A Methodology Based on Business Intelligence for the Development of Predictive Applications in Self-Adapting Environments

Marcelo Fabio Roldan
Universidad Nacional de La Rioja
La Rioja, Argentina
Marceloroldan@unlar.edu.ar

Joyati Debnath
Winona State University
jdebnath@winona.edu

Ana Funes, Germán Montejano, Daniel Riesco
Universidad Nacional de San Luis
San Luis, Argentina
{afunes, gmonte, driesco}@unsl.edu.ar

Abstract—**Based on different methodologies, we have developed a new methodology that synthesizes, accelerates and optimizes the application of Adaptive Business Intelligence (ABI) in medicine, genomics, pharmaceuticals and other related sciences. This methodology -in its final phase-culminates in a stage called "Predictive Software Development", where the model obtained may be used to generate a prediction oriented application that can be implemented in different programming languages. The work presents the methodology and also the steps followed in the development of a predictive application for self-adapting environments for a particular case study of hypothyroidism.**

Keywords-Methodology; Business Intelligence; Predictive Application; Self-Adapting

I. Introduction

In recent years there has been a great increase in generating and collecting data. This is due mainly to the processing power of computers and the low cost of storage.

Within these huge volumes of data there is information, which is strategically important and that is not normally accessible by classical techniques of information retrieval.

The discovery of this "hidden" information is made possible through data mining, which provides a set of techniques to find patterns and relationships in data.

This allows the creation of models, i.e. abstract representations of reality, as part of the Knowledge Discovery in Database Process (KDDP, for its acronym in English). As part of this process -among other things- is the data preparation and interpretation of the results giving meaning to the patterns found.

On the other hand, it is well known that the application of Business Intelligence [9], [12] provides benefits in business areas. This decision support system is enhanced when it is complemented with self-adaptive applications. Adaptive Business Intelligence (ABI) [10] is the discipline that combines prediction, optimization, and adaptability in a system capable of answering two fundamental questions: What is likely to occur in the future? And what is the best decision at this time?

To implement a technology in a business, a methodology is required. These methodologies are defined on the gained experience and the best of the most successful and popular procedures. Having a methodology, it has become as important and necessary as a letter of presentation for companies. With this premise in mind, we present here a new methodology to develop an Adaptive Business Intelligence system from the study of multiple profiles, specifically related to life sciences, which are far apart from the area of business, the usual field of Business Intelligence.

We have used a procedural synthesis to obtain the resulting methodology. The stages identified and here proposed cover the requirements of the applications. These stages have resulted from an analysis of a number of well-known methodologies: SEMMA [3], CRISP-DM [1], [3], [4], [6] and KDD Process [5], [7].

CRISP-DM consists of a cycle that includes six phases in terms of a process of hierarchical model, including a set of tasks described in six levels of abstraction (from the general to the specific): phase, generic task, specialized task and process instances. The sequence of the phases in its life cycle is not rigid, moving back and forth between different phases as needed [1].

The acronym SEMMA [2] (Sample, Explore, Modify, Model, Assess), refers to the process of conducting a data mining project. The SAS Institute considers a process cycle involving five stages: Sample, Explore, Modify, Model, and Assess. Sample: This stage consists of sampling the data, where a portion of the data set is taken to provide significant information, but it is small enough to manipulate it quickly. Explore is an exploration of the data for trends and unwanted anomalies. Modify facilitates the modification of the data by creating, selecting and transforming the variables for the modeling process. Model: This stage involves the application of various data mining techniques for finding a combination that perform a desired level of prediction. Assess: This stage consists of assessing data, evaluating the usefulness and reliability of the model found during the data mining process, as well as estimating its efficiency.

With respect to the KDD Process, it includes -as a constituent part- the use of data mining techniques. Given

Published by Atlantis Press, Paris, France.

that data mining is one of the most important stages in this process, usually both concepts are taken as synonyms. However, there are clear differences between both, being KDD a process consisting of a series of phases in which data mining is one of them. KDD involves other tasks for the successful completion of projects aimed to obtain new knowledge.

Besides presenting the formulation of a new methodology, which has been developed on the basis of the main steps for solving problems in Adaptive Business Intelligence and the methodologies described above, we also illustrate the applicability of this methodology through a case study of hypothyroidism. In particular, the problems of prediction in medicine have become a challenge for knowledge extraction from data, therefore a methodology based on Adaptive Business Intelligence will provide a set of solutions based on various methods and techniques (data mining, prediction, optimization, and adaptability), which will allow the extraction of knowledge not only in medicine but also in other disciplines.

