Framework for Data Model to Personalized Health Systems

Edwar Javier Girón Buitrón
Universidad del Cauca
Popayán, Cauca
+57 314 859 2503
edwardgb@unicauca.edu.co

Carolina Rico Olarte
Universidad del Cauca
Popayán, Cauca
+57 300 524 2526
carolinarico@unicauca.edu.co

Gineth Magaly Cerón Ríos
Universidad del Cauca
Popayán, Cauca
+57 313 754 6873
gceron@unicauca.edu.co

Diego Mauricio López Gutiérrez
Universidad del Cauca
Popayán, Cauca
+57 301 581 9362
dmlopez@unicauca.edu.co

ABSTRACT
When large amounts of data is handled, it is important to obtain the desired compatibility between such data to perform activities of access and storage of information; data models are a tool that helps to determine the structure of the information, in order to improve communication and accuracy in applications that use and exchange data with each other for a common purpose. Nowadays, there is no framework for health supporting the data modeling design, i.e. the existing models are generic and therefore are not suitable to support personalized systems and they do not consider the quality of clinical and personal data, required in health care. Based on the CRISP-DM methodology, a framework is proposed to design a data model for personalized health systems. This framework ensures the security of personal and clinical data to relate it with health standards, particularly with the Personal Health (PHR) ISO/TR 14292 standard, which addresses the recommendations of the parameters that must be within a personalized health system. To perform accurate recommendations it is important to make a data mining process, where the data is related to guarantee an accurate and reliable personalization; these relations generated by the model should be taken into account to apply them a data mining technique.

CCS Concepts
• Computing methodologies→ Modeling and simulation-Model development and analysis• Information systems→ Information systems applications-Data mining.

Keywords
Data model, personalized system in health, data mining, PHR.

1. INTRODUCTION
The ICT (Information and Communication Technologies) seek to promote healthy habits and lifestyles influencing in a positive way on the health of people, so are decreases the risk factors affecting the health of people [1]. The large amount of data that is stored on the cloud related to the ICT interventions and its users becomes a problematic issue, since it is important to discover through data mining, some useful, unexpected and understandable models and data patterns [2], to produce knowledge, which will be used in order to benefit the user.

For this reason, it is consider necessary to count with personalized systems to promote healthy habits and lifestyles based on a user model capable of collecting, managing and relating the personal information of each user in order to know the key aspects of that person and make a proper preprocessing of the data generated by the user according to a data model (its conceptual level); besides the user model must be according to a personal health record standard, since it is necessary to manage the clinical information of a person.

Having in mind the previous considerations, this paper describes a framework for a data model in health, where the user model used to generate the data model (conceptual) is in accordance with the ISO/TR 14292 standard [3] and is implemented in a personalized system through a data mining technique, all of this based in the CRISP-DM methodology for the preprocessing of information to support the promotion of healthy habits and lifestyles.

This paper is structured as follows: the next section illustrates the state of art to the topic at hand. Section 3 describes the methodology used in the research. Section 4 presents the description of the framework for a data model in health and Section 5 illustrates the results obtained by the implementation of the framework through a test with Weka. Finally, the paper presents the conclusions.

2. STATE OF ART
A user model allows a system to respond efficiently to the characteristics that represent knowledge and a preference of the user the system assumes that he/she has [4]. As a basic principle of operation, personalized systems requires acquiring user information and use it to make inferences on the abstract characteristics that are used in defining the behavior of the system to the user [5].

From a literature review are listed the following papers: [6] highlights the state of art of translational bioinformatics to design supported decisions and case-based reasoning systems and presents the design of a tele-health system, which is capable of combining text mining, search literature and case-based retrieval. On other side are the web based systems connected with an EMR, [7] presents a methodology for ontology engineering to generate case base in the medical domain using a based case reasoning system for a case study of diabetes diagnosis and finally [8] that shows a platform with an agent-based three-layer architecture supported on a multi-agent system for home care.

[9] and [10], where the first one describes a personalized system using advanced information algorithms, text and data mining and

978-1-5090-2435-3/16/$31.00 ©2016 IEEE

other computational techniques to support health decisions for patients with diabetes, and the second one presents an approach to medical knowledge recommendations based on collaboration.

