European Journal of Geography Volume 4, Issue 3: 38-49 October 2013

"DTH 1.0": TOWARDS AN ARTIFICIAL INTELLIGENCE DECISION SUPPORT SYSTEM FOR GEOGRAPHICAL ANALYSIS OF HEALTH DATA

Dimitris KAVROUDAKIS
University of the Aegean, Geography Department, University Hill, Mytilene, Lesvos, 81100, Greece
http://www.geo.aegean.gr/intro_en.htm, dimitrisk@geo.aegean.gr

Phaedon C. KYRIAKIDIS
University of the Aegean, Geography Department, University Hill, Mytilene, Lesvos, 81100, Greece
http://www.geo.aegean.gr/intro_en.htm, phkyriakidis@geo.aegean.gr

Abstract
The complexity of modern scientific research requires advanced approaches to handle and analyse rich and dynamic data. Organizations such as hospitals, hold a great number of health datasets which may consist of many individual records. Artificial Intelligence methodologies incorporate approaches for knowledge retrieval and pattern discovery, which have been proven to be useful for data analysis in various disciplines. Decision trees methods belong to knowledge discovery methodologies and use computational algorithms for the extraction of patterns from data. This work describes the development of an autonomous Decision Support System ("Dth 1.0") for the real-time analysis of health data with the use of decision trees. The proposed system uses a patient's dataset based on the patients' symptoms and other relevant information and prepares reports about the importance of the characteristics that determine the number of patients of a specific disease. This work presents the basic concept of decision trees, describes the design of a tree-based system and uses a virtual database to illustrate the classification of patients in a hypothetical intra-hospital case study.

Keywords: *Decision making, geographical analysis, artificial intelligence, data mining, health geography, decision trees.*

1. INTRODUCTION

The complexity of modern data analysis is constantly increasing as the number of variables involved increases. Modern scientific problems, require even greater computational power to handle and analyse available data in order to produce meaningful outputs for analysis and informed decision making. The more the variables and the characteristics of a problem, the greater the complexity in the association between problem and its characteristics. A modern group of methodologies is Artificial Intelligence (AI) which was introduced during the last decades in computational sciences. This group of methods include Machine Learning (ML) which is a category of methods for the extraction of knowledge from data to "train" a system which will later accumulate knowledge for analysis and prediction. The adaptation of AI methods to problem solving and data analysis is valuable to modern scientists as the complexity of scientific problems increases. Moreover, geographical applications require

advanced tools for spatial analysis. The variety, type and computational intensity involved in spatial data analysis, make AI methods valuable to the modern scientific arsenal of geographical analysis methods.

Data mining is a methodological group of AI which extracts information from data. Often the type, extent and complexity of datasets hides the underlying information and trends which are crucial for scientific analysis. Those trends and associations can be extracted with data mining operations. “Decision trees” is a common modern data mining methodology dealing with training and prediction. This approach predicts the value of a variable by knowing other available attributes and can be applied to variables of categorical nature. For example when information is available about individuals concerning: car ownership, house ownership, income category and age, then by training a decision tree model with data, a prediction model can be build to predict the car ownership (binomial variable) according to the values of the other variables. A very good exposure to the challenges and capabilities of the scientific field of data mining is the work of Witten et. al. [1] which analyses ML and data mining and provides practical recommendations.

Health geography is a field of geography focusing on the spatial characteristics of health related problems. Some of the topics of health geography can be approached by the use of AI methods. For example decision trees can be implemented towards the development of a health decision support system to understand and illustrate possible development of a disease by a patient via the use of patient's health record database. In other words a decision tree can be implemented for understanding the levels of a variable (development of disease) by parsing other data on the system. The underlying mechanism associates habits with a disease and offers a statistical model for the analysis of the development of the disease. This work illustrates the use of decision tree models for the understanding of health related datasets. After the presentation of the theory of decision trees, a model is constructed. Following the generation of a random health dataset (“arth2000”) the model is trained to predict the development of a disease in a sample of 1000 patients. Finally we discuss the potential use and extensions of the proposed decision support system.