The rest of the paper is organized as follows. Section II presents the proposed methodology that emerges from the analysis of a number of known methodologies. In Section II.B we give a brief description of the steps that form the proposed methodology, after presenting in Section II.A the complete life cycle of the methodology. In Section III we apply our methodology to a case study for the prediction of hypothyroidism. Finally, Section IV closes the paper with some conclusions and future work.

II. The ABI Methodology

The methodological steps for the development of applications based on Adaptive Business Intelligence include problem understanding, data preparation, modelling, approaching the objectives and implementation of a business application. To do this we can use different methodologies, such as CRISP-DM, SEMMA and KDD Process [3].

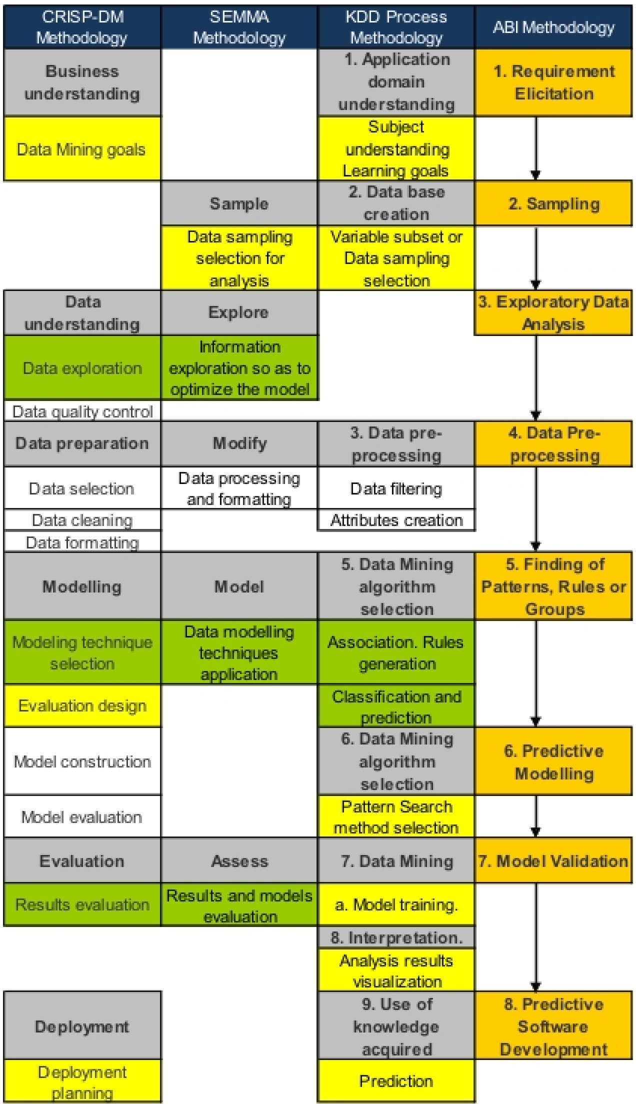
We think that a simple methodology -agile and effective-should make the necessary contributions in order to help the stakeholders (data miner, leader, expert, etc.) in their various phases, allowing a dynamic interaction -through the different phases- with the patterns that emerge from the data and that have significance as new knowledge.

The definition of this new methodology, which focuses and summarizes the virtues of the methods mentioned above, facilitates the production of knowledge through learning and discovery [5], especially when it comes to those areas of science such as medical, pharmaceutical or similar that can be conceptually considered far apart from the spirit of Adaptive Business Intelligence.

We could observe, through different applications, the presence of distinct-perceived stages. From a conceptual analysis and applying the criteria of three mature and accepted methodologies such as the CRISP-DM, SEMMA [2] and the KDD Process [7], we could clarify which the procedural or methodological steps necessary for our objective are. We evaluated exhaustively the different features corresponding to each stage for each of the different methodologies and we extracted the steps that are common, adding -at the same time- all the new steps we considered necessary.

The name for each stage has been assigned according to the emerging role in the methodological process. As a result of this work the scheme shown in TABLE I. came out. We can see, on the right column, the identified stages of our methodology, and the correspondence with the stages in each of the compared methodologies. These stages are briefly described in Section II.B, but first the complete life cycle of the methodology is presented in Section II.A.

