DDA_{AV} - Student Performance Detector

Andreia Rosangela Kessler Mühlbeier
UFSM – Universidade Federal de Santa Maria
Santa Maria - Brasil
andreiamuhlbeier@yahoo.com.br

Aderson de Carvalho
UFSM – Universidade Federal de Santa Maria
Santa Maria - Brasil
acarvalho@inf.ufsm.br

Fabiana Santiago Sgobbi
UFRGS – Universidade Federal do Rio Grande do Sul
Porto Alegre - Brasil
fabianasgobbi@gmail.com

Roseclea Duarte Medina
UFSM – Universidade Federal de Santa Maria
Santa Maria - Brasil
roseclea.medina@gmail.com

Liane Margarida Rockenbach Tarouco
UFRGS – Universidade Federal do Rio Grande do Sul
Porto Alegre - Brasil
liane@penta.ufrgs.br

Abstract - The accelerated development and increasing use of virtual learning environments (VLE) motivates a transformation in education. This article presents a study of techniques and data mining tools (MD), which aims to research and analyze the student's behavior in the virtual environment, real-time execution of a course, providing feedback to the teacher about the students' academic performance encouraging its participation to improve the performance and preventing the circumvention of the course. The results obtained from the research of tools and techniques demonstrate that it is possible to obtain these inferences during periods.

Keywords – data mining; student achievement; weka; knowledge discovery in databases.

I. INTRODUCTION

The progress and the widespread use of technologies open up new perspectives in terms of classroom teaching, semi-classroom and distance learning. With support tools and access through mobile devices, the Virtual Learning Environments (VLE) are highlighted with great expansion in the educational process. However, virtual learning environments bring a transformation in education, allowing greater interaction in the environment among students, teachers, tutors, content and interfaces; this fact takes interactions as an effective part of the learning processes [1].

The data store large volumes of information in environments that are very rich sources of knowledge which end up being left out, sometimes by lack of knowledge in how to interpret them. [2]

In this context, the process of evaluating the student's performance in virtual learning environments is held at the end of disciplines, scoring the cognitive performance of an often static way, practically without time for actions of retroactive recovery of this student. However, the need for these evaluation and monitoring actions during the course it is evident, to propose alternatives for its best achievement, in order to generate subsidies for early identification, in time to successfully complete the student's learning process.

This research aims to research and analyze the student performance in the virtual learning environment, using data mining (MD) methods, through efficient techniques for Knowledge Discovery in Databases (KDD) in the information stored in the database, providing perform the mapping of student performance, in real-time execution of the course. With this mapping, the teacher will have a feedback that may assist in stimulating the participation and improving the performance of the student in the course.

This paper is organized this way: Section II presents the virtual learning environments. Section III describes the theoretical basis for knowledge discovery in databases. Section IV describes the WEKA tool and its relevant characteristics. Section V presents the related work. Section VI presents the development of the research methodology. Section VII presents the Student Performance Detector (DDA_{AV}), considering its pedagogical and technological aspects, and results. Section VIII presents the conclusions of this paper.

II. VIRTUAL LEARNING ENVIRONMENTS

The Virtual Learning Environments (VLE) are softwares applications on web servers; they have a set of tools which allow the creation of courses and the development of learning. These environments often classify their users in three pre-set profiles: Administrator, Teacher and Student. However, it is valid to mention that there is another profile

 ISBN: 978-1-61208-418-3

as the Tutor, who works with the teacher, being responsible for pedagogical mediation [3].

One environment widely used in educational spaces is Modular Object-Oriented Dynamic Learning Environment (MOODLE); it is an Open Source software, which had its development started in the 1990s by Martin Dougiamas based on learning philosophies of constructivism and social constructivism, supporting the creation and management of courses with a focus on collaborative work and in a simple and intuitive environment to use [4].

