Evaluating Performance and Dropouts of Undergraduates Using Educational Data Mining

Laci Mary Barbosa Manhães
PESC/COPPE
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brazil
manhaes@cos.ufrj.br

Sérgio Manuel Serra da Cruz
PPGMMC/ICE
Universidade Federal Rural do Rio de Janeiro (UFRRJ)
Seropédica, Brazil
serra@ufrrj.br

Geraldo Zimbrão
PESC/COPPE
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brazil
zimbrao@cos.ufrj.br

ABSTRACT
Undergraduate students have different levels of motivation, different attitudes about learning, and different responses to specific instructional practices. Predicting the academic performance of students is an issue faced by many universities in emerging countries. Although those institutions store large amounts of educational data they are unsuccessful to detect which students are at risk of leaving the educational system. This paper presents an architecture that uses educational data mining techniques to predict and identify those who face the risk of dropping out. The approach may assist educational managers in supervising the development of students at the end of each academic term, identifying the ones with difficulties to fulfill their requirements. This paper shows experimental studies using real data about six undergraduate courses in one the largest Brazilian public universities.

Categories and Subject Descriptors
H.2 [**Information Systems**]: decision support, data mining.

General Terms
Experimentation, Algorithms, Performance, Design, Management.

Keywords
Educational Data Mining, Dropout, Feedback.

1. INTRODUCTION
The modernization of Brazilian society, as in other emerging countries, cannot be achieved without proper educational policies. It will not be possible to build a modern, internationally competitive economy, capable of incorporating and developing innovative technologies, productive processes, and organizational methods with a public higher education sector in a permanent state of crisis. Brazil has public and private universities, with 284 public universities fully financed by the Government, with no tuition costs and fees [1,2]; the total amount of undergraduate seats is increasing but still is not enough. The access to public universities is often skewed towards wealthier Brazilian students as these students tend to come from private high school education institutions which give them an advantage in the entrance examinations. Furthermore, it is true that a small number of Brazilian public universities are competitive globally. Some of the best institutions are located in the industrialized states of São Paulo and Rio de Janeiro, including the University of Campinas (UNICAMP), the University of São Paulo (USP) and the Federal University of Rio de Janeiro (UFRJ) amongst only four South American universities to make the 2013 QS World University Ranking list of the top 300 universities [3].

The Brazilian Government has a diagnosis about the country's economic problem. However, there is no similar consensus in the field of higher education, neither in the Government nor amongst the public and private educational sectors. There are several issues that contribute to reducing the efficiency of the Brazilian higher education system. For instance, on the primary and secondary school levels, educational quality remains low by global standards. Few Brazilians entering school attain the necessary skills to pursue higher education. Thus, Brazil has a small number of students in universities; the indicators of the Organization for Economic Co-operation and Development (OECD) shows that only 11% of the population have a higher education degree, with the average for other countries being 31% [4]. Finally, the latest census of Brazilian higher education shows that the number of students that do not finish the courses is higher than those who complete [1]; this phenomenon, also known as *dropout* [2], is also faced by many organizations around the world. High dropout rates have many undesired consequences not only for the students but also for Brazilian society. For instance, students have their professional lives limited and the society misuses considerable resources to support professors, employers, equipments, laboratories, libraries, and empty classrooms.

Several studies have been made to detect the causes for high dropout rates in the Brazilian educational system [5,6,7,8,9,10]. Some explanations try to elucidate such phenomenon: (i) difficulties to adapt to the academic environment; (ii) difficulties to attend the courses; (iii) poor academic background, or iv) the difficulties in balancing long hours of labour and study, to name but a few. However, if we consider the Computer Science point of view, the academic management software systems used by public Brazilian universities are not properly designed to support educational managers to investigate which students are at risk of dropping out. Despite keeping data on the students (records on examinations, assessments, marks, grades, and academic progression) the systems do not provide predicting or warning tools that can assist managers to extract knowledge on the performance of their students.

