Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study

Farhad Soleimanian Gharehchopogh
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, West Azerbaijan, Iran

Peyman Mohammadi
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, West Azerbaijan, Iran

Parvin Hakimi
Department of Obstetrics and Gynecology,
Tabriz University, Tabriz, Iran

ABSTRACT
By means of data mining techniques, we can exploit furtive and precious information through medicine data bases. Because of huge amount of this information, study and analyses are too difficult. We want some methods to exploring through data and extract valuable information which can be used in the future similar cases. One of these cases is accouchement. The mechanism of accouchement is a natural and spontaneous process without the need to any intervention. In some conditions, maybe mother, baby or both of them are in hazard and need help and support. This help is provided by Caesarian Section which saves mother and baby. Nevertheless, we need to know when we should use surgery. This study explains utilization of medical data mining in determination of medical operation methods. We render this with accumulating 80 pregnant women information. The results show that decision tree algorithm designed for this case study generates correct prediction for more than 86.25% tests cases

Keywords
Data Mining, Knowledge Discovery, Cesarean Section, Decision Tree.

1. INTRODUCTION
Recent findings in collecting data and saving results have led to the increasing size of databases. Medicine craft is one of cases that use huge databases. Main point in these massive databases is information and knowledge extracted from them. In today's world, collecting a wide variety of data about different illnesses has significant importance. Health centers collect these data for many reasons. The generic and main reason is facilitating medicament works. Using this information leads to patient receiving as best service in minimum possible time.

Data has been stored to be used in the future and doctors will gain from saved information in similar status. For example knows dues of medics on some of patients can use for other cases in future to prevent dues or we can predict hospitalization of patients after section also an analyzer could analyses healthcare quality indicator.

Some of stored data in clinical databases can be demographic information, disease and treatment status, the tests and their results, prescribed drugs, billing and administrative tasks [1]. Surgical operations are most sensitive between medicine acts, so need much more information and witting to choose the best method of execution.

The term "Data Mining" was introduced in 1990s, but data mining is the evolution of a field with a long history [2]. As datasets are growing in size, applications and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been achieved by other discoveries in computer science such as artificial neural networks (ANNs), clustering algorithms plans. Data mining techniques has been accomplished for genetic algorithm (GA) in 1950s, and for decision trees (DTs) in 1960s. Also it's supported vector machine (SVM) in 1990s methods [3]. The first use of data mining techniques in health information systems was fulfilled with the expert systems are developed since 1970s [4].

In healthcare, data mining is becoming gradually more desirable, and now it's more essential. For example, the existence of medical insurance fraud and abuse has led many healthcare insurers to attempt for reduce their losses by means data mining tools to help them find and track offenders [5]. Fraud detection using data mining applications is prevalent in commercial world, e.g. in the detection of fraudulent credit card transaction [6].

Clinical determinations are often focusing on physician's sense and experience based on the knowledge that comes from huge databases of hospitals. Data are in datasets and need techniques to discover them and use in clinical decisions. This information must be evaluated for medical researches to be used in health centers.

The organization of this paper is as follows. Section 2, describes data mining techniques and DTs algorithms and Section 3 explains data mining application in healthcare domain. The proposed method for discovering useful information in cesarean section are presented in section 4 and Section 5 contains discussions about proposed method. The paper ends in conclusion and proofs out come in section 6.

2. DATA MINING
Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques (Gartner Group). It likes mining gold through huge rocks and large amount of soils because data mining is extracting knowledge from bulky data sets. But it isn't exactly searching and gain knowledge, it's some different and is earning knowledge with analyzing and reconsidering. Data mining is the discipline of growing interest and importance, and an application area that can

provide significant competitive advantage to an organization by exploiting the potential of large data warehouses [7].

The goals of data mining divide it into two general types as predictive and descriptive [4]. By application of the predictive type, we understand that application of the model which attempts to 'predict' the value that a certain variable may take, given what we know at present [11]. Predictive data mining methods may be applied to the construction of decision models for procedures such as prognosis, diagnosis and treatment planning, which–once evaluated and verified–maybe embedded information systems [12]. Descriptive data mining tasks characterize the general properties of the data in the database [4]. At this stage, the model can be used from an applicative point of view only as a simple summarized (overall) description of the actual situation [11]. In descriptive methods, there are no hypotheses of causality among the available variables [13].