Amid other existing free softwares applications which allow the teaching/learning process, it was opted for MOODLE to conduct this research. The choice is given by virtue of being an environment with lots of tools constantly updated where there is a broad group of users who collaborated with their evolution, in addition to integrating other techniques in their repositories.

III. KNOWLEDGE DISCOVERY DATABASES

With the advancement of computer technologies which allows the storage and processing of a large volume of data, new technologies have been developed to assist in the extraction of information from these databases, through techniques such as Knowledge Discovery in Databases (KDD) and Data Mining (MD) [5].

The Knowledge Discovery process in Databases (KDD) presented by Fayyad [6], is "a non-trivial process of identifying valid standards, unknown, potentially useful and interpretable." It is to discover useful knowledge to the stored data from the application of modern techniques of data mining, evaluation of achieved standards and the interpretation of results.

The KDD process involves complex steps and each one must be performed carefully as it is very important that the established objectives and the overall success of the application are achieved. The steps are divided into: Pre-processing, data mining and post-processing [6].

A. Pre-processing

This stage is the identification and understanding of the problem, considering aspects such as goals and the data sources of which you want to extract knowledge. The next step is the selection of data from the sources, according to the objectives of the process, and the processing of data in order to be subjected to the methods and tools, the standards extraction phase.

B. Post-processing

At this stage, the extracted knowledge is evaluated about its quality and/or utility so it can be used to support a process of decision-making, whether by a human expert or an expert system.

C. Data Mining

The data mining phase is the central step that runs the knowledge discovery itself; its algorithms are the responsible to produce, semi-automatically, knowledge from existing data.

The MD process covers data selection, preparation, implementation tasks and/or techniques with their algorithms to make analyzes of the results in order to detect the extracted knowledge. The MD is divided into:

a) Association.
b) Classification.
c) Estimate.
d) Segmentation.
e) Summarization.

These tasks are performed by implementing algorithms with machine learning techniques such as:

a) Genetic algorithms.
b) Decision trees.
c) Association rule discovery.
d) Based reasoning cases.
e) Neural networks.

For the present study, the classification task that performs the decision tree technique, through the J48 algorithm [7] was chosen. The choice reflects the objective of searching, through the conditions offered to the 2 techniques, and identifying the student's performance with the result of some parameters. The decision tree respects a hierarchical test sequence, constructed along a tree structure with leaf nodes representing classes, where the algorithm expresses the rule through the path of the tree, from the root until a leaf node. The J48 algorithm is the ranking algorithm, implemented in Java, from the algorithm C 4.5 release 8; it builds a decision tree model based on a set of attributes, and it uses this model to classify instances in a cluster [7].

IV. WEKA TOOL

With the growing number of digital information, the interest in implicit knowledge discovery of informations grows. According to [8][9], there are some features that should be considered to choose a knowledge discovery tool.

a) Ability to access a variety of data sources, being online and offline.

b) Ability to include data models, object-oriented or non-standard models.

c) Processing capacity related to the maximum number of tables, records, or attributes.

d) Variety of attributes' types which can manipulate the tool.

e) Type of query language.

The Weka tool was developed at the University of New Zealand, in the Department of Computing. This tool uses

 ISBN: 978-1-61208-418-3

techniques to perform the following data mining tasks [10]: association, classification and clustering.

Mining begins by reading data from a file formatted especially for the tool, the ARFF (Attribute-Relation File Format). The ARFF is a text file that describes a list of instances which share a set of attributes [10].

In Weka, there is a variety of techniques for the listed tasks, such: ID3, PRISM, OneR and Naive Bayes [7].

The choice of the tool WEKA for this work is justified because it makes the system portable and it presents a cross-platform object-oriented language. The portability of language allows the tool to run on different platforms, and its object orientation produces advantages such as modularity, polymorphism, encapsulation, code reuse among others [11].