The dropout issue has many facets. Numerous studies attempted this. According to Márquez-Vera [11], it is a key problem with special characteristics that requires innovative approaches, such as the adoption of Educational Data Mining (EDM). Even if considering the international scenario, the studies that evaluate the performance and dropouts of students using EDM are still in their early stages [12,13,14,15]. For Romero and Ventura [14], several

open research questions can be investigated with EDM techniques. For instance, (i) the development of tools that can be easily handled and reused by non-specialists in Data Mining in any educational system; (ii) all the knowledge process (preprocessing, data mining, and post-processing) should be packed into a single application.

The goal of this paper is to provide educational managers in Brazilian public universities, EDM non-experts, with a user-friendly approach that provides useful feedback information on the performance of students and predicts the ones whose are at risk of dropping out of the educational system. The contributions of this paper are based on experimental studies that used real world student data of six undergraduate courses of the Federal University of Rio de Janeiro (UFRJ) during a period of 16 years. We chose UFRJ as it is the largest public federal Brazilian University with more than 100 undergraduate courses covering all areas of the Sciences and about 50,000 undergraduate students. Its courses are divided into two semesters (terms); at the end of each semester the educational managers need to plan a schedule for the next, offering courses according to the number of students expected. This activity is a complex task due to the erratic number of students that will attend the course in the next semester. Owing to the large amount of students and low effectiveness of the academic advisory services, it is hard to identify the students who are at risk of dropping out. Thus, the challenge faced in this work is to provide an approach that uses only the academic data available in current academic management software systems to aid managers to identify which students are at risk of dropping out in the next semester.

This paper is organized as follows: Section 2 defines dropout and EDM concepts and discusses related work. Section 3 describes the proposed three-tier architecture named WAVE. Section 4 describes the experiments and its results. Finally, conclusions and further research are presented in Section 5.

2. RELATED WORKS

Avoiding student dropout has been the focus of many higher education researches. Numerous studies have proposed several theoretical models to explain this phenomenon [16,17,18,19,20,21]. Among these studies, Tinto's student integration model [16, 17] is the most discussed. The author studied it by examining the students' personal attributes, educational goals, and commitment to the institution before entry. In this section, we present works that investigated the dropout phenomenon and performance prediction using EDM.

2.1 Student Dropout

There are several definitions about the concept of *dropout* in the literature and on how to measure it. In Brazil, this problem is not new, just the contrary, and in recent years the Public University has investigated dropout and evaluated it in different ways [4,5,6,7,8,9,10]. The works produced considered dropout according to the following perspectives: (i) social-economic (need to keep a job to live or to support a family); (ii) vocational (disappointment with a wrong course choice); and (iii) academic (failure in initial courses, poor educational background, difficulties with professor/student relationships or with colleagues).

Before starting to expose EDM related works, we assume that it is necessary to define some concepts used in this paper. We adopted the terminology defined by the Brazilian Department of Education to organize our research. It is assumed that *undergraduate course* (undergraduate programs) refers to the entire program (or curricula) of studies required to achieve a higher university degree. The concept *course* refers to a unit of instruction offered during the semester or academic year to build the undergraduate course. Some works consider the concept *dropout* when a student abandons the course [26,27]; in this paper *dropout* is used to classify the student who fails to complete, abandons or withdraws from one's undergraduate course.

2.2 Educational Data Mining

Baker and Yacef [12] presented EDM as a recent research area describing it as concerned with developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data. Educational systems are evolving and many technologies are being used to track people as they learn. Nowadays, several technologies are used to handle enormous amounts of data, spread across many databases. EDM comes to support educators, students, educational managers, governments and society to take advantage of such knowledge [12,13,14,15].

Many works that apply EDM techniques to predict student performance are focused on improving the quality of the course [22,23,24]. In other cases, monitoring student performance in a given course is critical to avoid problems in other courses of the curriculum [23]. The research made to predict student performance in courses varies a lot due to the different levels of student motivation, different attitudes about learning, and different responses to specific instructional practices. More, the amount of data available to perform such predictions could change significantly: it can be a small amount of data collected during a particular course, such as the number of assignments, intermediary and final student grades [24], or huge amounts of data stored, from student interaction with the e-learning environments [15,22,25].

