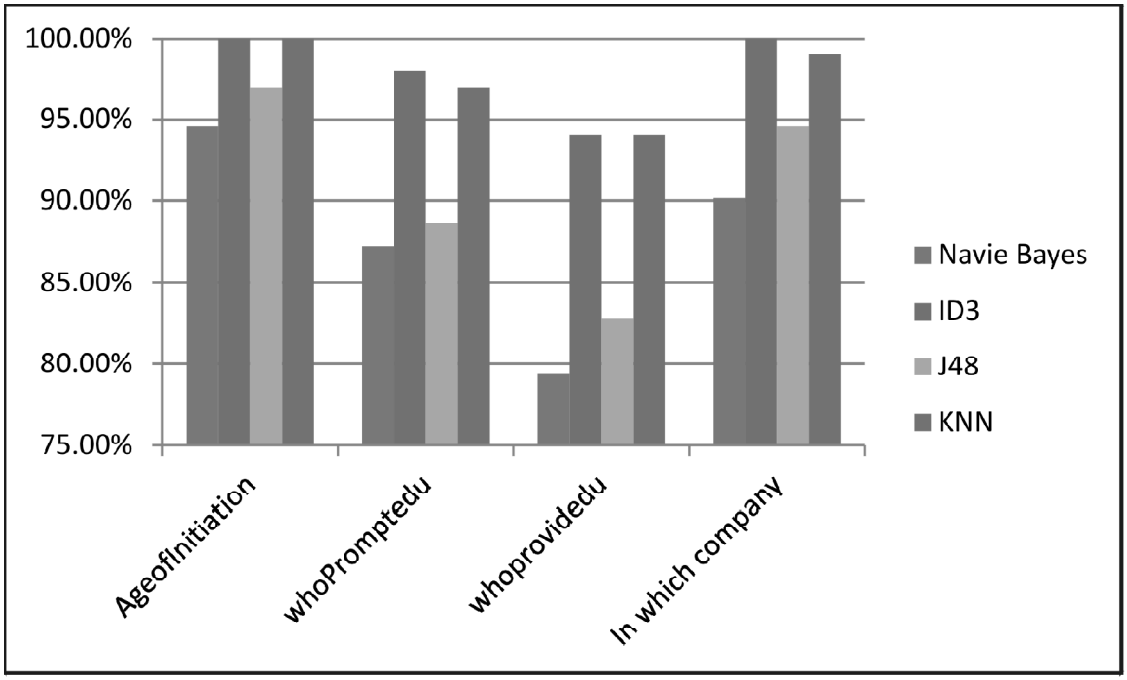

Figure 6: Accuracy interpretation of drug initiation parameters using various data mining techniques

XII. CONCLUSION

Drug addiction is a strong dependence on an illegal drug or a medication. In drug addiction, drugs hijack your brain, your mind, and your life [17]. They literally change your brain that is why we call addiction a brain disease. There are millions of people suffered and suffering because of the dual characteristics of opium, as well as of many other narcotic drugs and many psychotropic substances, both natural and synthetic 10]. High addictions lead to the dreadful reasons for drug abuse.

The results of data mining techniques conclude that Peer pressure is the foremost rationale of using drugs and there is urgent need to educate the youth which are falling in bad habit of abusing drugs. Drugs are bombarding the studies and academic career of the youth which needs special attention towards it.

We have seen among 204 patients, at various drug de-addiction centres there 118 rural area populations was there and their occupation is agriculture and due to their working conditions at extreme temperatures and working for long hours, it turn out to be a necessity to take drugs and it's also their family culture going. They came into drug de-addiction centres when they are caught by various diseases. For the most part they have their own house (not rented). Out of 204 patients, 61% patients suffered from accidents many times. 74% inhabitants does theft, borrow for intense craving for drugs. 68% patients have family history of drugs. By and large, 86% take drugs outside i.e. with friends, or at work/job.

Results of Navie bayes computed manually are at par with the weka implementation.

Age of initiation, who prompted you to take drugs, who provided you drugs at first time of initiation, and in which company it was taken are observed to beginners in one's life. There accuracies collectively bring into light the external sources are responsible for it in a larger area. There is an immediate prerequisite to have control on these parameters. Once this adolescence age is managed then in the later life they have very less chances to get lock into drugs. Today's youth are future of tomorrow. If they are not preserved today then they have no future. Drugs lose their potential to go for success. Drugs neither give enjoyment and nor relieves stress, it's a temporary till we have taken drugs. But it has long lasting effects on one's physical and psychological effects.

De-addiction is the process of becoming drug free. And this can be achieved by reducing drugs day by day because stopping addiction is impossible in one day. This is possible when the person admits himself and acknowledges the problem and this process is known as recovery. Recovery is a way to *regain* control. It is very hard situation because extreme craving prevents to do so. Only strong realization and determination can do so. Here teen age is to study more to have a good job.