2.1 Data Mining Techniques

Knowledge Discovery (KD) uses data mining generally consists of seven phases: Data cleaning, Data integration, Data Selection, Data transformation, Data mining, Pattern evaluation, Knowledge presentation.

In data cleaning process noise, inconsistent and nonessential data must have been removing. Data integration combines multiple data sources which maybe needful.

Data selection process is to form a data file by selecting data from the database or other data sources which b being worked on [8]. In some machine learning schemes such as ANN and SVM, each instance data is represented as a vector of real numbers. Therefore, we have to convert the nominal attributes into numeric data before feeding into classifiers [9]. Thus in data transformation phase we have to transform or consolidate some data into acceptable forms.

Data mining step is an essential process where intelligent tricks and appropriate algorithm is selected and applied to prepared data in order to extract data patterns.

Knowledge discovery in databases phases divided into four following levels: Defining the Problem, Data Pre-processing, Data Mining and Post Data Mining. Pattern evaluation is in post data mining step and it's typically employs filters and thresholds to discover patterns [10]. Final phase, knowledge presentation, performs when the final data are extracted some techniques visualize and report the obtained knowledge to the users.

2.2 Decision Tree Algorithm

DT represent rules, which can be axiomatic by humans and can be used in knowledge systems such as database, it's a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions [4]. In principle, DTs are used to predict the membership of objects to different categories (classes), taking into account the values that correspond to their attributes (predictor variables). As we have mentioned above, the DT method is one of the main data mining techniques [11].

The DT algorithm is a classification and regression algorithm provided by Microsoft SQL Server Analysis Services (SSAS) for use in predictive modeling of both discrete and continuous attributes. You can see a simple DT in fig. 1 which is a Univariate Tree.

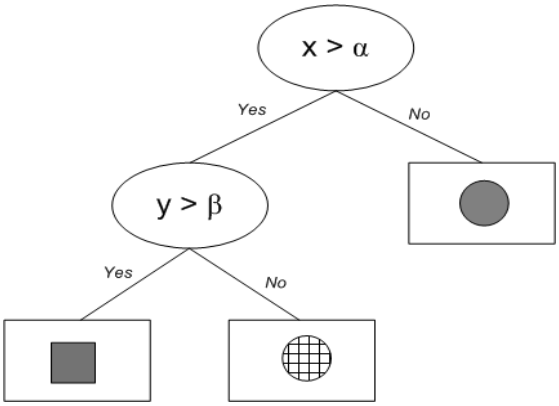

Fig.1. A Univariate DT

We show the results of this DT in fig. 2. that tree has two decision nodes (specify a test on a single attribute) and tree Leaf nodes (indicates the value of target attribute). As you see, the tree classifies instances by starting at the root of the tree and moving through it until a leaf node.

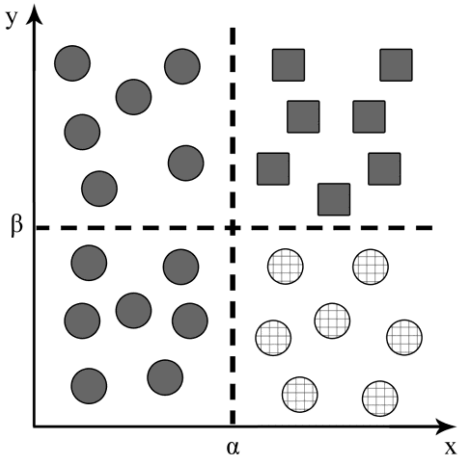

Fig. 2.Result of apply Univariate DT

A research done with combined ANN and DTs model for prognosis of breast cancer relapse in 2002. This presents a decision-support tool for the prognosis of breast cancer relapse using clinical–pathological data. They hope that clinicians will be able to use ANN combined with DTs to search through large datasets seeking subtle patterns in prognostic factors, and that may further assist the selection of appropriate adjuvant treatments for the individual patient [14].

DTs are one of the most popular methods that are used for Data Mining purposes. DTs can be constructed using a variety of methods. For example, C4.5 uses information-theoretic measures [25] and Classification and Regression Trees (CART) [26] uses statistical methods [15]. In one of the last-mentioned ascertainments in current year (2012) S.S. Moon et al. enquired about DT models for characterizing smoking patterns of older adults [16]. The association between college-education and smoking was corroborated as previous study shows [27] that people without college education less likely to try quit smoking compared to the people with college education and negative correlation between education and life time nicotine dependence [28].