Kotsiantis et al. [26] focused on the prediction of student dropout in undergraduate course at the Hellenic Open University. They analyzed 354 students' records collected for the module "Introduction to Informatics". The data used related to the students´ personal records and the results of the assignments. Such work compared six data mining techniques. The authors asserted that Naive Bayes algorithm was the most appropriate and the conclusions could be widespread in the majority of distance education curricula in the university. In [27], the Electrical Engineering undergraduate course at the Eindhoven University of Technology was evaluated. This study examined the data collect from 2000 to 2009 from no more than 648 students in the first year of their undergraduate course. The authors considered the results of several data mining techniques using three datasets (pre-university data, university grades, and both datasets). The overall result showed the decision tree algorithm as the more suitable to solve such problem. Pal [28] predicted the dropout of Engineering students in the first academic year. The author used demographic records (gender, age, and so on) collected during the admission of student and pre-university information (high school marks, family income of students, parents' qualification). The author used four data mining algorithms, but the emphasis was on the analysis of student data.

As far as we are concerned, the related works share similarities: (i) identifying and comparing algorithm performance to find the most appropriate EDM techniques, or (ii) identifying the relevant attributes associated with the problem of dropout. Those studies can be divided according to the type of data used. Some works use past time-invariant student records (demographic and pre-university student data). Others include time-varying data, and in this case the data is collected as the course progresses (scores, test results, activities) [22].

The key differences in such works to our approach are related to the data used to execute the prediction and focus on the problem. We focus on predicting student dropout in undergraduate courses with the use of EDM techniques. And we adopt time-varying data: semester courses with a final grade, GPA (Grade Point Average) and derived attributes. Besides, we are able to continuously monitor the progress of the students in each academic semester.

3. ARCHICTETURE

Our research presents an architecture named WAVE. It was designed to keep a continuous record of student academic progress based on the EDM requirements stated by [12,13,14]. The target is to predict and identify students that are at risk of dropping out of the educational system. Our work was conceived to enhance the Academic Management Software System (AMSS) of the public federal universities in Brazil. Currently, these systems are legacy software that store huge amounts of academic data, but lack the functionalities to perform a systematic processing of student information from the perspective of evaluating student performance or predict those who are at risk of dropping out.

The architecture aims to extend the AMSS, and we saw that the approach is less risky and more economical than implementing a new system. WAVE is presented in Figure 1. Due to the nature of the AMSS, one of the most appropriate approaches to add new analytical functionalities while retaining the existing systems is to adopt a multi-tier architecture model. WAVE was designed as a three-tier architecture consisting of a *presentation tier*, *an application tier*, and a *data tier*.

The presentation tier is the topmost level of the architecture. It is responsible for handling the interaction with the user. The educational manager can access the system directly by using a Graphical User Interface (not presented in this paper)

The application tier manages the key functionalities of the architecture and the rules for processing data. The tier consisted of three components: *ETL*, *EDM*, and *Knowledge Management Repository (KMR)*.

The ETL is responsible for loading and preparing the records for the next components. The initial preprocessing data phase is time-consuming and was improved by using the Pentaho Data Integration (Kettle) an open source product for Extraction, Transformation and Loading (ETL). The ETL component has iterative steps to run data extraction, data cleaning, data selection, and data transformation. The extraction stage establishes the connection with the AMSS or other student data source and extracts the data. The transformation stage of the ETL performs data cleaning, data selection, and data transformation. The loading stage is used to format and load the data produced by the previous stages to be used by the EDM component and knowledge management repository.

The EDM is the second component of the application tier. In order to predict student performance, we evaluated several classification algorithms. We evaluated a set of well-known public data mining algorithms available from the Weka library [29]. WAVE adopted the classification techniques as they have the best performance according to our previous experiments [30, 31]. The following classification algorithms were considered: Naïve Bayes (NB), Multilayer Perceptron (MLP), Support Vector Machine (SVM) with polynomial kernel and RBF kernel and Decision Tree (DT).

