Using Machine Learning Algorithms in Cardiovascular Disease Risk Evaluation

A.V. SITAR-TĂUT[1], D. ZDRENGHEA[1], D. POP[1], D.A. SITAR-TĂUT[2]

[1] Cardiology-Rehabilitation Department, University of Medicine and Pharmacy "Iuliu Haţieganu", Cluj-Napoca, Romania

[2] Faculty of Economics and Business Administration, "Babeş-Bolyai" University, Cluj-Napoca, Romania

***Abstract* – Even if Medicine and Computer Science seem apparently intangible domains, they collaborate each other for few decades. One of the faces of this cooperation is Data Mining, a relative new and multidisciplinary field capable to extract valuable information from large sets of data. Despite this fact, in cardiology related studies it was rarely used. We assume that some data mining tools can be used as a substitute for some complex, expensive, uncomfortable, time consuming, and sometimes dangerous medical examinations. This paper aims to show that cardiovascular diseases may be predicted by classical risk factors analyzed and processed in a "non-invasive" way.**

***Keywords*: cardiovascular disease, machine learning algorithms**

I. BACKGROUND

Diversity of information made that useful data processing and acquisition to become very ample processes, this being the main cause for appearing and developing of "data mining" concept. As we know, data mining represents an analytical process that explore a very large data sets seeking for new patterns and relationships between variables, generalizing this relationships in a new model, formula, or decision tree. It is capable to discover new knowledge without previous hypothesis, the goal being to discover new, unexpected, unintuitive knowledge [1][2], analyzing data from different point of views and summarize them in new and useful information. Data mining has become a tool for improving the classical statistical tools used in future tendency's prediction [3]. There have already been some tries to use this tool in medicine (E.g. Herron's study that explores the usefulness of data mining in discrimination between benign and malign tumors [4]).

Machine learning is a subfield of artificial intelligence. Its field aims to design and develop algorithms that allow computers to improve their performance over time based on data [5][6]. To learn, the machine must analyze its past experience in order to find useful regularities, patterns, even ones that a human might miss [7][8]. A major focus of machine learning research is to automatically produce (induce) models, such as rules and patterns, from data. Machine learning is much related to fields such as "data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science" [9] and data warehouses [6]. Machine learning is inspired from the animal and human learning patterns, but now it is able teaching us how the animals and human being learn. How the machines do learn and why do we need this is the big question? Nilsson says that a "machine learns whenever it changes its structure, program or data" – as a reaction to or from its environment – "in such a manner that its future performance improves". These changes can be insertions or updates in a database. We need a machine that learns because some facts are heavy to be explained and offering an example is a better opportunity; the machines must to adapt to some environments or environment variations, which were not predicted at the time of its design or development; large amount of data may contain hidden relationships that humans may not discover due their limited storing and computing capabilities; etc [10].

Weka (Waikato Environment for Knowledge Analysis) is a cross-platform open source, and probably the most popular machine learning based application. It was written in Java and developed by the University of Waikato [11]. We will use version 3-6-0 in this material.

A decision tree is a tool supporting decision process. It uses a tree-like graph or model of decisions and their possible consequences. These trees are used in decision analysis, to help identifying the proper strategy to reach a goal, or they depict the way how to calculate conditional probabilities for complex processes.

In data mining and machine learning fields, a decision tree acts as a predictive model. Here, the tree-based model names are classification tree, for discrete outcome, or regression tree for continuous outcome. In their constructions, the leaves represent classifications and branches are conjunctions of facts or characteristics that lead to such classifications. The machine learning technique used to generate a decision tree from data is called decision tree learning [12].

For the moment, cardiovascular diseases represent the first mortality cause in women [13][14][15][16][17], in Europe approximately 55 percent of women's deaths being

caused by cardiovascular diseases (especially coronary disease and stroke) [15]. Number of deaths caused by cardiovascular diseases is greater that the aggregated next seven causes [14], and for the moment, cardiovascular diseases kill more women than men [13]. Framingham study has revealed the impact of smoking, hypertension, dyslipidaemia, diabetes mellitus, obesity, male gender, and age on developing of cardiovascular disease. Since 2004, the guidelines had revealed the importance of cardiovascular risk factors' recognition in women, but also the importance of classifying women according to cardiovascular risk (low, medium, and high) [13][15].

In our country, we do not have enough data for an objective characterization of women from cardiovascular point of view. But we already know that, in women, the cardiovascular mortality and morbidity are some of the highest from Europe.