TABLE I. STAGES OF THE PROPOSED METHODOLOGY



CRISP-DM Methodology	SEMMA Methodology	KDD Process Methodology	ABI Methodology
Business understanding		1. Application domain understanding	1. Requirement Elicitation
Data Mining goals		Subject understanding Learning goals	
	Sample	2. Data base creation	2. Sampling
	Data sampling selection for analysis	Variable subset or Data sampling selection	
Data understanding	Explore		3. Exploratory Data Analysis
Data exploration	Information exploration so as to optimize the model		
Data quality control			
Data preparation	Modify	3. Data pre-processing	4. Data Pre-processing
Data selection	Data processing and formatting	Data filtering	
Data cleaning		Attributes creation	
Data formatting			
Modelling	Model	5. Data Mining algorithm selection	5. Finding of Patterns, Rules or Groups
Modeling technique selection	Data modelling techniques application	Association. Rules generation	
Evaluation design		Classification and prediction	
Model construction		6. Data Mining algorithm selection	6. Predictive Modelling
Model evaluation		Pattern Search method selection	
Evaluation	Assess	7. Data Mining	7. Model Validation
Results evaluation	Results and models evaluation	a. Model training.	
		8. Interpretation.	
		Analysis results visualization	
Deployment		9. Use of knowledge acquired	8. Predictive Software Development
Deployment planning		Prediction	

a. Yellow indicates a task that has some degree of incidence in the developed methodology, green that the incidence is high and no background color that it has little significance.

A. Life Cycle

As in the different methodologies we have studied and compared, the succession of phases in our methodology is not rigid. Each phase is structured in several second-tier general tasks. The general tasks are projected to specific tasks, which ultimately describe the actions to be developed for particular situations; however we do not give prescriptions about how they must be performed.

The life cycle in Figure 1. shows synthetically those aspects that describe the phases to follow for the realization of a deployment model, useful for the development of a predictive application. These phases have been defined aiming to follow a coherent and progressive development that allows arriving –for a given problem– to an acceptable solution.

Figure 1. also shows the correspondence between the steps of our methodology and the different phases in the life cycle. These steps indicate, in turn, a path to be followed by the analyst to find hidden patterns in data, exposing the activities that have to be achieved sequentially during the process of developing adaptive applications based on Business Intelligence.

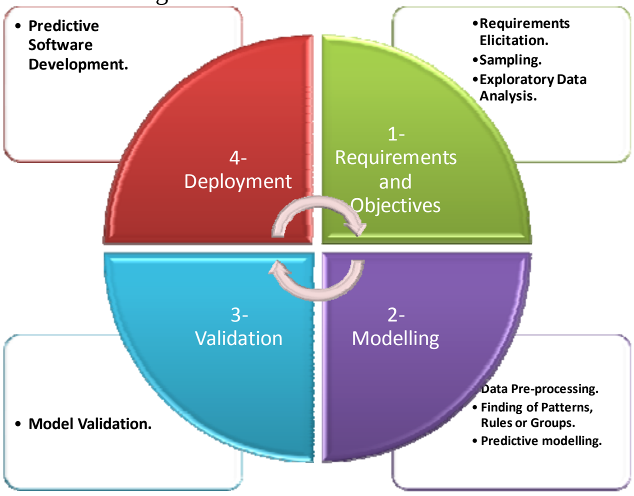

• Predictive Software Development.
•Requirements Elicitation.
•Sampling.
•Exploratory Data Analysis.
4- Deployment
1- Requirements and Objectives
3- Validation
2- Modelling
• Model Validation.
Data Pre-processing.
• Finding of Patterns, Rules or Groups.
• Predictive modelling.

Figure 1. Life cycle of the proposed ABI methodology and its steps.

As we can see in Figure 1. , the life cycle iterates on its different phases and over the identified steps, starting at the phase 1 referred as Requirements and Objectives.

Note that the deployment phase could appear as separated from the rest to represent the time in which it is possible to construct the software. However, given the self-adaptive nature of the applications emerging from the use of this methodology, this step is also part of the cycle as this is the way the application optimize its prediction, reaching so the aim of Adaptive Business Intelligence:." as Michaelewicsz has indicated "Adaptive Business Intelligence systems include elements of data mining, predictive modeling, forecasting, optimization, and adaptability, and are used ... to make better decisions [10].