XIII. FUTURE WORK

The need to develop a system where every internet user can check his/her level of addiction if he/she is addicted to drugs and proposed system should be developed online, then research work will be of great help for every mobile user and drug de-addiction centers by helping them in taking the inputs from the suspicious person through an interactive computer interface as drug addicts hesitate in discussing the problem with anybody.

REFERENCES

[1] Rob McBride and James F. Mosher J.D, "Public Health Implications of the International Alcohol Industry: Issues Raised by a World Health Organisation Project", vol 80, pp 141-147, 1985.

[2] G. W. Neat, Rensselaer Polytech. Inst., Troy, H. Kaufman, R. J. Roy, "Expert adaptive control for drug delivery systems", (Wiley), vol.28, no.4, pp. 443–458, 2000.

[3] Imran Kurt, A. Turhan Kurum, Kazim Ozdamar , "Comparing classification techniques for predicting essential hypertension", Expert Systems with Applications (Elsevier), vol 29, pp 583–588, October 2005.

[4] Tae-Suk Kim, Dai-Jin Kim , Heejin Lee , Yong-Ku Kimb, "Increased plasma brain-derived neurotrophic factor levels in chronic smokers following unaided smoking cessation", vol 423, pp 53–57, 9 August 2007.

[5] Baezconde-Garbanati L, Beebe L, Perez-Stable E, "Building capacity to address tobacco-related disparities among in American Indian and Hispanic/Latino communities: Conceptual and systemic considerations", Journal of Addiction (wiley), vol 102, pp. 112–122, 2007.

[6] Seral Şahan , Kemal Polat, Halife Kodazb, , Salih Güneş, "A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis", Computers in Biology and Medicine (Elsevier), vol 37, pp 415–423, March 2007.

[7] Imran Kurt ,Mevlut Ture , A. Turhan Kurum , "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease", vol 34, Pp 366–374, January 2008.

[8] Imran Kurt, Mevlut Ture, A. Turhan Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease", Expert Systems with Applications (Elsevier), vol 34, pp 366–374, January 2008.

[9] Mevlut TureMolly Magill and Lara A. Ray, "Cognitive-Behavioral Treatment with Adult Alcohol and Illicit Drug Users: A Meta-Analysis of Randomized Controlled Trials" Alcohol Drugs.; vol 70, pp. 516–527, July 2009.

[10] Pauline Hussaarts, Hendrik G. Roozen PhD , Robert J. Meyers PhD, Ben J.M. van de Wetering and Barbara S. McCrady, "Problem Areas Reported by Substance Abusing Individuals and Their Concerned Significant Others , vol 21 , pp. 38–46, 2012.

[11] Karen J. Hartwell , Sudie E. Back, Aimee L. McRae-Clark, Stephanie R. Shaftman, Kathleen T. Brady, "Motives for using: A comparison of prescription opioid, marijuana and cocaine dependent individuals", Addictive Behaviors (Elsevier),Vol 37, pp 373–378, April 2012.

[12] P.K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", Journal of King Saud University - Computer and Information Sciences (Springer), vol 24, pp 27–40, January 2012.

[13] Chaitrali S. Dangare Sulabha S. Apte," Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, vol 47, pp.44-48, June 2012

[14] M. Akhil Jabbar, B. L Deekshatulu ; Priti Chandra," Heart disease prediction using lazy associative classification", International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013, March 2013, pp. 40 – 46.

[15] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, Sebastián Ventura,"Predicting students' final performance from participation in on-line discussion forums", Computers & Education (Elsevier), vol 68, pp 458–472, October 2013.

[16] Luqman M, Javed MM, Daud S, Raheem N, Ahmad J, Khan AU, "Risk factors for lung cancer in the Pakistani population", vol 15, pp 3035-3039, 2014.

[17] Sneha Chandra, Maneet Kaur, "Creation of an Adaptive Classifier to enhance the classification accuracy of existing classification algorithms in the field of Medical Data Mining", 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, March 2015, pp 376 – 381

[18] Raye Z. Litten, Megan L. Ryan, Daniel E. Falk, Matthew Reilly, Joanne B. Fertig and George F. Koob, "Heterogeneity of Alcohol Use Disorder: Understanding Mechanisms to Advance Personalized Treatment", Alcoholism, Wiley online library, vol 39, pp 579–584, April 2015.

[19] Shubpreet kaur, Dr. R. K. Bawa, "Future Trends in Medical Healthcare System: Implementation and Analysis of Data Mining Techniques among Drug Abusers", International Journal of Energy, Information and Communications vol.6, pp.17-34, 2015.

[20] Noreen Kausar, Sellapan Palaniappan, Brahim Belhaouari Samir, Azween Abdullah, Nilanjan Dey, "Systematic Analysis of Applied Data Mining Based Optimization Algorithms in Clinical Attribute Extraction and Classification for Diagnosis of Cardiac Patients", Applications of Intelligent Optimization in Biology and Medicine, Springer, vol 96, pp 217-231, 2016.