In clinical research, the tests are used in order to establish the presence or absence of some diseases [18]. Thus, for confirming or infirming the presence of cardiovascular disease (expressed as coronary heart disease, stroke, or peripheral artery disease), the patients are submitted to different tests (biochemical tests, rest ECG, stress test, echocardiography or angiography). Some of them are invasive, uncomfortable for patients, and mainly expensive and time consuming. Also, their sensitivity and specificity differ from 100 percent (to a certain measure, depending by test type).

Data mining techniques are able to identify the high risk patients, to define the most important variables in cardiovascular patients, but, in the same time, they have the capacity to build a model in order to distinguish, in a simple and understandable way, the relationships between any two variables [6]. The purpose of present study was to compare the capacity of different data mining methods to evaluate and quantify the relationships between cardiovascular risk factors and cardiovascular disease, differently by the gender of patients.

Present study (MENOCARD) was conducted to an 825 people sample, from Cluj county. The data has been collected from general practitioner's files, for every patient being registered the blood pressure, if he/she is hypertensive (blood values more than 140/90 mm Hg, or used of antihypertensive drugs); glycemia, if he/she is diabetic; body mass index (BMI); serum lipids fractions (a patient being considered as dyslipidaemic if total cholesterol is greater than 200 mg/dl or if seric triglycerides are greater than 150 mg/dl); smoking status; presence or absence of cardiovascular disease (defined as coronary artery disease – CAD –, stroke or peripheral artery disease – PAD).

Data mining procedures has been supported by Weka. As Lee showed, an important problem of data mining application is represented by the fact that they are functioning just with "clean databases", and missing values represent a crucial problem for a data mining soft [6] and for the quality of the output.

At the beginning, the database included 145 attributes and 825 instances. Initially, database's cleaning up process has been performed for previous statistical processing. But this was not enough for our current purposes. The irrelevant attributes have been eliminated, just 10 being considered significant for the present study. After an evaluation of these attributes from relevance's point of view (using Evaluator: InfoGainAttributeEval, Ranker method), no more eliminations needed. Instances containing missing data have also been eliminated, from initially 825 patients (instances) comprised in the study, just 303 (36.73%) of them (145 males, 158 females) have been kept, being included in data mining processing.

Data mining techniques (Naïve Bayes) and decision trees (J48) have been applied on our database. Mean age of included subjects was 72.85 ± 6.26 years of age (values between 56 and 96 years). No significant differences were registered regarding mean age between two genders (72.59±6.86 years in male vs. 73.09±5.67 in women, p=NS).

The percentages of patients in which cardiovascular factors are present are synthesized in Table 1.

Table 1 – Comparative cardiovascular risk factors by gender

	Women No (%)	***Men No (%)***
Smoking	12 (7.6)	29(20)
Obesity/overweight	169 (69)	87(60)
Diabetes mellitus	39 (24.7)	28(19.3)
Dyslipidaemia	89(56.3)	60(41.4)
Hypertension	149 (94.3)	128(88.3)

Using InfoGainAttributeEval, Ranker method, we evaluated the cardiovascular risk factors' importance in presence/absence of cardiovascular disease (data being presented in Table 2).

Table 2 – Ranking by disease attribute evaluation process result

RANKING	***CAD***	***AVC***	***PAD***
Age	1	4	4
Diabetes (DM)	2	1	5
Obesity/ overweight	3	6	2
Hypertension (HBP)	4	7	6
Smoke	5	5	1
Gender	6	3	3
Dyslipidaemia	7	2	7

Within the entire sample, no significant differences have been recorded between the two methods regarding the capacity to realize a right prediction of coronary heart disease, stroke or peripheral artery disease based on presence /absence of risk factors.

Naïve Bayes correctly classified 62.03% of instances regarding presence/absence of coronary heart disease,

79.21% regarding presence/absence of stroke, respectively 94.06% regarding presence/absence of peripheral artery disease. The results obtained by J48 were similar, the registered percentages being 60.40%, 79.87%, and 94.06% respectively.

Going further with the analyze, we had observed that if for the patients with coronary heart disease, Naïve Bayes has succeed to capture 80.44% from relevant information, the percentages registered for those with stroke and peripheral artery disease are incomparable lower (being equal to zero in both cases).

On the other side, for those without strokes or peripheral artery disease, relevant information was captured in 99.17% and 100%, respectively, of the cases.

Practically, we can say that models realized with Naïve Bayes had provided medium results regarding identification of patients with coronary artery disease (F-measure=0.715) and very good results in identification of patients without stroke (F-measure=0.884) or without peripheral artery disease (F-measure=0.969).