V. RELATED WORK

The work of Maia et al. [12] focuses on the future performance of students in disciplines of an undergraduate degree, are made from the grades achieved in subjects taken already. In this model, students and course subjects were modeled as nodes and their representation as the edges that make up a graph. The authors reported that, among the subjects there is a large variation in the values of the average errors analyzed, ranging from 3.6% to 100%. However, the authors conclude that a significant mean error for a discipline could indicate: that it does not have great relationship with the other subjects in the curriculum, or the assessment has some degree of disconnection with the results obtained in other disciplines.

In [13], to see high rates of dropout students in distance courses, one through an interview fieldwork was performed with a professional distance education, to identify some evidence of evasion courses. Based on the identified attributes, a prototype was designed to identify with the user log records stored in the database, information from these students. The work follows the KDD online database and used the WEKA tool, in particular the J48 algorithm that identifies behavioral prediction by the decision trees show. The author concludes the research, saying it can be identified through access to AVA, use patterns and certain diagnoses with evasion evidence thus propose corrective measures to ensure that a pass student to have a material behavior in the use of a VLE.

Accordingly, VLE [14] used to support classroom courses, are characterized by storing a large volume of data. These environments need tools to filter useful information to detect student performance. The research investigated the data stored in the VLE to extract information related to student performance. To detect this information was necessary to select a set of attributes, considering three dimensions: usage profile of VLE, student-student interaction and two-way interaction student-teacher. The form used RandomForest [7] and MultilayerPerceptron [7] ranking algorithms available in the Weka tool is pointed out that in all the experiments we used the method K-fold Cross-Validation [7] as data layering technique. The results of using the MD techniques on the selected set of attributes demonstrated that it is possible to obtain inferences regarding student performance with overall accuracy rates ranging from 72% to 80%, but leaves specific that the accuracy rate may be insufficient to evaluate the quality of the classification model, since the number of instances of classes is unbalanced in the case study, due to each being in different scenarios.

No analyzes focusing on student performance in the virtual learning environment and real-time course of execution were not identified in the current researches. However, there are indications that this type of analysis is important for the teacher to assist in stimulating participation and improvement in students' learning performance in MOODLE.

VI. RESEARCH METHODOLOGY

The nature of the research is ranked as a field research of qualitative-descriptive type. According to Lakatos and de Marconi [15], a field survey aims to obtain information about an issue, for which it is searched for a response in order to discover the relation between them.

In the first stage of development, to the data mining application in VLE, a literature research was performed in order to have knowledge of how the knowledge discovery in databases was conducted to understand and analyze the operation of data mining steps (tasks and techniques) and the functionality of available data mining tools.

The second stage involved two moments: the assembly of a hardware infrastructure that supported the installation, development and implementation of this work, consisting of a Dell Power Edge T300 server, with Intel Xenon Quadcore X3363 2.83 GHz processor with 4 physical cores and four virtual cores, 8GB of RAM, two 500GB hard drives and Windows Server 2008 operating system 64-bit. On this server, the following programs were installed: WampServer version 2.2, which provides softwares on its package that is necessary for the operation of MOODLE, as Apache version 2.2.22 server; the database MySQL version 5.5.24; PHP version 5.2.13 and phpMyAdmin version 3.4.10.1. After, the installation of MOODLE VLE version 2.5.2 was done. For the development, editing, and environmental manipulation, a Philco laptop with Intel Pentium Dual-Core processor was used, SU 4100 1.3GHz, 2GB of RAM, a 320GB hard drive and the operating system Windows 7 Ultimate 64-bit.

After the installation of MOODLE environment, to compose the research scenario it was worked with the database of the discipline Introduction to the Media Integration in Education. This which composes the curricular base of the course Specialization in Media in

 ISBN: 978-1-61208-418-3

Education, Post-Graduate *lato sensu*, of Universidade Federal de Santa Maria (UFSM), offered in distance education mode, during the second half of 2012. The course integrates in this edition 134 (one hundred and thirty four) students divided into five (5) campuses (Cachoeira do Sul, Cruz Alta, Panambi, Restinga Seca and Santana do Livramento) and it is composed by 10 activities.