In [11] an algorithm for classifying personal health records is introduced and it is designed a personalized search engine matching PHR systems to search the web user so he/she can get answers really relevant for him/her and his/her health status. It is important to highlight the work done by Meyer in [12], since this work comprises concepts covered in this article: presents a personalized system based on a user model that is eHealth oriented to a standard of PHR to promote healthy habits and lifestyles. However, it differs in the way the information is obtained to create the user model, because it comes from his/her context and not from the inherent characteristics from his/her personality, where a data analysis technique is used to infer such contextual characteristics.

3. METHODOLOGY DESCRIPTION

For the development of this article, the Engineering Research Methodology [13] was used. Following this methodology through its stages: 1)It is obtained a conceptual basis through a literature review. 2)The Delphi methodology [14] is used for the selection of items from the user model based on the recommendations of the ISO/TR 14292 standard to design the data model. 3)For system development, the user-centered design (UCD) methodology [15] is used, and the CRISP-DM methodology [16] describes the data mining tasks, in order to distribute the information management and concretize appropriate data mining models to the context this project. 4)To make the evaluation, the DESMET methodology is used to ensure the quality of the development system.

1. At last, it is important to conclude, socialize and present the obtained results.

Figure 1. CRISP-DM reference model phases [16]

In figure 1, the CRISP-DM methodology reference model is observed. This methodology was followed in the next way: the Business Understanding and the Data Understanding phases are used to understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives; and to start collecting data, then get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses about hidden information. The conceptual data model level from the proposed framework is created using these two phases.

The logical data model level is based on the Data Preparation phase that includes all activities required to construct the data set from the initial raw data. Tasks include table, case, and attribute selection as well as transformation and cleaning of data for modeling tools. Following the methodology, the Modeling phase that selects and applies a variety of modelling techniques, and calibrate tool parameters to optimal values [17], is used to create the physical data model level from the framework that is going to be describe in Section 4.

4. FRAMEWORK DESCRIPTION

A framework is a standardized set of concepts, practices and criteria to focus on a particular type of problem that serves as a reference, to confront and solve new problems of a similar nature [18]. Thus, a framework is created for a data model based on a user model according to the ISO/TR 14292 standard for a personalized system in health; the data model consists of three levels as shown in figure 2.

Figure 2. Data model levels

4.1 Conceptual Data Model

Describes the semantics of a domain, being the scope of the model. A conceptual schema specifies the types of facts or propositions that can be expressed using the model [19].

According to what was stated in [4], it starts to make the characterization of the user model from a generic profile, a psychological profile and other features. Through the methodology of user-centered design, several characteristic elements classified are considered. Using the Delphi methodology, comparison charts (x vs. y) are created to define preliminarily the relations between each of the characteristic elements considered.

ISO/TR 14292 standard should be characterized for the user model conformation. In the first characterization, topics covered in this technical report, which makes a classification of personal health records according to six dimensions classified within the topics: PHR structure, order and parameters, security, communication and architecture. The items obtained from the characterization of the standard, are 29 resulting items, then this items are categorized as presented in Fig. 3.

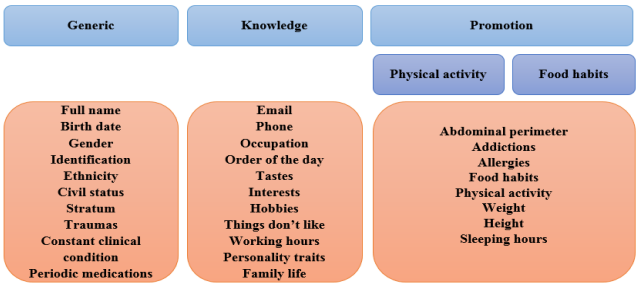

Figure 3. Items resulting from the process in accordance with ISO

4.2 Logical Data Model

Describes the data in as much detail as possible, without regard to how they will be physical implemented in the database [19].

As a first step, the relation between items in a relationships table is structured as a basis for the system learns the model, i.e., the relationships table becomes to the model in the way to learn and execute their processes of inference. The relations between items are made from the information comparative tables and respecting the proposed metrics [20]. Once the necessary user data is obtained, with the help of a classification algorithm the inference process can be perform, in order to determine the physical activity and healthy diet proper intervention for the user according to his/her health status, needs and interests.