2. DECISION TREE CLASSIFICATION

Classification is a training methodology in ML. It assigns class labels to cases based on models linking known class labels with attribute levels. Some of the most common data classification techniques to date are: neural networks [2], [3] , Bayesian networks [4–7] and Support Vector Machines [8–11]. Those approaches have their advantages and disadvantages and are suitable for different types of data analysis. Decision tree methodology is mostly suitable for classifying datasets with nominal variables and exploring relationships between the standardized variations of their attributes. Some of the advantages of this methodology are the fast learning algorithms that can be used such as ID3 [12] and C4.5 [13] and the robustness of the methodology to noise such as missing values and attribute noise. Some of the disadvantages of decision trees are the difficulty to represent the parity of values in a relationship and proportional complexity of the output diagram which sometimes can be misleading if not followed by expert analysis. In other words when data are complex and the output graph has a substantial amount of nodes and leafs, the human eye can be misled. This will not happen if the tree graph is followed by a detailed explanation of each part of the tree. A decision tree accepts a dataset of nominal data and produces a dendrogramatic representation [14] of the data variables according to their levels. The generic algorithm used for such a classification is the following:

1. **A** is the best decision attribute for the next node
2. Assign **A** as decision attribute for the node

3. For each value of **A**, create a new descendant of the node
4. Sort data by leaf nodes
5. Iterate over leaf nodes, until data are classified

The result of this algorithm produces a number of nodes and leafs illustrating the number of potential decisions from the attributes of the data. This categorization of potential decisions groups all possibilities, and counts the most prominent ones. The modelling process splits the data into two subsets: one for learning and the other for prediction. The first subset of the data is used as a trainee for the model and teaches the model to understand the relationship between attributes. The second subset is used for evaluating this knowledge. The error calculation is based on the number of successful predictions of the model. The splitting of the data depends on the number of cases. The error quantification is also associated with the number of cases in the data, the number of variables and the number of levels in the ordinal scale, discretizing the range of variability of continuous explanatory variables.

Decision trees can process data involved in various disciplines and develop a knowledge discovery tool to predict levels of ordinal dependent variables. In the broader scientific field of health sciences there have been some interesting attempts to use decision tree methodologies. The work by Andreescu et.al. [15] illustrates the use of decision trees in the prediction of patients respond to treatment of late-life depression. With a number of 461 records, the authors developed a hierarchy of predictors with decision trees. Additionally, the work of Mann et. al [16] aims of determining the risk for a suicide attempt in psychiatric patients with the analysis of multiple risk factors. Decision trees method has been used in a health dataset of 408 patients with mood schizophrenia or personality disorders to distinguish possible attempters. Another interesting use of decision trees in health sciences is the work by Zhang et.al [17] which is an attempt to demonstrate the effectiveness of one treatment against another with respect to pregnancy in poly-cystic ovary syndrome (PPCOS). That work used a dataset of 445 women who ovulated in response to treatments among a dataset of 626 participants. Decision trees was used to reflect treatment results between types of the syndrome. Furthermore, Koko et. al. [18] described the evaluation of various decision tree methods on problems of orthopaedic fracture data and concluded that there are some limitations on the accuracy of the model and the sensitivity of the decision tree size. Tsien et. al. [19] in their research about classification trees for diagnosing myocardial infraction, concluded that ML methods such as decision trees can be used in medicine for supporting early diagnostic decisions. Jones et. al. [20], illustrate the use of decision trees in the identification of signals of possible drug reactions and concluded that data mining methods, such as decision trees can be a promising tool for identifying new patterns in medical datasets. Dantchev et. al. [21] argue that decision trees are still in experimental stages and remain difficult to apply to clinical practice in psychiatry. Nevertheless in the same report they argue that those tools allow researchers to see epidemiological data from a more generalized perspective and focus on new priorities. Letourneau et. al. [22] focused on decision making techniques for chronic wound care and concluded that decision trees can help decision making by guiding trained personnel through assessment and treatment options. Another notable example of use of decision trees in health sciences can be seen in the work of Alemi and Gustason [23] , who describe some analytical tools that aid decision making such as decision trees, also includes a number of examples with decision making scenarios.