In the third stage, the process of modeling the functioning of the block began. The proposed model was executed with the Astah Community tool, which allows the construction of Unified Modeling Language (UML) [16] diagrams, such as use case diagrams, activity diagrams, among others. The Astah Community [16] is afree modeling software for object-oriented systems design, based on the diagrams and in the notations of UML, which can generate code in Java language.

The fourth step involved the installation of WEKA version 3.7.8 tool, developed in the programming language JAVA, which offers several pre-data processing algorithms and results analysis. In the software, files were generated in the extension (* .arff) with their respective rules, to run the J48 algorithm. This algorithm allows the construction of decision trees that classify and display in their branches the most relevant attributes, as name, campus, discipline, notes of the activities and situation.

In the fifth step, it was executed the translation of the rules generated in WEKA software with the extension (* .arff) for the PHP language. The informations were extracted from the MOODLE environment database, in the form of an Excel electronic spreadsheet (note, campus and situation). After, they were processed in WEKA tool, which originated a file in notepad generated in the extension (* .arff). As a result, the generated file has been converted to PHP programming language [17] through the software PHP Editor.

In the sixth stage, the construction of the block was performed, receiving the number of any proposed activities in the discipline to be analyzed, and the integration of it in the MOODLE learning environment. The block which was developed, works by a plug implemented through an application, allowing its application in the environment's interface.

In the seventh step, tests were performed to validate the integration in each development stage, through the white box test (test performed by the developer). According to Sommerville [18], the tests are derived from the knowledge about the structure and implementation of the software, in other words, the developer seeks to test and to know all the system code, by examining the logical path to verify the operation of the tool. For this development, the following was used: basis path testing - aims to check if each statement of the system was performed at least once during the testing activities; condition testing - is based on verify if all logical conditions contained in the system, i.e., common error condition such as parentheses, relational operators and arithmetic expressions [19].

The first test was made after the generation of rules in the format (* .arff), when the consistency of J48 algorithm was verified. The second test was done after translating the rules to the PHP language. In the final test, the block was validated after its integration into the MOODLE environment.

With the active plug-in, the teacher tells the number which corresponds to the activity proposed in the discipline and in response a report is submitted as a web page, informing only the students who obtained low performance. The result is stored in tables in the SQL - based database of MOODLE environment with information about the activities.

Finally, in the eighth stage, monitoring reports were generated about student performance (creation with PHP language) and creation of the decision tree and graphics (WEKA software).

VII. DDA_{AV} – STUDENT PERFORMANCE DETECTOR

The environment aims to detect the student's performance during the execution of the course, using data mining technique in each executed and evaluated activity.

The research scenario consists of the discipline "Introduction to the Media Integration in Education", with 134 students divided into five campuses and subdivided into 10 activities with educational content in correlation to the main subject.

The relevant attributes extracted from the environment are: name, campus, discipline, and the grades of the activities performed by the students and registered by the teacher. In the context of the executed work, performance can mean the evaluation of student interaction in the environment, the learning level in the proposed activities, level of participation and level of difficulty in interpreting the task.

Figure 1 illustrates the course of the environment, displaying the proposed content and activities, the block "Student Performance" and the option of sending the number for any of the activities proposed in the discipline.

As output results, WEKA tool presented the attributes (name, campus, discipline, observation of the activities and situation) after running the J48 algorithm, and some other information relevant for classifying attributes of this work. The selected attributes totaled 134 cases, representing 100% of all the stored in the database; the "Properly Classified Instances" where the attributes were correctly classified contains 123 bodies which means 91.79%, and the "Incorrectly Classified Instances" shows the misclassification of 11 cases which is only 8.21%.