These relationships were made between all items of the figure 3, with the purpose of guaranteeing quality of information because the relationships declare data structures and it allows avoid redundancy of knowledge.

Consider figure 4 to see the learning process of the model based on the items relations, it can be observed the indirect relations that rise in the flow lines, such as the way it affects the user's tastes in food. Another finding is the dependency level between items, the black lines represent unidirectional relationships and the blue ones represent bidirectional relationships. Thus, the relations between all items are structured, achieving a successful possibility of inference in promoting health.

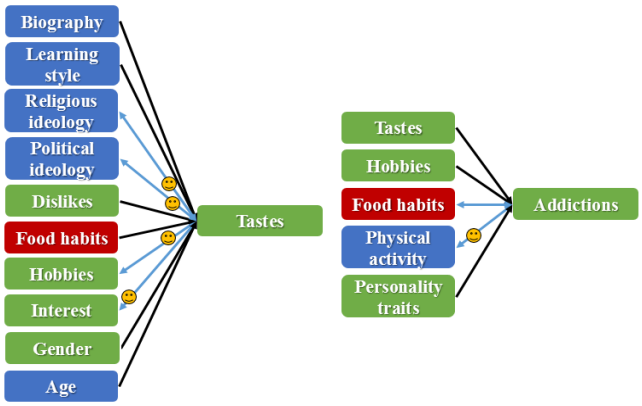

Figure 4. Relationship's items about tastes and user addictions

4.3 Physical Data Model

Describes the physical means by which data are stored [19]. Table 1 specifies the characteristics of each attribute that is considered for a physical activity recommendation. It is important to prepare the information according to the tool that is going to be used to create a mining model. Since the software used is the Weka tool, the data format must be a file type .ARFF, the structure of these files is exemplified below in the figure 5.

```
@relation UserModel

@attribute BMI{low_weight, normal_weight, overweight, obesity}
@attribute Life_cycle{teenagers, adult_teenagers}
@attribute Ethnicity{indigenous, afro-Colombian, other}
@attribute Trauma{mobility, visual, auditory, without_trauma}
@attribute System_usage{health, beauty, sport}
@attribute Cardiovascular_disease{diabetes, hypertension, without_risk}
@attribute Intervention{dance, walk, bodily_exercises, stretch, stretch_eyes, limbs,
personal hygiene, HIIT, labours, labours_limbs, labours_eyes, LISS, swim, eyes,
relaxation, SCC, jog}
@data
Low_weight,adult_teenagers,indigenous,mobility,health,diabetes,stretch_eyes

@relation: the name of the relationship is indicated.
@attribute: specifies a variable or attribute. Attributes can be from various types ac-
cording to their characteristics.
@data: in this section the data enter in the model, either for analysis or as mining
model training set.
```

Figure 5. Data format

Table 1. Mining model attributes

Attribute	Description	Values
BMI	Represents a previous inference in the user's physical condition. It is a nominal fact type.	Low weight, normal weight, overweight, obesity.
Life cycle	It is a model fact inferred from their date of birth. Despite being an integer, is taken as a nominal value in the relationships table.	Teenagers, adult teenagers.
Ethnicity	Indicates the racial group a person belongs. It is a nominal value.	Indigenous, afro-Colombian, other.
Trauma	It represents a person with disability. It is a nominal fact.	Mobility, visual, auditory, without trauma.
System usage	It is the aim of the system user. It is a nominal fact.	Health, beauty, sport.
Cardiovascular disease	Indicates a user clinical condition. It is a nominal fact.	Diabetes, hypertension, without risk.
Intervention	It is the class of dataset. It is a nominal fact.	Dance, walk, bodily exercises, stretch, stretch_eyes, limbs, personal hygiene, HIIT, labours, labours_limbs, labours_eyes, LISS, swim, eyes, relaxation, SCC, jog.

4.4 Architecture

The architecture is divided into three parts: (1) servers, (2) control (3) view. Following there is a brief explanation of each [20].