3. USE OF DECISION TREES METHOD

This part describes the use of Decision Trees in the proposed health decision support system. An artificial dataset is used for the training and validation of the method. The dataset consists of five independent variables (city, age, sex, activity, the milk) and one dependent variable

(hypothetical disease **AD**). The cases of the dataset represent visits of individuals to a health care facility (hospital). It is assumed that during each visit the medical personnel examined the individuals and recorded information about those variables. The total cases in the dataset is n=2000 (1000 training, 1000 evaluating) as can be seen in table 1. Table 2 shows the probabilities for each variable's level, used for the generation of the "arth2000" dataset. Those probabilities are also used for the evaluation of the accuracy of the classification. This work assumes that independent variables influence the dependent variable and via the decision tree methodology illustrates the type and amount of this influence. Subsequently, the model will be able to predict the probability of existence of disease AD by processing the levels of the five dependent variables of an individual. The process of prediction can be complex and may also depend on the type and quality of the available dataset. The variables of the data are: the city variable which indicates the area of living (5 levels) of the patient, the age variable is a categorization of the age group of the patient (4 levels). Activity indicates the type of patient's active or passive type of living (2 levels) and milk variable indicates the daily consumed milk units (3 levels).

We prepare a model which uses information gain as a quality measure to populate a dendrogram. Information gain is represented as the entropy value of the data passed from the model. If the data values are new and haven't been processed earlier from the model, then they add to the overall information scheme. This information gain is calculated by counting the number of previous occurrences of the particular combination of data values, in the dataset. For example if the model processes a list of 10 individuals consisting of 9 males and one female, the information gain increases by one when the model is processing the 10^{th} individual because up until the 9^{th} individual the model only knew about the existence of just one sex. After handling the new sex-level (female) the model creates a new category of individuals and assigns all new females to that.

Table 1. Overview of the "arth2000" dataset

City of residence	Age category	Sex	Daily activity	Milk consumption	Suffer from AB disease
Athens	mature	female	high	high	Yes
Mytilini	old	female	average	average	Yes
Mytilini	mature	male	low	low	No
...	...	...	...	...	...
...	...	...	...	...	...

Additionally the model uses no pruning mechanism as this would limit the extent of the tree and because the number of cases in the "arth2000" dataset is limited. When the model processes the individuals, the minimum records per node is set to 5. In other words, the information gain weight should reach 5 individuals before creating a new leaf. This is an empirical value and varies depending on the type of data or the type of analysis and the required complexity for the results.

Table 2. Probabilities of levels for the variables of the "arth2000" dataset

city		**age**		**activity**		**milk**		**AD disease**	
Athens	0.1	middle	0.2	active	0.5	average	0.3	No	0.4
Chios	0.2	old	0.4	passive	0.4	high	0.2	Yes	0.6
Crete	0.2	young	0.1			low	0.5		
Mytilini	0.3	mature	0.3						
Rodos	0.2								

Figure 1 depicts the overall structure of the developed model. Initially, health data are inserted via a csv reader. Those data could be in principle fetched from the examination rooms of a hospital and following the model processing procedure, they can be presented to decision makers in real time. The proposed procedure then splits the database in two parts for cross validation. In other words, the model will learn from the first part of the data and then evaluate the quality of the knowledge upon the second half of the data. The process of learning is held in module number 3 and the linkage of the dataset is taking place in module number 4. The linkage evaluates the leaning ability of the model with the remaining data.