J48 succeeded to capture 72.6% of relevant information in patients with coronary artery disease (as it is showed in figure 1) but unfortunately, it was also incapable to capture relevant information for those with strokes or peripheral artery disease (percentages being also equals to zero). For those ones without strokes or peripheral artery disease, J48 delivered similar results as Naïve Bayes, relevant information being correctly classified in 100%, respectively 99.6% of the cases.

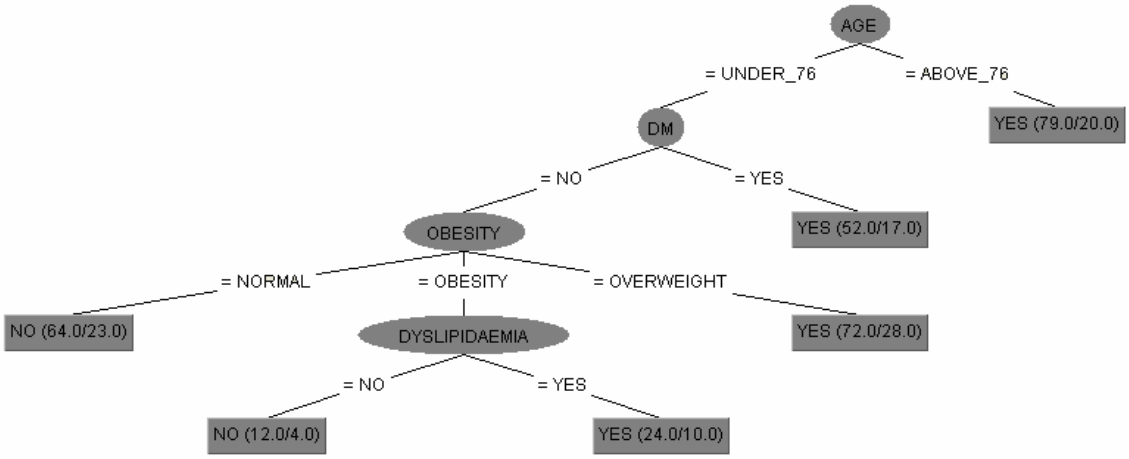

Figure 1 – Decision tree classifying patients with coronary artery disease

As in the previous model achieved with Naïve Bayes, good results have been obtained for those with ischemic disease (F-measure=0.668), but also for those without strokes (F-measure=0.888) or peripheral artery disease (F-measure=0.968). Thus, we can say that both methods had succeeded to predict, in a mean percentage, the patients with coronary artery disease, but they did not succeed to identify those ones with strokes or peripheral artery disease. In the same time, on behalf of presence/absence of cardiovascular risk factors, both methods have succeeded to identify the ones without strokes or peripheral artery disease.

Going further, we tried to see if the gender of the patients can be involved in capacity of two methods to classify correctly the instances. The obtained results have been similar with those obtained for the entire sample, no significant differences being registered between two methods regarding correct classification of patients, no matter of patient's gender. The tendencies to capture approximately 70 percent of relevant information in those with coronary artery disease have been maintained, but, similar with previous analyses, no relevant information has been captured in those with strokes or peripheral artery disease.

Comparing the instances in both genders, we observed that capacity of capture relevant information in ischemic patient was higher in women, no matter of used method (Naïve Bayse – in women 79.6% vs. 74.1% in men-, J48 in women 89.8% vs. 71.6% in men). In the same time, even the percentages are not very high, in women it has been discovered a higher capacity to capture relevant information in patients with stokes (Naïve Bayes – 3.7%, J48 0% in women, 0% for both methods in men) or peripheral artery disease (Naïve Bayes – 16.7%, J48 – 50% in women, but 0% for both methods in men).

In both genders, relevant information has been captured in 90 percent in those having neither stroke nor peripheral artery disease.

II. DISCUSSION

Although modern IT&C tools have been applied in medical field in the last few years, the number of

researches is still low. In cardiology the use of data mining is very poor, almost equal to zero.

It is well-known the importance of cardiac risk factors in cardiovascular disease's development. According to our knowledge, no studies focused on comparative importance of risk factors in developing a cardiovascular disease, have been made before. The only possible way to hierarch cardiovascular risk factors is to use relative risks determined by the presence or absence of respective risk factor. In the same time, the studies found in literature are various, using different methodology and different end-points.