Figure 2 shown the performance of students in a report generated in the MOODLE environment. In the report, one can see the individual performance (student name) and general (discipline), with all the activities.

 ISBN: 978-1-61208-418-3

The proposed work is justified by the importance of the teacher to follow, throughout the course, the implementation process, avoiding a posteriori analysis, i.e., with the results, the teacher is able to give special attention to students with difficulties in constant learning, directly, without requiring more than a tool for analysis.

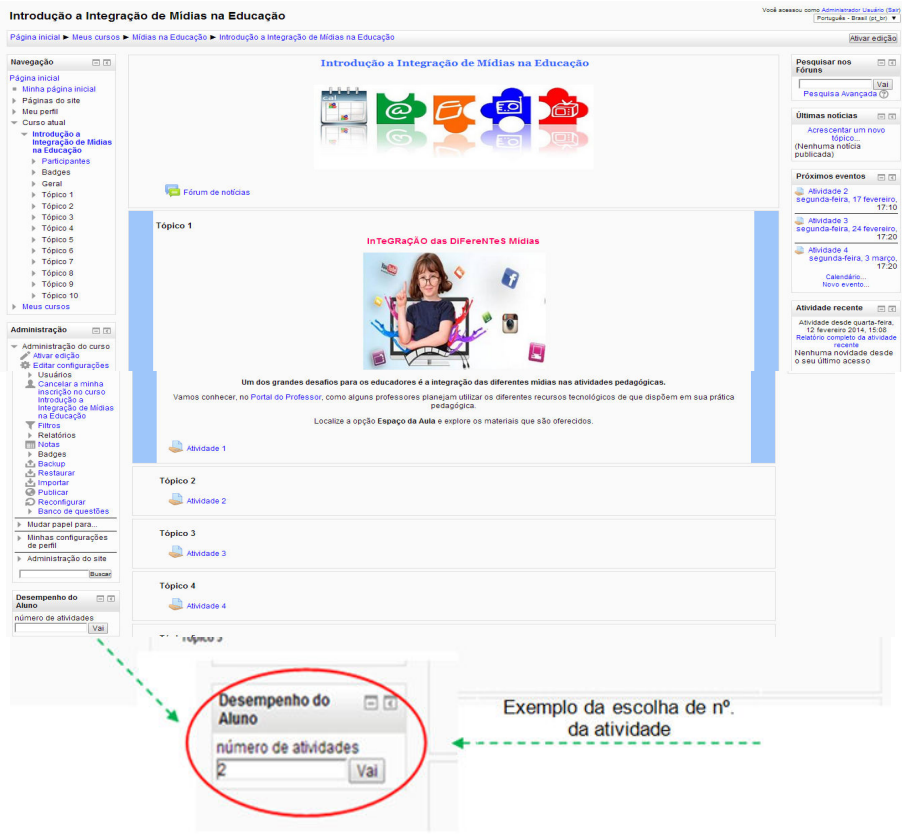
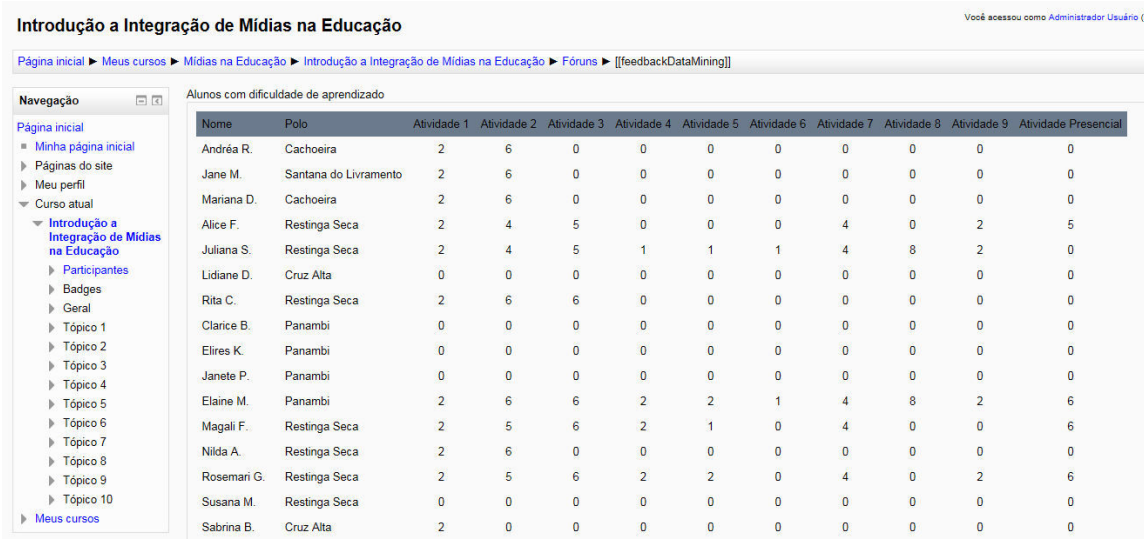