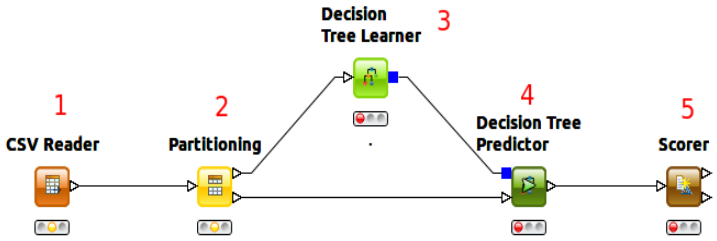

Figure 1. Work-flow process diagram for data mining of the arth2000 dataset

Decision Trees approach, helps solve a problem, which in this case-study is the understanding of possible future existence of disease AB to a number of patients by knowing a limited number of information about each patient. This approach is used to represent the various decision points along the examination of a potential patient. As can be seen in table 3, (Rule 3) according to the arth2000 dataset, if the patient is from Chios city, it has an increased probability to suffer from AB disease. The examination personnel, then needs to ask the patient about his/her age as in Rule 6, patients of mature, old and young age, have an additional increased probability to suffer from disease AB. The resulted decision tree, offers a list of characteristics which have increased probability over the possible existence of AB disease, according to the arth2000 dataset. It describes the logical steps required for determining whether an individual has increased probability to suffer from disease AB (dependent variable) by knowing the value of a number of other variables (independent variables).

The decision tree methodology consists of a root node split by a single variable into partitions. In turn those partitions become nodes to be split further. This divide-and-conquer approach continues until no further splitting would improve the performance of the model. The performance is the ability of the model to understand the possible categorization of a case, based on its attributes. This ability of the model, increases as the statements incorporate

additional knowledge about the training dataset. In other words, the more the information gain from a categorization, the higher the ability of the model to categorize cases with less information. The categorization statements of the produced decision tree are depicted in Table 3.

Table 3. The categorization statements for the "arth2000" dataset

Rule				Appearance of disease AB	cases	percentage	categories
1				No	450/1000	45%	
	2			No	348/791	43%	milk=high,low
		4		No	239/569	42%	city=Athens,Crete,Mytilini, Rodos milk=high,low
		5		No	109/222	49%	city=Athens,Crete,Mytilini, Rodos milk=average
			10	No	51/120	42%	city=Crete,Rodos milk=average
			11	Yes	44/102	43%	milk=average city=Athens,Mytilini
	3			No	102/209	48%	city=Chios
		6		No	83/180	46%	city=Chios age=mature,old,young
		7		Yes	10/29	34%	city=Chios age=middle

For each statement (row) the first column is the id of the statement. The disease AB column indicates that this statement can categorize cases that may suffer from disease AB. The "cover" column shows the number of cases that have been categorized with this rule. The percentage column shows the percentage of the categorized cases with respect to the total number of cases. Finally, the "categories" column shows the information gain from each rule. For example rule 11 indicates that cases from Athens or Mytilene with average consumption of milk, have 43% probability to suffer from disease AB. The combination of those rules can categorize this particular "arth2000" dataset with 100% accuracy.

In order for the model to judge how good a potential split (node-leaf) is, the information gain rule is used, which creates a new split at the attribute with the highest information gain. This approach creates new splits only when they will create concrete partitions of the dataset. The split function strategy for this model is the entropy reduction strategy. The greater the information from a categorization, the greater the knowledge of the model for future categorizations. Finally the minimum number of observations in a node before attempting a split (splitting factor) is 100 cases for 1000 cases. This reduces the tree complexity and produces a more readable representation. The selection of the splitting factor depends on the number of total cases and the amount of detail required in the results. The more complex the results, the more the leafs of the decision tree.

The very same categorization process can be used with any other categorical dataset to prepare categorization rules. This makes the approach generalizable and flexible. A factor that determines the accuracy of the results is the number of levels for each variable. Binomial variables can be categorized with less rules than variables with 4 or 5 levels. This is making sense if we consider that entropy increases for variables with larger number of levels. This forces the model to produce more rules in order to fully categorize a dataset. On the other hand, binomial variables can be easily categorized and require less rules. Finally, a fully

randomized dataset, where there is no relationship between variables, may require a great number of rules to be fully categorized. The tendency of the dataset towards a random distribution is directly associated with the inability of the decision tree to categorize all cases with less rules.