In this context, the contribution of the data miner as the main analyst and the expert in the knowledge area will provide the new sources that support the maintenance of self-adaptive system as a whole.

It's worth noting that it is possible to implement a predictive application through specific modules, such as Business Intelligence does [12]. These modules can incorporate capabilities of cost-impact analysis, cost and confusion matrices, hypothesis testing techniques, Boosting, Bagging, Randomization, and other sophisticated technologies that are not part of this work.

Besides that, the structural logic of the methodology considers as well the feature of self-adaptability as it is shown in Figure 2.

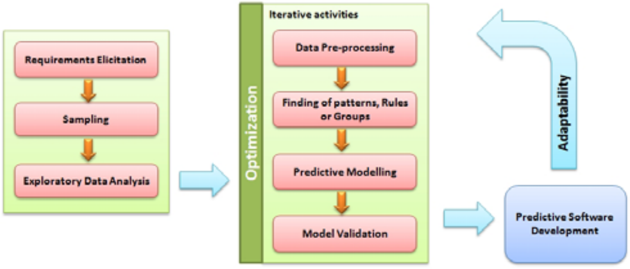


Figure 2. Implementation of optimization, prediction and adaptability in the methodology

B. *Stages of the ABI Methodology*

In this section we summarize the scope of each of the identified stages in the proposed ABI methodology shown in Figure 1.

1)Requirements Elicitation. As in any software process, we work closely with the domain experts to define the problem and determine the objectives of the project. These project objectives are translated into hypotheses about the initial selection of the data mining techniques to be used later in the process that will take place.

2)Sampling. It includes collecting sample data and deciding what data, including format and size will be needed. It checks the data integrity, redundancy, missing values, the plausibility of attribute values, etc. In addition, the stage includes verification of usefulness of the data with respect to the goals of data mining.

3)Exploratory Data Analysis. Since such systems are driven by data, it is important to have a good understanding of them. The aim is to identify the most important fields related to the problem and determine what values derived may be useful.

Stages 2 and 3 may be considered complementary and can work together in a corrective cycle of improvement.

4)Data Pre-processing. This stage is where data cleaning is done, which includes checking the integrity of the data records, dealing with noise and missing values, removals and duplications, format adjustments, adding numeric expressions and balanced formulas that improve the quality of significant or relevant missing data. At this stage, data may be subjected to supervised or unsupervised filters reducing data that are considered non-related to project objectives.

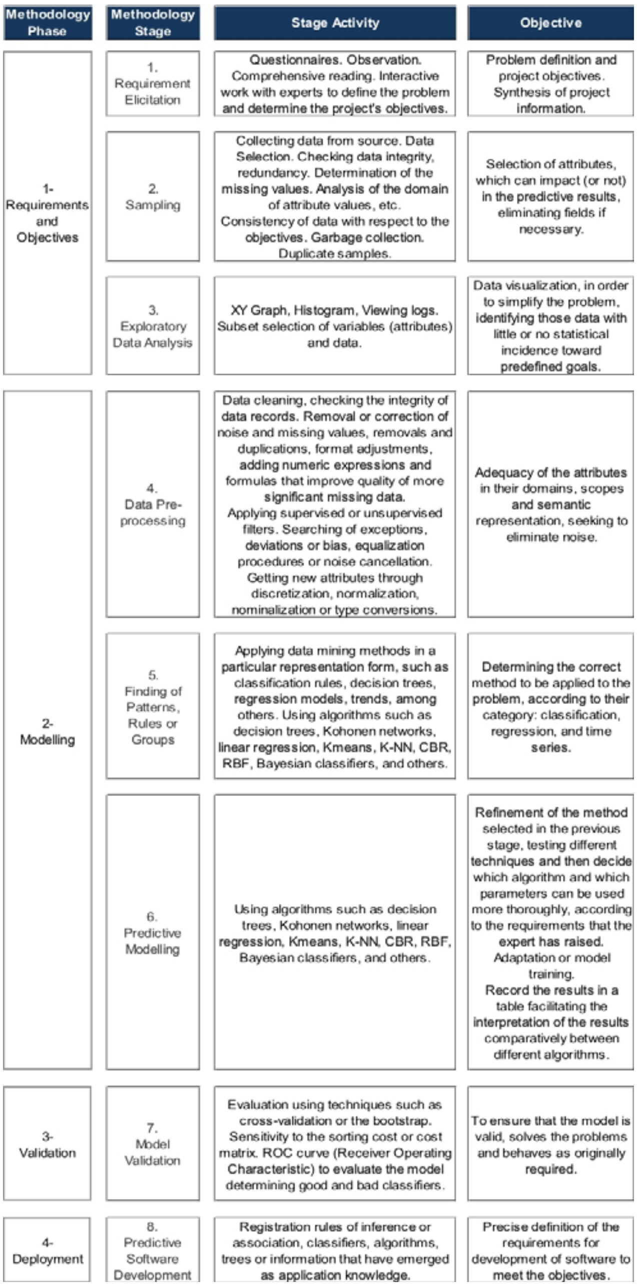
5)Finding of Patterns, Rules or Groups. At this stage it is intended that, through different methods of data mining, the results match or come close to the objectives defined in stage 1. This is where we choose the data mining technique that allows finding patterns in data.