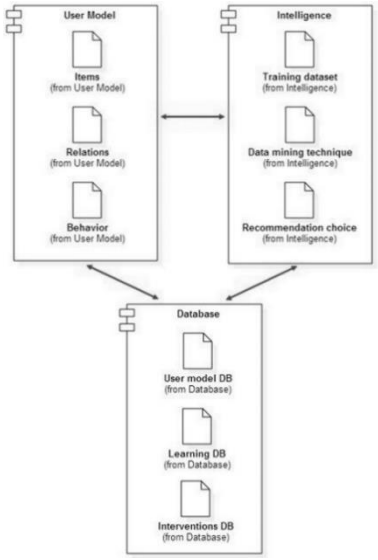

Figure 6. Diagram component of the architecture

In figure 6 is possible to see the relations of components described in the control part that is highlighted in figure 8. a) User Model Component, it considers the selected items from the user characterization through the model, the relations between items so structured a prior knowledge to determine the inference rules. b) Intelligence Component, this is where the training dataset supported by the relations between items becomes relevant as an essential part of detecting patterns in order to achieve health inferences. Data mining techniques and inference decisions are

also part of the intelligence. From this component is achieved to promotion Physical Activity and Healthy Diet recommending personalized interventions. c) Database Component, block database becomes in the user model and intelligence components support, since storing the attributes of the learning processes and feedback and the inferred interventions are collected in databases included in this space.

5. RESULTS: TEST WITH WEKA

The results according to the data mining process within the user model are produced by a machine learning tool called Weka. Weka is a freeware tool developed in Java that allows the implementation of different data mining techniques; it also can perform data filtering and preprocessing tasks [21].

It is known that several data mining algorithms can be used to accomplish an aim, however given the data model characteristics, three known algorithms were examined in the literature: decision trees, particularly the J48 algorithm, the k-nearest neighbor algorithm and Bayesian networks based in network topologies [22]. The data model evaluation is made with the three mentioned algorithms through Delphi methodology, making tests of try and error and choosing the best algorithm for the data model.

As a first step, it is defined that dataset does not present missing, anomalous data or little use information that will obstruct the data mining process. In general, all initial variables are trusted to be use in the creation of data mining model. Prior to running the test algorithm, the Cross-validation the main class to analyze is the recommendation of an intervention of physical activity, so it is been specified within the previous use of the algorithm configurations.

```
Number of Leaves  :     163

Size of the tree :      244


Time taken to build model: 0.31 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         382               44.213  %
Incorrectly Classified Instances       482               55.787  %
Kappa statistic                          0.3832
Mean absolute error                      0.0743
Root mean squared error                  0.2242
Relative absolute error                 69.0742 %
Root relative squared error             96.7396 %
Total Number of Instances              864
```

Figure 7. J48 algorithm results window

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         348               40.2778 %
Incorrectly Classified Instances       516               59.7222 %
Kappa statistic                          0.3307
Mean absolute error                      0.0908
Root mean squared error                  0.2123
Relative absolute error                 84.4184 %
Root relative squared error             91.5819 %
Total Number of Instances              864
```

Figure 8. Knn algorithm results window