The methodology of data categorization discussed, can be potentially useful to health facilities, such as hospitals for categorizing patient records and present statistics based on the patient profile. For example a large hospital with a great number of daily visits, can produce a good amount of data related to patient characteristics and health problems. Those data can be obtained from a patient upon his/her arrival and stored in a database. Then, a centralized computational decision support system can process this dataset and prepare correlations between patient's characteristics. A decision tree can be implemented in R statistical language, with the use of the "*part*" library [24], to prepare a statistical description for such a database. The administration of a hospital may access the results of the process and use visualizations for decision support. The results may indicate trends and patters that are not initially visible and can offer a centralized and categorized view over the characteristics of each disease.

4. AUTOMATED HEALTH DECISION SYSTEM

The proposed system makes decisions based upon health related datasets and evaluates the possibilities of occurrence of a control variable. The application of the proposed system is the statistical analysis of health databases from non-expert uses such as managers and directors of hospitals. This analysis can also be helpful in early examination procedures or in decision making in epistemological analysis.

Considering the lack of centralized statistical tool in the Greek Health Sector, this work is innovative because it describes and proposes the use of such a tool and advocates the use of not only descriptive statistics but also the use of AI for advanced statistical analysis. Sometimes it is difficult to fully understand the big picture of a cause-effect relationship especially when it is hidden deep into a great amount of data. The proposed system is using data mining methods and databases to construct a "hospital-oriented" computer system that will prepare on the fly data statistics. More specifically, this work describes the design and the details of a proposed system that can be installed in a health facility such as hospitals, and prepare on-the-fly statistics about the patient's database. As illustrated in figure 2 the flow of information starts from the examination room (point 1) where the basic description of the characteristics and the medical record are transformed into a digital record (point 2) and inserted into a database (point 3). This process can be facilitated by modern palmtops with live html forms that can either create a new medical record or update an existing one with new medical information. The database of the system can retrieve data with queries. Instead for the researchers/managers (point 5) to use traditional SQL queries to retrieve tabular results, they can use a series of predefined actions that will call a number of queries. The pre-defined actions will present the results in graphical form (point 4). One of those predefined actions may include the use of a decision tree to analyse the data and observe the relationship between variables.

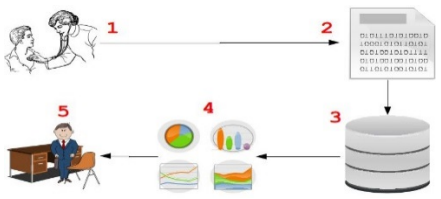

Figure 2. The abstract design of the proposed decision system

This proposed system requires a number of "examination-room computers" which will be used for data entry during examination. The software required for those machines, should be just basic intra-net browsing and an html compatible client (web browser). The main database will be hosted on an in-house server which will be an average computer with an installation of a database (PostGreSQL, MySql). In this computer the freely available statistical environment R is required which will prepare statistical analysis, and graphs. Finally the managerial and administrative staff will need average computers with basic intra-net browsing for selection of predefined actions and visualization of results.

The preparation of predefined actions saves time and effort and provides a relatively error-free environment for the generation of statistics. The following code snippet (code 1) shows an example of a predefined action which is triggered by a managerial staff to prepare a description of the age categories of the patients in the database. The following code snippet (code 1) is written in the R statistical programming language and its output is depicted in figure 3.