6)Predictive Modeling. It is a refinement of the method selected in the previous stage, testing different techniques and then deciding which algorithm and which parameters can be used, according to the requirements posed by the expert.

7)Model Validation. It includes understanding the results, checking whether the discovered knowledge is novel and interesting. This interpretation of the results by experts in the field is significant when trying to verify the impact of discovered knowledge. Due to its iterative nature, it is possible to replicate the earlier stages using other methods and data mining algorithms, identifying which alternative actions can be taken to improve results.

8)Predictive Software Development. Once the model is obtained, it must be deployed. This means leaving the development environment in a form that can be used by external software.

TABLE II. CORRESPONDENCE BETWEEN PHASES, METHODOLOGICAL STAGES AND ACTIVITIES



Methodology Phase	Methodology Stage	Stage Activity	Objective
1- Requirements and Objectives	1. Requirement Elicitation	Questionnaires. Observation. Comprehensive reading. Interactive work with experts to define the problem and determine the project's objectives.	Problem definition and project objectives. Synthesis of project information.
	2. Sampling	Collecting data from source. Data Selection. Checking data integrity, redundancy. Determination of the missing values. Analysis of the domain of attribute values, etc. Consistency of data with respect to the objectives. Garbage collection. Duplicate samples.	Selection of attributes, which can impact (or not) in the predictive results, eliminating fields if necessary.
	3. Exploratory Data Analysis	XY Graph, Histogram, Viewing logs. Subset selection of variables (attributes) and data.	Data visualization, in order to simplify the problem, identifying those data with little or no statistical incidence toward predefined goals.
2- Modelling	4. Data Pre-processing	Data cleaning, checking the integrity of data records. Removal or correction of noise and missing values, removals and duplications, format adjustments, adding numeric expressions and formulas that improve quality of more significant missing data. Applying supervised or unsupervised filters. Searching of exceptions, deviations or bias, equalization procedures or noise cancellation. Getting new attributes through discretization, normalization, nominalization or type conversions.	Adequacy of the attributes in their domains, scopes and semantic representation, seeking to eliminate noise.
	5. Finding of Patterns, Rules or Groups	Applying data mining methods in a particular representation form, such as classification rules, decision trees, regression models, trends, among others. Using algorithms such as decision trees, Kohonen networks, linear regression, Kmeans, K-NN, CBR, RBF, Bayesian classifiers, and others.	Determining the correct method to be applied to the problem, according to their category: classification, regression, and time series.
	6. Predictive Modelling	Using algorithms such as decision trees, Kohonen networks, linear regression, Kmeans, K-NN, CBR, RBF, Bayesian classifiers, and others.	Refinement of the method selected in the previous stage, testing different techniques and then decide which algorithm and which parameters can be used more thoroughly, according to the requirements that the expert has raised. Adaptation or model training. Record the results in a table facilitating the interpretation of the results comparatively between different algorithms.
3- Validation	7. Model Validation	Evaluation using techniques such as cross-validation or the bootstrap. Sensitivity to the sorting cost or cost matrix. ROC curve (Receiver Operating Characteristic) to evaluate the model determining good and bad classifiers.	To ensure that the model is valid, solves the problems and behaves as originally required.
4- Deployment	8. Predictive Software Development	Registration rules of inference or association, classifiers, algorithms, trees or information that have emerged as application knowledge.	Precise definition of the requirements for development of software to meet the objectives.

This final stage implies planning how to use the discovered knowledge if necessary. It is presented as a final stage of this methodology, but as an initial stage of a software development process, since it raises the requirements of a new software product, which uses the rules of inference, binders, trees or information that have emerged as knowledge of the application developed by the methodological steps so far described.