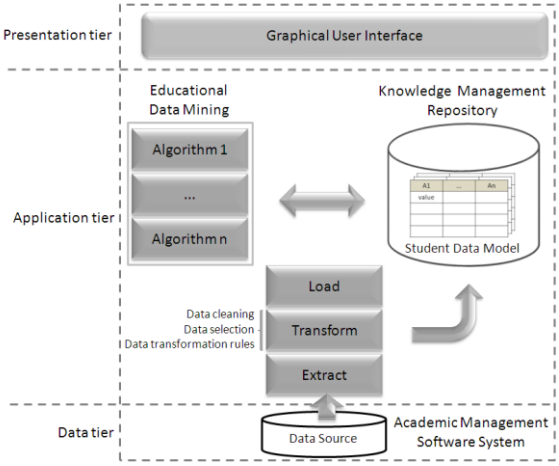

Figure 1. The three-tiers of WAVE architecture.

The third component of the tier is the KMR. It consists of a collection of student data items. The repository is populated at the end of each semester after final exams, the updated data extracted from the AMSS. Thus, new data set is added to the KMR. Each data set has different attributes that vary according to the number of semesters completed by the student. We define a Data Model based on student features (attributes) in each semester. The last tier in our architecture is the data tier, consisting of the database provided by the AMSS.

4. EDM EXPERIMENTS

In this paper, we investigated six undergraduate courses (Law, Pharmacy, Physics, Civil, Mechanical and Production Engineering). The courses were chosen because they belong to distinct departments, and have different a student profile to them. Besides, they have different numbers of entrances by year/semester, have different historical levels of dropout rates and also different responses to specific instructional practices. We emphasize that all the data was previously anonymized and was provided by the official the AMSS at UFRJ. The AMSS keeps registers about students enrolled in courses, stores the grading scheme, registers the examination results, calculates the GPA (Grade Point Average), cumulative CGPA (Cumulative Grade Point Average) and the number of credits per semester.

4.1 Definition of EDM Attributes and Data Model

In this work, we have evaluated a large amount of data about individual student´s academic progress in each semester from 1994 to 2010. Academic progress assumes a student's successful completion academic requirements towards a degree.

The data model used by WAVE consists of the following attributes: 1) *student id* is the key to identifying the student in the datasets; 2) *semester id* is used to identify the data in the semester;

3) *number of courses* is used to store the number of courses in which a given student is enrolled in the semester; 4) *number of courses approved* in the semester; 5) *average grade* of the courses approved; 6) *number of courses* in which a student failed due to absence or low grades; 7) *number of course* a student failed in due to a low grade; 8) *first semester status*; 9) *value* of the GPA; 10) *second semester status*, it is a class label attribute. In this work, we considered two status values for *semester status*. The first class value is assigned when a student failed (*no progress*) in the semester. The second class value refers to a student who has had *progress* in the semester. The progress level could be configured according to the educational manager. The rule of progress level could be a number of courses approved (one or two as a minimum), number of credits, or GPA. For instance, a student has progress when one's performance in a given semester is above the minimum value defined by the manager.

At UFRJ, every freshman is automatically enrolled in the initial set of introductory courses (about 6 or 7 courses). For instance, in Engineering the introductory courses are Physics, Experimental Physics, Calculus, Chemistry, Computer Science and an introductory course in a specific Engineering. Student performance in those courses is critical. Thus, three attributes were used to store the values for each introductory course: course id, the numeric grade and course status (pass, low grade, absence fail and low grade).

Table 1. Attributes of the Data Model

Attribute	Value/Type
student id	id code (string)
semester id	id code (string)
number of courses in which a given student is enrolled in the semester	{1 to n} (numeric)
number of courses approved in the semester, course status pass	{0 to n} (numeric)
average grade of the approved courses in the semester	{0 to 100} (numeric)
number of courses with status equal to absence fail and low grade (AFLG)	{0 to n} (numeric)
number of courses with status equal to low grade (LG)	{0 to n} (numeric)
first semester status (value defined by the rule of progress level)	{no progress, progress} (string)
value of GPA in first semester	{0 to n} (numeric)
introductory course id (about 6 or 7 courses)	id codes (string)
numeric grade for each introductory course	{0 to 100} (numeric)
student status in the course for each introductory course: pass (PA), low grade (LG), absence fail and low grade (AFLG)	{PA,LG,AFLG}(string)
second semester status *class label attribute*	{no progress, progress} (string) *{?,?} predicted value*

We defined around 31 attributes as shown in Table 1. Those attributes correspond to the data model of the Wave architecture.