- Lines 1-3 import the required libraries and packages for the analysis.
- Lines 5 and 6 activate the database driver and connect to the database.
- Lines 8 and 11 discover the available tables in the database and list its fields.
- Lines 14 and 15 retrieve all available data from the database table
- Lines 17-20 convert the data according to age categories
- Lines 22-24 prepare the diagram variables
- Lines 26-28 generate the bar-plot diagram

The generated bar-plot show the absolute number of patients and the percentage for each age group. This predefined statistical analysis generates output that informs the user (managerial and admin staff) about the age groups of the patients that visit the hospital.

```
1 library("RSQLite")
2 library("DBI")
3 library("tools")
4
5 drv <- dbDriver("SQLite")
6 db <- dbConnect(drv, "data.sqlite")
7
8 print(dbListTables(db))
9 #arth2000
10
11 print(dbListFields(db, "arth2000"))
12 #"id" "activity" "milk" "age" "city" "arth"
13
14 ola <- dbSendQuery(db, "select * from arth2000")
15 d <- fetch(ola, n=2000)
16
17 yo <- length(subset(d, d$age=="young")[[1]])
18 ma <- length(subset(d, d$age=="mature")[[1]])
19 ol <- length(subset(d, d$age=="old")[[1]])
20 mi <- length(subset(d, d$age=="middle")[[1]])
21
22 counts <- c(yo,ma,ol,mi)
23 per <- (counts/sum(counts))*100
24 names <- c("young", "Mature", "Old", "Middle")
25
26 mp <- barplot(counts, names.arg=names)
27 text(mp, counts + 50, format(paste(per,"%")), xpd = TRUE)
28 text(mp, counts - 50, format(counts), xpd = TRUE, col = "blue")
```

Code 1. Snippet of R code for the generation of bar-plot with the age structure of the patients. The generated plot is figure 3.

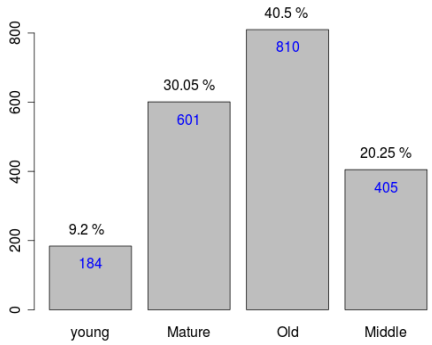

Figure 3. Result of an example predefined action showing the age groups of patients in the dataset

Another predefined statistical "action" could be the profiling of the dataset with decision trees methodology. This methodological approach may serve as an information categorization process that generates a dendrogram of groups of cases. Code 2 depicts the R code of such a process which may be used in the proposed computational system.

```
df=createData(2000)
library("rpart")
tree <- rpart(arth ~ city + activity + age + milk, method="class",
              data=df,control=rpart.control(minsplit=100,cp=0.001))
summary(tree)
plot(tree)
```

Code 2. Snippet of R code for the generation of sample data and the subsequent construction of a decision tree from them.

After the installation and calibration of this autonomous health decision support system, the users can use it without any knowledge on statistics or artificial intelligence. The predefined actions will run code snippets that will prepare various statistical analysis. It is an autonomous system as it uses only the predefined statistical actions and does not require any input by the user. This enables the end user to focus only on a number of important and error-free statistical analyses. Of course as with other similar systems, the quality of the imported data will influence the quality of the exported statistics.

5. GEOGRAPHICAL ANALYSIS OF HEALTH DATA

The data that can be collected form the heath care facilities, include place of residence of the patient. This is important as can provide spatial attributes to health records for further analysis. The health data that have spatial records can be presented in maps and analysed by area in order to provide a better insight on the spatial distribution of health events along with other information about the patient such as proximity to health facilities, socio-economic characteristics etc. Those type of information could be used for geographical analysis of a disease outspread and map the areas which have significant amount of incidents. The understanding of spatial characteristics of a disease outspread can help decision makers to provide better health services and information to the general public for the protection of public health. Geographical epidemiological studies aim to understand the spatial characteristics of health data and formulate hypotheses regarding the spatial causes and effects of a disease [25]. Some of the different branches of spatial epidemiology are disease mapping, cluster identification and spatial socio-economic analysis of a disease[26]. Understanding the greater spatial trends of a disease as well as mapping the spatial distribution of a disease from health records of a hospital, can be a challenging task especially

due to privacy and ethical reasons. Nevertheless, if handled with care, health data records that have spatial attributes can be very useful in early warning epidemiology systems and provide a different approach on understanding the causes and effects of a disease outspread in an area. Automated systems of identifying underlying trends such as decision trees, provide the necessary lay of data mining, which extracts knowledge from individual records. The use of such an autonomous system can spare financial and other resources by avoiding the manual examination of data records and by quickly indicating the geographical characteristics of an event.