Data mining methods, applied on our sample, have permitted to distinguish that if in developing ischemic disease the most important risk factors are represented by age, diabetes mellitus and obesity, this fact is not true for strokes or peripheral artery disease. Therefore, for stroke the major risk factors seem to be diabetes mellitus and dyslipidaemia, while for peripheral artery disease, the smoking and obesity seem to be the most important risk factors. On the other hand, we have to emphasis that the patients' gender presents a greater importance in development of strokes and peripheral artery disease. Our study reveals that practically no important differences have been found between Naïve Bayes and J48 according the capacity to identify cardiovascular disease patients. We have to remark that for the patients with coronary artery disease the two methods are able to capture the relevant information. On the other side, neither one has been capable to extract relevant information for stroke of peripheral artery disease patients, but are capable to present relevant information for those without these diseases. Why? Probably because the effect of risk factors is more than just a simple summation of the individual parts.

The limit of the study is represented by the small number of accurate instances.

III. Conclusions

The relationship between cardiovascular risk factors and cardiovascular disease is not linear, being necessary other studies in order to test new artificial intelligence methods, to be assessed more patients and more cardiovascular risk factors.

In order to fulfill these desiderates, we intend developing following researches – supported by e-Procord grant – these studies being projected to be addressed to diverse people category (according to the age, gender, pathology, with classical or non-classical risk factors).

Acknowledgement

MENOCARD – CEEX Grant no 98/2006, "Optimization of degenerative cardiovascular diseases' treatment in postmenopausal women" and "**e-ProCord** – New Medical and Modeling Approaches in IT&C Age Applied on Cardiovascular Profile Evaluation at Molecular Level. Differences Implied by Gender, Age, and Existing Pathology", PN II Program, IDEI, ID_2246/2009 CNCSIS Code.

References

[1] D.-A. Sitar-Taut, "Baze de date distribuite", Risoprint Publishing House, pp291-295, 2005.
[2] V.P. Bresfelean, "Implicații ale tehnologiilor informatice asupra managementului instituțiilor universitare", Risoprint Publishing House, pp 127-170, 2008.
[3] P. Andreeva, "Data Modelling and Specific Rule Generation via Data Mining Techniques", International Conference on Computer Systems and Technologies - CompSysTech' 2006.
[4] P. Herron, "Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning Classification Algorithms", Data Mining. Spring 2004.
[5] L. Zhang, W.M. Kim Roddis, "Machine Learning in Updating Predictive Models of Planning and Scheduling Transportation Projects", Transportation Research Record, TRB, No. 1588, pp. 86-94, 1997 .
[6] I.-N. Lee, S.-C. Liao and M. Embrechts, "Data mining techniques applied to medical information" . Med. inform. vol. 25, no. 2, pp81- 102, 2000.
[7] http://www.cs.kuleuven.ac.be/~dtai/ml/research
[8] I. Harris, J. Denzinger and D. Yergens,"Application of the Weka Machine Learning Library to Hospital Ward Occupancy Problems", Canada Technical Report 2007fi884fi36. December 12, 2007.
[9] http://en.wikipedia.org/wiki/Machine_learning
[10] N.J. Nilsson, "Introduction to Machine Learning", http://www.wepapers.com/Papers/12236/INTRODUCTION_to_machine_learning
[11] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005
[12] http://en.wikipedia.org/wiki/Decision_Tree
[13] L. Mosca, C.L Banka, E.J. Benjamin, K. Berra, C. Bushnell, R.J. Dolor et al, for the Expert Panel/Writing Group,"Evidence-Based Guidelines for Cardiovascular Disease Prevention in Women: 2007 Update" JACC (49):1230–1250, 2007.
[14] F. Alfonso, J. Bermejo, J. Segovia, "Cardiovascular disease in women. Why now?" Review, Rev Esp Cardiol; 59(3):259-263. March 2006.
[15] M. Stramba-Badiale, K.M. Fox, S.G. Priori, P. Collins, C. Daly, I. Graham, et al, "Cardiovascular diseases in women: a statement from the policy conference of the European Society of Cardiology", Eur Heart J (27); 994-1005, 2006.
[16] N.K. Wenger, L.J. Shaw, V. Vaccarino, "Coronary heart disease in women: update 2008", Clin Pharmacol Ther; 83 (1):37-51, 2008.
[17] L. Pilote, K. Dasgupta, V. Guru, K. Humphries, J. McGrath et al. „A comprehensive view of sex-specific issues related to cardiovascular disease", CMAJ; 176(6) S1-41, 2007.
[18] M.A. Turkman, "Predictive tools in the assessment of diagnostic tests", Third Workshop on Statistics, Mathematics and Computation First Portuguese-Polish Workshop on Biometry. Lisbon, 21-22 July Universidade Aberta – PORTUGAL, 2008. www.univ-ab.pt/wemc2008