```
=== Summary ===

Correctly Classified Instances         351               40.625  %
Incorrectly Classified Instances       513               59.375  %
Kappa statistic                          0.3373
Mean absolute error                      0.0856
Root mean squared error                  0.2082
Relative absolute error                 79.6391 %
Root relative squared error             89.8415 %
Total Number of Instances              864
```

Figure 9. BayesNet algorithm results window

The figure 7 shows the total percentage of correct classified instances, this value means there is a 44,213% of success in the classification of an intervention according to its training dataset. Although this percentage does not overcome the 50% of success, such value is expected given the previous data analysis, because it is evident that the interventions relevant to the class are varied, which means more than one intervention works for a determined combination case of user attributes. The decision tree allows performing a deeper analysis about the model behavior.

In general, the model results under the J48 algorithm shows that the attributes *trauma*, *BMI* and *life cycle* accomplish a decisive role in the selection of a defined intervention; this result is consistent with the data exploratory analysis. Likewise, the classification distribution of the instances indicates that the interventions are not subject to a specific user profile, some user characteristics defined the selection of an intervention but in general, the interventions are varied for varied users.

Applying the k-nearest neighbor algorithm, using five (5) neighbors for classification (figure 8).The Bayesian networks provide the possibility of making a qualitative and quantitative analysis of the information; these properties allow complementing the dataset analysis. It is obtain the percentage of correct classified instances by the algorithm as a first step (figure 9)

Figures 8 and 9 indicates a little more than 40% of correct classified instances, just as the previous algorithms, the rightness percentage does not overcome the 50%. Analyzing the network structure generated by the algorithm, it is observed that there is a relation between the *trauma*, *BMI* and *life cycle* variables, such relation had been pointed across the analysis and the structure confirms it. Reviewing the probability tables of some interest nodes, it is evident some similar behaviors founded in the decision tree, e.g., the high variation rate of an intervention from a user trauma or his/her BMI.

Table 2. Probabilities distribution table of trauma node

Intervention	BMI	mobility	visual	auditory	without_trauma
Dance	low_weight	0,045	0,045	0,045	0,864
Dance	norm_weight	0,029	0,029	0,029	0,912
Dance	overweight	0,25	0,25	0,25	0,25
Dance	obesity	0,25	0,25	0,25	0,25
Walk	low_weight	0,01	0,01	0,402	0,578
Walk	norm_weight	0,009	0,009	0,361	0,62
Walk	overweight	0,011	0,011	0,633	0,344
Walk	obesity	0,25	0,25	0,25	0,25
Bodily_exercises	low_weight	0,036	0,893	0,036	0,036
Bodily_exercises	norm_weight	0,033	0,9	0,033	0,033
Bodily_exercises	overweight	0,017	0,95	0,017	0,017
Bodily_exercises	obesity	0,015	0,456	0,015	0,515
Stretch	low_weight	0,25	0,25	0,25	0,25
Stretch	norm_weight	0,25	0,25	0,25	0,25
Stretch	overweight	0,25	0,25	0,25	0,25
Stretch	obesity	0,071	0,071	0,786	0,071
Stretch_eyes	low_weight	0,85	0,05	0,05	0,05
Stretch_eyes	norm_weight	0,893	0,036	0,036	0,036
Stretch_eyes	overweight	0,893	0,036	0,036	0,036
Stretch_eyes	obesity	0,9	0,033	0,033	0,033
Limbs	low_weight	0,05	0,35	0,45	0,15
Limbs	norm_weight	0,1	0,7	0,1	0,1

Table 3. Results summary from the apply algorithms to dataset

The distribution table verifies that for more than one intervention there are more than one user profile that can realize the intervention activity. Hereafter, a table is presented with the results summary thrown by the applied algorithms to the user model dataset.

Algorithm	Correct classified instances	Kappa statistic
J48 (Decision tree)	44.213%	0.3832
IBk (Knn)	40.2778%	0.3307
BayesNet	40.625%	0.3373

From the previous analysis, the following contributions are presented: a)The algorithm with a better response to the data model dataset is the decision tree. b)The trauma, BMI and life cycle attributes are closely related between them, dividing the interventions by groups. This indicates that even the interventions are diverse for more than one user kind, these are more suitable to certain kind of users defined by the mentioned attributes. c)The percentage of correct classified instances suggest the majority of interventions are not attached to a certain user kind, although some of them are quite specific for a user, in general, most of them are generic to any user. d)The model results are consistent with the hypothesis from the exploratory analysis, this implies that it is a reliable model in its behavior, for that matter it is a consistent data model.

6. CONCLUSIONS AND FUTURE WORK