6. DISCUSSION AND CONCLUSIONS

The automated system discussed in this work may provide a basic mechanism for real-time analysis of health related data. The statistical analysis of databases may not substitute the actual medical diagnosis from specialized personnel in any case. Nevertheless it is a first step towards the automation of diagnosis and a valuable tool for decision makers and researchers to understand and identify data patterns.

The arth2000 dataset is but an arbitrary dataset consisting of just 2000 cases. No real data have been used in this stage of the research as there is lack of free health database for Greece. In a future state of this research, there is the possibility to use anonymized data from Greek Health Public System. One of the benefits of using decision tree methodology is that by scaling up the number of the cases and using a real medical dataset consisting of more than 10000 cases, the accuracy of the model will increase, and the model will be able to learn faster. Additionally, the geographical analysis of the model results can be facilitated by analysing the effects of disease AB by city which is the second determinant factor of disease AB. This may indicate that disease AB is more common in some geographical areas than others. This may be the basis for a more advanced spatial analysis and may lead to more sophisticated results, such as the identification of the reasons and effects of disease AB on other characteristics of daily life of individuals in those areas. Another interesting point is that this analysis can be automated. This means that an intelligent system can be constructed with very basic hardware such as a home-range personal computer which accepts data from a hospital and prepares instantly (in real time) statistical reports about the profile of the patients. Finally, This proposed system may be used for research and decision support processes and act as a data exploration tool.

With the use of ML methods such as decision trees, one can better reveal information which is hidden in data. Complex relationships that may exist in very large datasets are sometimes difficult to understand and may require a great number of computations and cross-tabulations. Analysis of rich and complex datasets can be valuable to the health sector as it may reveal underlying geographical patterns for diseases, symptoms and characteristics of patients. Finally, decision trees in general may act as a framework to consider the probability of events and *pay-offs* of decisions in various data analyses, not only in health-related sectors but also from in sectors such as geography marketing and logistics.

7. SUMMARY

- What was already known to the topic
  - Artificial intelligence techniques can give a very good insight in medical data and provide good understanding in related variables
  - Health geography is a scientific area which under good knowledge of the underlying mechanisms, can provide meaningful information about geographical distribution of health-related data

- Decision trees is an artificial intelligence approach which has been successfully used in a number of scientific areas and provides a good understanding on the relationship of variables in a dataset.

- What this study added to our knowledge
  - Computerised systems with the incorporation of AI, may give medical personnel a very good understanding regarding the topic in research
  - Decision trees can be used in such systems in order to facilitate the understanding of the relationship between patients characteristics
  - Hospitals in Greek health sector do not yet have expert centralized computer systems for aggregate depiction of patients data.
  - A novel expert system in Greek health Sector may give a very good understanding on patients data and interconnect patient's characteristics such as geographical area of residence.