TABLE II. shows the correspondence between life cycle phases, methodological stages and performed activities.

III. APPLICATION OF THE PROPOSED METHODOLOGY FOR THE PREDICTION OF HYPOTHYROIDISM CASES

The data used come from a health center -the Garavan Institute and J. Ross Quinlan Sydney in Australia [11]. They correspond to patients classified as non suffering hypothyroidism, suffering primary hypothyroidism, secondary hypothyroidism and compensated hypothyroidism. For space considerations, we present a summary of the different stages. We have used Weka [8] as data mining tool.

Stage 1. A model with a mean absolute error lesser or equal to 2% is wanted.

Stage 2. We worked with a database with 3,772 records. It has 30 attributes, 7 of them have numerical values (1 integer and 6 reals), and the rest are nominal (discrete).

Stage 3. The classes are showed in the TABLE III.

TABLE III. DISTRIBUTIONS BY CLASS

Class	Number of samples
Negative	3481
Compensated hypothyroidism	194
Primary hypothyroidism	95
Secondary hypothyroidism	2

Stage 4. From the analysis of the data distribution, we could observe the possibility of selecting a subset of data, based on some correlations estimated prior the modeling. We determined that from the 30 initial attributes, only 5 were the most representative.

Stage 5. We applied two different tree classification algorithms, SimpleCart and J48. From the application of SimpleCart we got 99.54% of correctly classified instances, while in J48 3,756 instances were correctly classified (99.57%). When we applied the same algorithms to the selected set of attributes we found in the previous stage SimpleCart yielded 97.37% versus 96.81% of J48, getting the classification tree shown in Figure 3.

Stage 6. During the process of model training, we used cross-validation of 10 folds with J48 algorithm which yielded 3,652 samples correctly classified (96.81%).

Stage 7. The model evaluation was performed using the area under the ROC curve (Relative Operating Characteristic). The area under the ROC curve obtained was 0.9838.

Stage 8. At this stage we defined the requirement features to be met by the software that implements the model learn through the previous steps. The discovered rules were expressed as follows (note that hypothyroidism is negative in case no rule applies). See Figure 4.

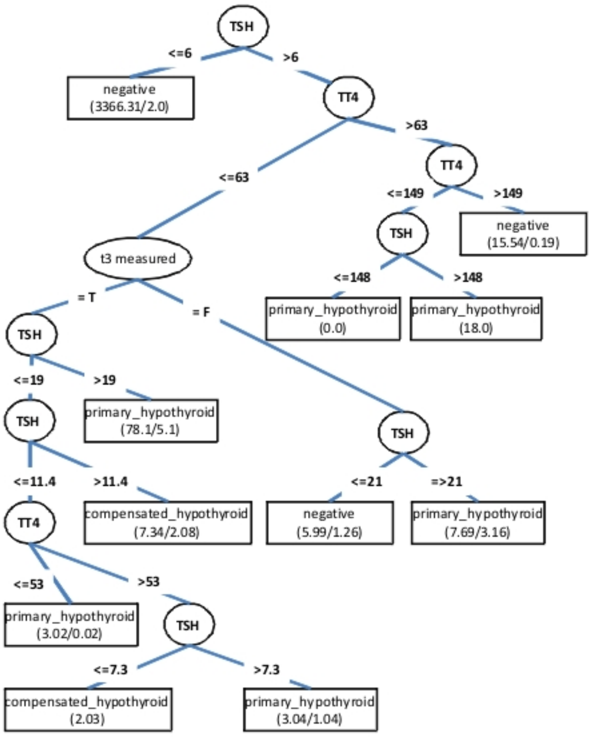
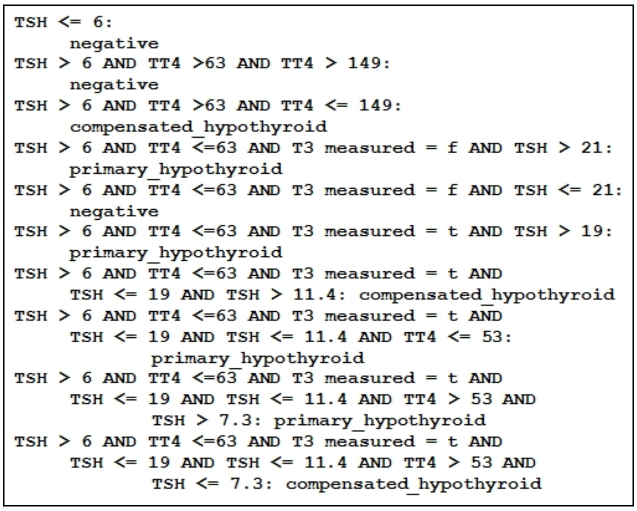