Also, Exarchos et al. [17] gain some beneficial information about predicting Parkinson's symptoms. They used partial

DTs to predict Parkinson's symptom, which is a new approach for diagnosis and therapy in patients suffering from Parkinson's disease. They have presented the data miner module of the PERFORM system for developing PD symptom prediction models and for discovering new knowledge in the form of association rules. The accuracy of the symptoms' prediction ranges from 57.1 to 77.4%, depending on the symptom.

Depending on the situation and desired outcome, there are various types of DTs that you can use. They are: Classification Tree, Regression Tree, DT Forests, Classification and Regression Tree, K-Means Clustering. For example Classification Tree would find its application in probability and statistics or Regression Tree is used in calculations for real estate [18]. Some of useful machine algorithms for DTs are CHIAD, QUEST, ID3, and IDA [18, 19].

3. DATA MINING IN HEALTH CARE

Target of data mining is discovering new valid sketches which are traceable by statistical tools and artificial intelligence in large amount of data. In recent years, data mining is so successful in scientific and commercial communities. It's used for tracing people and group's behavior, processing medicine information, supporting services to clients, vindicate concluding and some other acts. For example, principles of data mining concluded from some special discoveries, like the relations between Estrogen and Alzheimer.

The use of statistical techniques for analyzing data is upswing in recent years. Generally, statistical methods used in order to positivity of a precept, thus first we build a theory and then with statistic methods we evaluated its truth. Unlikely data mining don't follow past known patterns to characterize unknown patterns. So it has most benefits in medicine field. Some instances are: perusing affections, dues of drugs, diagnose and predict of most kind of disease, determination medication method, predicting success rate of medicine operations like surgeries, health information system (HIS) analyzing, medical image analyzing and processing.

3.1 Caesarian Section

Caesarian section is the most commonly performed obstetric operation in the world. While the debate continues about the ideal caesarian section rate, safety and cost remain the main areas of concern. Numerous references to caesarian section appear in ancient scripts. However, early history of this procedure is of dubious accuracy. Indeed, the commonly held belief that Julius Caesar himself was delivered this way is unlikely to be true [20]. To enable clear communication between healthcare professionals, four categories of urgency (Table 1) have been recommended by the UK National Confidential Enquiry into Patient Outcome and Death (NCEPOD), and endorsed by the UK Royal College of Obstetricians and Gynecologists (RCOG) and the UK Royal College of Anesthetists (RCA) [21].

Table 1.Categories of emergency for caesarian section

Category	Criterion
1	Immediate threat to the life of the woman or fetus
2	Maternal or fetal compromise that is not immediately life threatening
3	No maternal or fetal compromise but early delivery required
4	Delivery timed to suit woman and staff (elective)

Almost a quarter of deliveries in the UK are performed by caesarian section [22]. The procedure itself has changed little over the years, although evidence based refinements have resulted in reduced morbidity, and further research continues to trying to adapt techniques for safety further improvement [20].

Table 2.Caesarian Section Classification Dataset

No	Age	Delivery NO	Delivery NO	Blood of Pressure	Heart Problem	Caesarian
1	22	1	Timely	High	apt	No
2	26	2	Timely	Normal	apt	Yes
3	26	2	Premature	Normal	apt	No
4	28	1	Timely	High	apt	No
5	22	2	Timely	Normal	apt	Yes
6	26	1	Premature	Low	apt	No
7	27	2	Timely	Normal	apt	No
8	32	3	Timely	Normal	apt	Yes
9	28	2	Timely	Normal	apt	No
10	27	1	Premature	Normal	apt	Yes
11	36	1	Timely	Normal	apt	No
12	33	1	Premature	Low	apt	Yes
13	23	1	Premature	Normal	apt	No
14	20	1	Timely	Normal	inept	No
15	29	1	Latecomer	Low	inept	Yes
16	25	1	Latecomer	Low	apt	No
17	25	1	Timely	Normal	apt	No
18	20	1	Latecomer	High	apt	Yes
19	37	3	Timely	Normal	inept	Yes
20	24	1	Latecomer	Low	inept	Yes
21	26	1	Premature	Normal	apt	No
22	33	2	Timely	Low	inept	Yes