Figure 1. Environmental progress and block.



Nome	Polo	Atividade 1	Atividade 2	Atividade 3	Atividade 4	Atividade 5	Atividade 6	Atividade 7	Atividade 8	Atividade 9	Atividade Presencial
Andréa R.	Cachoeira	2	6	0	0	0	0	0	0	0	0
Jane M.	Santana do Livramento	2	6	0	0	0	0	0	0	0	0
Mariana D.	Cachoeira	2	6	0	0	0	0	0	0	0	0
Alice F.	Restinga Seca	2	4	5	0	0	0	4	0	2	5
Juliana S.	Restinga Seca	2	4	5	1	1	1	4	8	2	0
Lidiane D.	Cruz Alta	0	0	0	0	0	0	0	0	0	0
Rita C.	Restinga Seca	2	6	6	0	0	0	0	0	0	0
Clarice B.	Panambi	0	0	0	0	0	0	0	0	0	0
Elires K.	Panambi	0	0	0	0	0	0	0	0	0	0
Janete P.	Panambi	0	0	0	0	0	0	0	0	0	0
Elaine M.	Panambi	2	6	6	2	2	1	4	8	2	6
Magali F.	Restinga Seca	2	5	6	2	1	0	4	0	0	6
Nilda A.	Restinga Seca	2	6	0	0	0	0	0	0	0	0
Rosemari G.	Restinga Seca	2	5	6	2	2	0	4	0	2	6
Susana M.	Restinga Seca	0	0	0	0	0	0	0	0	0	0
Sabrina B.	Cruz Alta	2	0	0	0	0	0	0	0	0	0

Figura 2. Performance report of students.

 ISBN: 978-1-61208-418-3

VIII. CONCLUSION

Currently, teachers and higher education institutions have faced a huge challenge which is to propose quality education and more individualized, to a growing number of students in various courses offered in different modes (presence, semi-presence and distance education). To assist in this process, VLE have been frequently used because they allow a greater control, various types of interaction, and the adoption of different methodologies and strategies. However, all this multiplicity and complexity of information can hamper the task of following and evaluating the student performance.

In this sense, the KDD process that aims to discover new knowledge, assists in the exploration of large volumes of data and detects useful information, through the application of techniques and tasks which implement MD algorithms.

The objective of this research was to apply data mining techniques in a VLE, presenting to the teacher a student's performance report during the execution of a course, which entails the prevention of the student's failure and consequently that occurs the evasion of the course. The report was extracted by means of integration of rules J48 classification algorithm, with the relevant attributes of the environment database.

The DDA_{AV} had as the research scenario, data from VLE MOODLE, which involves pedagogical processes between teachers and students in courses offered in virtual environments. The survey endorsed the difficulty of analyzing a large amount of data, which are available in the VLEs database and then highlighted the importance of using tools that help the teacher to follow the trajectory of the student, and monitor their performance within the course.