The salient contributions to the development of this work are: a)The CRISP-DM methodology suggests that prior treatment of information that can be analyzed is vital to the expected results quality when applying any data mining technique. b)The inference relationships obtained from the user model elements, which are covered by the logical data model allows the system learns the user personality inherent characteristics in a more efficiently way so the personalization process could be more accurate and appropriate. c)The proposed components architecture resulting from the framework application is generic and flexible to be implemented in any personalized system in health promotion. d)Through the test with weka, the relations of the data model inference are evaluated verifying the validity of the framework for a personalized system, so the efficiency and satisfaction of the given recommendation are guaranteed. e)The percentage obtained in the algorithms tests is good consider the size of the evaluated dataset. With more variables and relations in the dataset, this percentage could be higher.

As futures work it is proposed: 1) Run the former algorithms with a bigger dataset formed by more variables such as the ones described in figure 5 and relations generated from the user model that conforms the conceptual data model.2) Evaluate the personalized system in a study case that reveals the importance of the information obtained through the user model and the process executed from the data model proposed.

7. ACKNOWLEDGMENTS

This work was performed under the doctoral work “Adaptive System to support the Promotion of Physical Activity and Healthy Eating” financed by Colciencias.

8. REFERENCES

[1] Leka, S., Griffiths A., Cox, T.: La organización del trabajo y el estrés: estrategias sistemáticas de solución de problemas para empleadores, personal directivo y representantes sindicales. In: Protección de la salud de los trabajadores, vol. 3 (2004)

[2] Hernández, J.: Análisis y Extracción de Conocimiento en Sistemas de Información: Datawarehouse y Datamining, http://users.dsic.upv.es/~jorallo/cursoDWDM/

[3] International Organization for Standarization: ISO/TR 14292: Personal Health Records — Definition, Scope and Context (2012)

[4] Martins, A., Faria, L., De Carvhalo, C., Carrapatoso, E.: User Modeling in Adaptive Hypermedia Educational Systems. In: Educational Technology & Society, vol. 11, nº 1, pp. 194-207 (2008)

[5] Ifeachor, Z., Hu, P., Sun, L., Hudson, N. Zervakis, M.: Bioprofiling over Grid for Personalised eHealthcare for AD. In: Proceedings of the 2009 conference on Computational Intelligence and Bioengineering: Essays in Memory of Antonina Starita, pp. 155-165 (2009)

[6] Bellazzi, R., Larizza, C., Gabetta, M., Milani, G., Nuzzo, A., Favalli, V., Arbustini, E.: Translational bioinformatics: challenges and opportunities for case-based reasoning and decision support. In: Case-Based Reasoning. Research and Development, pp. 1-11 (2010)

[7] El-Sappagh, S., El-Masri, S., Elmogy, M., Riad, A., Saddik, B.: An Ontological Case Base Engineering Methodology for Diabetes Management. In: Journal of medical systems, vol. 38, nº 8, pp. 1-14 (2014)

[8] Isern, D., Moreno, A., Sánchez, D., Hajnal, A., Pedone, G., Varga, L.: Agent-based execution of personalized home care treatments. In: Applied Intelligence, vol. 34, nº 2, pp. 155-180 (2011)

[9] Chen, H., Compton, S., Hsiao, O.: DiabeticLink: a health big data system for patient empowerment and personalized healthcare. In: Smart Health, pp. 71-83 (2013)

[10] Huang, Z., Lu, X., Duan, H., Zhao, C.: Collaboration-based medical knowledge recommendation. In: Artificial intelligence in medicine, vol. 55, nº 1, pp. 13-24 (2012)

[11] Yadav, N., Poellabauer, C.: An architecture for personalized health information retrieval. In: Proceedings of the 2012 international workshop on Smart health and wellbeing, pp. 41-48 (2012)

[12] Meyer, J., Çakır-Turgut, E., Helmer, A.: Supporting a healthy lifestyle by re-using personal online data. In: ACM SIGHIT Record, vol. 2, nº 1, pp. 13-13 (2012)

[13] Kahani, M.: Engineering Research Methodology, http://www.slideserve.com/donnan/engineering-research-methodology

[14] Carreño, M.: El método Delphi: cuando dos cabezas piensan más que una en el desarrollo de guías de práctica clínica. In: Redalyc, vol. 38, nº 1, pp. 185-193 (2009)

[15] International Organization for Standarization: Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems (2010)

[16] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 Step-by-step data mining guide (2010)

[17] Data Mining Concepts, “Data Mining Process”, https://docs.oracle.com/cd/B19306_01/datamine.102/b14339/5dmtasks.htm.

[18] Whatls.com, http://whatis.techtarget.com/definition/framework.

[19] Agile Data: Data Modeling 101, http://www.agiledata.org/essays/dataModeling101.html

[20] Rico, C., Girón, E., Cerón, G., López, D.: Towards a Standardized User Model for Personalized Systems in Health. In: Revista S&T, vol. 13, no. 34, pp. 9-29 (2015)

[21] Machine Learning Group at the University of Waikato: Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/index.html

[22] Bhavsar, H., Ganatra, A.: A Comparative Study of Training Algorithms for Supervised Machine Learning In: International Journal of Soft Computing and Engineering – IJSCE, vol. 2, n° 4, pp. 74-81 (2012)