Figure 3. Classification tree obtained in stage 5 of the proposed methodology.



```
TSH <= 6:
      negative
TSH > 6 AND TT4 >63 AND TT4 > 149:
      negative
TSH > 6 AND TT4 >63 AND TT4 <= 149:
      compensated_hypothyroid
TSH > 6 AND TT4 <=63 AND T3 measured = f AND TSH > 21:
      primary_hypothyroid
TSH > 6 AND TT4 <=63 AND T3 measured = f AND TSH <= 21:
      negative
TSH > 6 AND TT4 <=63 AND T3 measured = t AND TSH > 19:
      primary_hypothyroid
TSH > 6 AND TT4 <=63 AND T3 measured = t AND
      TSH <= 19 AND TSH > 11.4: compensated_hypothyroid
TSH > 6 AND TT4 <=63 AND T3 measured = t AND
      TSH <= 19 AND TSH <= 11.4 AND TT4 <= 53:
               primary_hypothyroid
TSH > 6 AND TT4 <=63 AND T3 measured = t AND
      TSH <= 19 AND TSH <= 11.4 AND TT4 > 53 AND
               TSH > 7.3: primary_hypothyroid
TSH > 6 AND TT4 <=63 AND T3 measured = t AND
      TSH <= 19 AND TSH <= 11.4 AND TT4 > 53 AND
               TSH <= 7.3: compensated_hypothyroid
```

Figure 4. The discovered rules in the methodological process

IV. CONCLUSIONS AND FUTURE WORK

We have reviewed different extant methodologies and from this review we have devised a methodology that can be applied in medical areas.

By applying the new methodology to a case study we have found that in a simple but effective way we were able to achieve a satisfying knowledge extraction to be useful in the application of Adaptive Business Intelligence.

As future work we envisage:

- Applying the methodology in the biochemistry area, particularly in an ongoing laboratory renal research where a great amount of data is generated.
- Expand an existing data bank related to hospital-induced sepsis so as to be able to determine its causes.
- Determination of the neural centers involved in states of fear, stress, risk using EEG. This can also be used to update an avatar to represent emotions in virtual learning environments.
- Analyze the causes affecting student desertion.

REFERENCES

[1] Arancibia J. G.: Metodología para el Desarrollo de Proyectos en Mineria de Datos CRISP-DM. http://yoshibauco.wordpress.com/ (2011)

[2] SAS Institute. United Kingdom. http://www.sas.com/ - Último acceso: (2011)

[3] Azevedo A., Santos M.F.: KDD, SEMMA y CRISP-DM: A Parallel Overview. IADIS European Conference Data Mining 2008. Part of MCCSIS (2008)

[4] Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., Wirth R.: CRISP-DM 1.0 Step-by-step data mining guide, CRISP-DM consortium (1999, 2000).

[5] Cios K. J., Pedrycz W., Swiniarski R.W., Kurgan L.A.: Data Mining. A Knowledge Discovery Approach. Springer. (2007)

[6] Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., Wirth R. : Guía paso a paso de Minería de Datos. (2007)

[7] Fayyad U., Piatetsky-Shapiro G., Smith P., Uthurusamy R.: From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining. pp. 1-29. California: AAAI Press / The MIT Press. (1996)

[8] Hall M., Frank E., Holmes G., Pfahringer B.: The WEKA Data Mining Software: An Update. Pentaho Corporation. (2009)

[9] Watson H. J., Wixom B. H.: The Current State of Business Intelligence. Computer Magazine, Vol. 9 Issue 40, Page(s): 96 – 99. (2007)

[10] Michalewicz Z., Schmidt M., Michaelewicz M., Chiriac C.: Adaptative Business Intelligence. Springer (2007)

[11] Machine Learning Data Set Repository. MLdata. http://mldata.org/repository/data/viewslug/datasets-uci-hepatitis/ - Ultimo acceso 2012.

[12] Moss L. T., Atre S. : Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Addison Wesley. (2003)