In the survey, the difference between the virtual environments of learning existing is that the DDA_{Av} has the advantage of the unification into a single report, information to the teacher about the trajectory of the student, consisting in a set of relevant data so that the teacher can prepare teaching strategies that meet the individual needs of students. Additionally, DDA_{AV} obtains semi-automatically, the variables "Enough" and "Insufficient" which characterize the performance of the student, for the inference of measures by the teacher.

REFERENCES

[1] R. Donnelly, “Interaction analysis in a 'learning by doing' problem-based professional development context”. Computers & Education, vol. 55 no. 3, p. 1357-1366, 2010.

[2] C. Romero, S. Ventura, M. Pechenizkiy, R. Backer, S. J. D., “Handbook of Educational Data Mining”, Ed. C R C, p. 535, 2012.

[3] A. P. Rodrigues, "Virtual Environment Integration with Digital Learning Repository" 2012. Thesis (Ph.D. in Education.) - Federal University of Rio Grande do Sul - UFRGS, Porto Alegre, p.188.

[4] MOODLE. “Statistics Documentation Moodle”. 2011. Available at: <http://docs.MOODLE.org/22/en/Statistics>. Accessed: Mar 2015.

[5] R. Goldschmidt, E. L. Passos, “Data Mining: um guia prático”. Rio de Janeiro: Elsevier, 2005. 2ª. Reimpressão.

[6] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, “The KDD process for extracting useful knowledge from volumes of data”. Communications of the ACM, New York, vol. 39, no. 11, p.27-34, 1996.

[7] WEKA. Waikato environment for knowledge analysis. 2013. Available at: <http://www.cs.waikato.ac.nz/ml/weka/>. Accessed: Abr 2015.

[8] M. Goebel, L. Gruenwald, “A survey of data mining and knowledge discovery software tools”. In: *SIGKDD Explorations*, June, 1999.

[9] L. A. Vieira, “Tools to estimate missing values in a database in the Pre-Processing Step of a KDD”. Work and Conclusion Course (Computer Science), University of Vale do Itajaí, 2008.

[10] I. H. Witten, E. Frank, M. A. Hall, “Data mining: Practical machine learning tools and techniques”. San Francisco: Morgan Kaufmann, 3 ed., 2011.

[11] D. Jacomini, “Entrants of Base Analysis in UNIDAVI”. Work Completion Course in Information Systems. New South Wales, in 2008.

[12] R. F. Maia, E. M. Spina, S. S. Shimizu, “System Student Performance Forecast for Assisted Learning and Course Rating”. Proceedings of the XXI SBIE -XVI WIE, 2010.

[13] C. S. de Afiune, “Educational Data Mining: Prediction Behavior in Distance Education Environments (DE)”. Term paper. State University of Goias, Anapolis, 2012, p. 108

[14] E. Gottardo, “Academic Performance estimation Students in A AVA using Data Mining Techniques”. Dissertation (Master of Applied Computing, Federal Technological University of Paraná (UTFPR), 2012, p.85.

[15] E. M. Lakatos, M. A. de Marconi, “Scientific Methodology fundamentals”. 5th . Ed . Editora Atlas . Faculty of Arts, 2003.

[16] ASTAH, “Astah Community”. 2010. Available at: <http://astah.change-vision.com/en/product/stah-community.html>. Accessed: Mar. 2014.

[17] C.A. J. Oliviero. "Make a site with PHP 5.2 MySQL 5.0, E-Commerce Driven project. "1st Edition. Ed. Erica Ltda. São Paulo, p.412, 2013.

[18] I. Sommerville, “Software Engineering”. Edição 6: Addison-Wesley, 2003.

[19] R. Pressman, “Software Engineering - A Professional Approach ”. 7th Edition, 2011.

 ISBN: 978-1-61208-418-3