REFERENCES

[1] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Morgan Kaufmann Pub, 2011.
[2] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1994.
[3] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*. PWS Boston, MA, 1996.
[4] D. B. Rubin, "The bayesian bootstrap," *The Annals of Statistics*, pp. 130–134, 1981.
[5] F. V. Jensen, *An Introduction to Bayesian Networks*, vol. 210. UCL press London, 1996.
[6] F. V. Jensen and T. D. Nielsen, *Bayesian networks and decision graphs*. Springer Verlag, 2007.
[7] A. Darwiche and E. Corporation, *Modeling and Reasoning with Bayesian Networks*. Citeseer, 2009.
[8] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines*. Cambridge University Press, 2000.
[9] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 2002.
[10] A. W. Moore, "Support Vector Machines," *http://jmvidal. cse. sc. edu/csce883/svm14. pdf*, vol. 31, p. 2001, 2007.
[11] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer Verlag, 2008.
[12] J. Cheng, U. M. Fayyad, K. B. Irani, and Z. Qian, "Improved decision trees: A generalized version of ID3," in *Proceedings of the Fifth International Conference on Machine Learning: June 12-14, 1988, University of Michigan, Ann Arbor*, 1988, p. 100.
[13] J. R. Quinlan, *C4.5: Programs For Machine Learning*. Morgan Kaufmann, 1993.
[14] J. B. Phipps, "Dendrogram topology," *Systematic Biology*, vol. 20, no. 3, p. 306, 1971.
[15] C. Andreescu, B. H. Mulsant, P. R. Houck, E. M. Whyte, S. Mazumdar, A. Y. Dombrovski, B. G. Pollock, and C. F. Reynolds, "Empirically Derived Decision Trees for the Treatment of Late-Life Depression," *Am J Psychiatry*, vol. 165, no. 7, pp. 855–862, 2008.
[16] J. J. Mann, S. P. Ellis, C. M. Waternaux, L. XINHUA, M. A. Oquendo, K. M. Malone, B. S. Brodsky, G. L. Haas, and D. Currier, "Classification trees distinguish suicide attempters in major psychiatric disorders: a model of clinical decision making," *The Journal of clinical psychiatry*, vol. 69, no. 1, pp. 23–31, 2008.

[17] H. Zhang, R. S. Legro, J. Zhang, L. Zhang, X. Chen, H. Huang, P. R. Casson, W. D. Schlaff, M. P. Diamond, S. A. Krawetz, C. Coutifaris, R. G. Brzyski, G. M. Christman, N. Santoro, and E. Eisenberg, "Decision trees for identifying predictors of treatment effectiveness in clinical trials and its application to ovulation in a study of women with polycystic ovary syndrome," *Human Reproduction*, vol. 25, no. 10, pp. 2612 –2621, 2010.

[18] P. Kokol, M. Zorman, M. Stiglic, and I. Malcic, "The limitations of decision trees and automatic learning in real world medical decision making," in *Proc. 9th World Congr. Med. Inform. (MEDINFO-98)*, 1998, vol. 52, pp. 529–533.

[19] C. L. Tsien, H. S. Fraser, W. J. Long, and R. L. Kennedy, "Using classification tree and logistic regression methods to diagnose myocardial infarction," *Stud Health Technol Inform*, vol. 52 Pt 1, pp. 493–497, 1998.

[20] J. K. Jones, "The role of data mining technology in the identification of signals of possible adverse drug reactions: value and limitations," *Current Therapeutic Research*, vol. 62, no. 9, pp. 664–672, Sep. 2001.

[21] N. Dantchev, "Decision trees in psychiatric therapy," *Encephale*, vol. 22, no. 3, pp. 205–214, 1996.

[22] S. Letourneau and L. Jensen, "Impact of a decision tree on chronic wound care," *J Wound Ostomy Continence Nurs*, vol. 25, no. 5, pp. 240–247, 1998.

[23] F. Alemi and D. H. Gustafson, *Decision Analysis for Healthcare Managers*. Health Administration Press, 2006.

[24] T. M. Therneau, B. Atkinson, and B. Ripley, "Rpart: recursive partitioning," *R package version*, vol. 3, pp. 1–23, 2005.

[25] M. Rezaeian, G. Dunn, S. S. Leger, and L. Appleby, "Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary," *J Epidemiol Community Health*, vol. 61, no. 2, pp. 98–102, Feb. 2007.

[26] P. Elliott and D. Wartenberg, "Spatial epidemiology: current approaches and future challenges," *Environmental health perspectives*, vol. 112, no. 9, p. 998, 2004.