23	25	1	Premature	High	apt	No
24	27	1	Timely	Low	inept	Yes
25	20	1	Timely	High	inept	Yes
26	18	1	Timely	Normal	apt	No
27	18	1	Premature	High	inept	Yes
28	30	1	Timely	Normal	apt	No
29	32	1	Timely	High	inept	Yes
30	26	2	Premature	Normal	inept	No
31	25	1	Timely	Low	apt	No
32	40	1	Timely	Normal	inept	Yes
33	32	2	Timely	High	inept	Yes
34	27	2	Timely	Normal	inept	Yes
35	26	2	Latecomer	Normal	apt	Yes
36	28	3	Timely	High	apt	Yes
37	33	1	Premature	Normal	apt	No
38	31	2	Latecomer	Normal	apt	No
39	31	1	Timely	Normal	apt	No
40	26	1	Latecomer	Low	inept	Yes
41	27	1	Timely	High	inept	Yes
42	19	1	Timely	Normal	apt	Yes
43	36	1	Premature	High	apt	Yes
44	22	1	Timely	Normal	apt	Yes
45	36	4	Timely	High	inept	Yes
46	28	3	Timely	Normal	inept	Yes
47	26	1	Timely	Normal	apt	No
48	32	2	Timely	High	inept	Yes
49	26	2	Latecomer	Normal	apt	No
50	29	2	Timely	Low	inept	Yes
51	33	3	Latecomer	Normal	inept	No
52	21	2	Premature	Low	inept	Yes
53	30	3	Latecomer	High	apt	No
54	35	1	Premature	Low	apt	No
55	29	2	Timely	Normal	inept	Yes
56	25	2	Timely	Normal	apt	No
57	32	3	Premature	Low	inept	Yes
58	21	1	Timely	Low	apt	Yes
59	26	1	Timely	High	apt	Yes
60	30	2	Premature	High	inept	Yes
61	22	1	Latecomer	High	apt	No
62	19	1	Timely	Normal	apt	Yes
63	32	2	Timely	Low	apt	Yes
64	32	2	Timely	Normal	inept	Yes
65	31	1	Latecomer	High	inept	No
66	35	2	Timely	Normal	apt	Yes
67	28	3	Timely	Normal	apt	Yes
68	29	2	Timely	Normal	inept	No
69	25	1	Timely	Low	apt	Yes
70	27	2	Latecomer	Low	apt	No
71	17	1	Timely	Low	apt	Yes
72	29	1	Latecomer	Low	inept	Yes
73	28	2	Timely	Normal	apt	No
74	32	3	Timely	Normal	inept	No
75	38	3	Latecomer	High	inept	Yes
76	27	2	Premature	Normal	apt	No
77	33	4	Timely	Normal	apt	Yes
78	29	2	Premature	High	apt	Yes
79	25	1	Latecomer	Low	apt	Yes
80	24	2	latecomer	Normal	apt	No

One of the main complications of caesarian section is hemorrhage. The average blood loss at caesarian section is between 0.7 and 1.01. This is often underestimated, particularly, when the loss has been large [23].

Classifying caesarian section into ‘Emergency’ and ‘Elective’ is no longer common practice as it does not convey the degree of urgency and in fact the ‘Emergency’ category is too broad as it may include procedures done within minutes to save the life of mother or baby, as well those in which no immediate danger is anticipated [22].

There are many reasons why a health care provider might feel that you need to have a cesarean delivery. Some cesareans occur in critical situations, some are used to prevent critical situations and some are elective. Some of these medical reasons are Placental Abruption, Uterine Rupture, Breech Position, Fetal distress and some other reasons. The mentioned arguments are for an absolute caesarian delivery, if there is one of these cases the choice of caesarian section is approximately necessary.

4. PROPOSED METHOD

Design of DT for pregnant women’s information in health center of Tabriz is studied in this section that composed of five attributes. Among the most important characteristics of delivery problems in the medical field we choose age, number of pregnant, delivery time, blood of pressure and heart status. We classify delivery time to Premature, Timely and Latecomer. As like as delivery time we consider blood of pressure in three statuses of Low, Normal and High moods. Also in this research, we destine existence or absence of heart problem.

To implement DT, C4.5 algorithm was adopted due to it abilities to apply negotiation strategies, diagnose delivery

method in pregnant women and high accuracy in medical applications, and etc. We choose five attributes and then we use DT C4.5 algorithm to discover worth data among all of information saved in patient's records. Now, we exert Weka3.6.6 software in our dataset that we collect of 80 pregnant women information. In Table II, we depicted some of pregnant women's information. The C4.5 algorithm builds DTs using a top-down, greedy search procedure and represents the core of Quinlan's highly successful DT algorithm [24].

5. RESULTS AND DISCUSSION

In our real training dataset, we collected pregnant women's information that referred to delivery in Tabriz health center, and then applied DT C4.5 algorithm to exploit operational and drastic information by genuine data. Now, in two below tables, factors error rate is calculated and computed. Our created DT size is 31 and it has 21 leaves node.

Table 3. Accuracy of Computing on Dataset in C4.5 Tree

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.971	0.217	0.767	0.971	0.857	0.946	No
0.783	0.029	0.973	0.783	0.867	0.946	Yes
0.863	0.109	0.886	0.863	0.863	0.946	-

We simulated all pregnant women's information then the results indictor which we obtain is the fact as was direct relevance between heart failure rate and deliveries those ends to caesarian section. Heart status play key role in our designed system bye DT. Other results indicate that more than 75% women with inept heart status didn't have a natural delivery and over 65% of those have inept heart case savors an abnormal pressure of blood.

Table 4. Evaluation Results on Training Data

Accuracy	Number	Percent
Correctly Classified Instances	69	86.25 %
Incorrectly Classified Instances	11	13.75 %
Kappa statistic	0.7281	-
Mean absolute error	0.1767	-
Root mean squared error	0.2972	-
Relative absolute error	-	36.1264 %
Root relative squared error	-	60.1217 %
Total Number of Instances	80	-

As you see, this decision tree for 86.25% cases predicted correct results. So we can gain these useful information obtained in important decision in determination of medicine operations.

6. CONCLUSION

Development in usage of computer systems and data stores in databases cause to exiting huge databases. In this study, the data mining and its phases are defined and data mining studies which are done in medical and health areas are introduced. Decisions and elections in medicine areas and operations are the most important part of identifying and medicaments of a patient. Determination of a delivery type is so vital because the life of mother and fetus depended to this choice. Then, we collected some factors among pregnant women referred to health center. There was an important relevance between heart failures, blood of pressure and delivery type. This system is predicated more than 86.25% cases is correct. The success of DT is measured by its accuracy, and we can increase this factor with enlargement of dataset and appending some other grave attributes in the future.

7. REFERENCES

[1] F.S. Gharehchopogh, Z.A. Khalifelu, "Application Data Mining Methods for Detection Useful Knowledge in Health Center: A Case Study Using Decision Tree", 2011 International Conference on Computer Applications and Network Security.

[2] A.A. Walter, "Data Mining Industry: Emerging Trends and New Opportunities", Massachusetts Institute of Technology, pp. 13-15, 2000.

[3] Kantardzic, Mehmed. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, 2003.

[4] H. Jiawei and K. Micheline, Data Mining: Concepts and Techniques, vol. 2, Morgan Kaufmann Publishers, 2006.

[5] Christy, T. (1997). Analytical tools help health firms fight fraud. Insurance & Technology, Vol .22(3), pp 2-26.

[6] Biafore, S. (1999). Predictive solutions bring more power to decision makers. Health Management Technology, Vol.20 (10), pp 12- 14.

[7] Indranil Bose, Radha K. Mahapatra, Business data mining — a machine learning perspective, Information & Management, Volume 39, Issue 3, 20 December 2001, Pages 211-225, ISSN 0378-7206, 10.1016/S0378-7206(01)00091-X.

[8] H. Witten and F. Eibe, Data Mining: Practical Machine Learning Tools and Techniques, vol. 2, Diane Cerra Publishers, 2005.

[9] Chia-Ming Wang, Yin-Fu Huang, Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data, Expert Systems with Applications, Volume 36, Issue 3, Part 2, April 2009, Pages 5900-5908, ISSN 0957-4174, 10.1016/j.eswa.2008.07.026.

[10] Zhang, Shichao and Zhang, Chengqi and Yang, Qiang, Data preparation for data mining, Applied Artificial Intelligence, Volume17, Issue 5-6, 2003, Pages 375-381.

[11] Florin Gorunescu, Data Mining: Concepts, Models and Techniques, Intelligent Systems Reference Library, Vol. 12, Springer Publication, 2011.

[12] Riccardo Bellazzi, Blaz Zupan, Predictive data mining in clinical medicine: Current issues and guidelines, International Journal of Medical Informatics, Volume 77, Issue 2, February 2008, pp. 81-97.

[13] Paolo Giudici, Silvia Figini, Applied Data Mining for Business and Industry, A John Wiley and Sons, Ltd., Publication, 2009.

[14] José M. Jerez-Aragonés, José A. Gómez-Ruiz, Gonzalo Ramos-Jiménez, José Muñoz-Pérez, Emilio Alba-Conejo, A combined neural network and decision trees model for prognosis of breast cancer relapse, Artificial Intelligence in Medicine, Volume 27, Issue 1, January 2003, Pages 45-63.

[15] S. Piramuthu, "Input data for decision trees", Expert Systems with Applications", Expert Systems with Applications, Volume 34, Issue 2, 2008, Pages 1220-1226.

[16] Sung Seek Moon, Suk-Young Kang, Weerawat Jitpitaklert, Seoung Bum Kim, Decision tree models for characterizing smoking patterns of older adults, Expert Systems with Applications, Volume 39, Issue 1, January 2012, Pages 445-451.

[17] T.P. Exarchos, A.T. Tzallas, D. Baga, Dimitra Chaloglou, Dimitrios I. Fotiadis, Sofia Tsouli, Maria Diakou, Spyros Konitsiotis, Using partial decision trees to predict Parkinson's symptoms: A new approach for diagnosis and therapy in patients suffering from Parkinson's disease, Computers in Biology and Medicine, Volume 42, Issue 2, February 2012, Pages 195-204.

[18] A. Navada, A. N. Ansari, Overview of Use of Decision Tree algorithms in Machine Learning, Control and System Graduate Research Colloquium(ICSGRC), Shah Alam, IEEE, 2011

[19] Mevlut Ture, F. Tokatli, I. Kurt, Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, and ID3) in determining recurrence-free survival of breast cancer patients, Expert Systems with Applications, Volume 36, Issue 2, Part 1, March 2009, Pages 2017-2026.

[20] A. Simm, P. Ramoutar, Caesarian section: Techniques and complications, Current Obstetrics & Gynaecology, Volume 15, Issue 2, April 2005, Pages 80-86.

[21] Andrew Simm, Darly Mathew, Caesarian section: techniques and complications, Obstetrics, Gynaecology & Reproductive Medicine, Volume 18, Issue 4, April 2008, Pages 93-98.

[22] Elizabeth A. Bonney, Jenny E. Myers, Caesarian section: techniques and complications, Obstetrics, Gynaecology & Reproductive Medicine, Volume 21, Issue 4, April 2011, Pages 97-102.

[23] K.R. Hema, R. Johanson, Caesarian section: techniques and complications, Current Obstetrics & Gynaecology, Volume 12, Issue 2, April 2002, Pages 65-72.

[24] Mevlut Ture, Fusun Tokatli, Imran Kurt, Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, and ID3) in determining recurrence-free survival of breast cancer patients, Expert Systems with Applications, Volume 36, Issue 2, Part 1, March 2009, Pages 2017-2026.

[25] S.L., Salzberg,"C4.5: Programs for Machine Learning", Machine Learning journal, Springer Netherlands, Vol: 16, No: 3, PP: 235-240., 1994.

[26] Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone. 1984. Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books and Software, Pacific Grove, CA. 358 pp.

[27] Molly P. Green, Kristen L. McCausland, Haijun Xiao, Jennifer C. Duke, Donna M. Vallone, and Cheryl G. Healton. A Closer Look at Smoking Among Young Adults: Where Tobacco Control Should Focus Its Attention. American Journal of Public Health: August 2007, Vol. 97, No. 8, pp. 1427-1433.

[28] Mei-Chen Hu, Mark Davies, and Denise B. Kandel. Epidemiology and Correlates of Daily Smoking and Nicotine Dependence among Young Adults in the United States. American Journal of Public Health: February 2006, Vol. 96, No. 2, pp. 299-308.