4.2 Definition of EDM Algorithms

WAVE consists of a set of traditional classification algorithms: Naïve Bayes (NB), Multilayer Perceptron (MLP), Support Vector Machine with polynomial kernel (SVM^1) and RBF kernel (SVM^2), and Decision Table (DT). We adjusted our architecture to use a composition of classification algorithms and identify the number of classifiers with the same prediction, to give more reliability and reinforce the overall results.

Table 2 shows an example of the layout of a report that can be analyzed by the educational manager. The report shows individually, in lines, the results for each student. The columns show the prediction results of each classifier. Value '1' indicates *progress* and '0' *no progress*. The last column (?) shows the result of the composition. Value '1' is used when the majority of the classifier prediction was *progress* and '0' when the majority indicated *no progress*. However, the educational manager has the autonomy to interpret the results.

Table 2. Layout of report with prediction of the classifiers for *n* students

Student ID	NB	MLP	SVM^1	SVM^2	DT	?
Student 1	0	0	0	0	1	0
…	…	…	…	…	…	…
Student n	1	1	1	1	1	1

In our preview work [30,31], we evaluated and compared the behaviour of several of the above mentioned classifiers. All of them have good results when applied to the data model. Therefore, in this work, we present a deeper analysis in the application of the Naïve Bayes algorithms, as the Naïve Bayes classifier presented the highest true positive rate for all datasets used in the experiments.

4.3 Description of the EDM Experiment

In this section, we evaluate the Naïve Bayes algorithm used in the architecture, considering the student's datasets of the first semester in the year to obtain a prediction for the next semester. To check the functionalities of the architecture we reproduced the real conditions of the academic requirements to analyze the performance of students after their first academic semester.

As proposed in the architecture, the KMR consists of a different student data model. Indeed, the KMR holds the datasets that are used to train the classifiers. In this experiment, for each course, we merged the student´s dataset of even years (1994-1, …, 2008-1) which were used as a training sets (see Table 3). Those datasets have the last attribute, second semester status (class label), completed with the current value for the student performance in the second semester of the year. We considered the following rule of progress to assign a value to the class label (*semester status)*: *progress* when a student finished the semester with a minimum one course pass and *no progress* when all course results had low grades or absent fail.

Tables 3, 4, and 5 show the quantitative information on the datasets used in the experiments. The tables use the following keys: Civil Engineering (CE), Mechanical Engineering (ME), Production Engineering (PE), Pharmacy (Pha), Physics (Phy), and Law (Law). Table 3 shows the number of instances for the training set, for each undergraduate course, those numbers representing the total number of students enrolled in even years

(1994-1, …, 2008-1). In addition, we identified a number of students in the two classes.

Table 3. Number of students in the two classes in the training sets for each undergraduate course

Class	CE	ME	PE	Pha	Phy	Law
No progress	81 (17%)	58 (14%)	25 (8%)	63 (10%)	326 (52%)	335 (16%)
Progress	408 (83%)	358 (86%)	290 (92%)	548 (90%)	297 (48%)	1785 (84%)

We also defined the same data model for the test sets, and selected the student data from the first academic semester in the period of (1995-1, 1997-1,…, 2009-1). As a result, we built 48 datasets using the proposed attribute for the data model. In brief, differently from a single training set we specified 48 datasets to be used as individual test data files, those files being referred to as supplied test sets [25]. Table 4 shows the distribution of students in the undergraduate course and the entrance in each year/semester.

Table 4. Number of the students in each undergraduate course in each year/semester in the test sets

Y/S	CE	ME	PE	Pha	Phy	Law
1995-1	53	60	34	72	84	266
1997-1	68	50	41	70	87	285
1999-1	63	47	38	69	55	285
2001-1	49	45	51	71	34	272
2003-1	70	60	42	78	75	264
2005-1	71	65	41	74	61	273
2007-1	67	66	43	73	74	253
2009-1	60	61	39	96	37	262

Table 5 shows the percentage of students distributed in the two classes; those values are based on the test sets. For example, 19% of students in Civil Engineering belong to the *no progress* class. Those are the students with a higher probability of dropping out of their courses.

Table 5. Percentage of students in both classes in the test sets

Class	CE	ME	PE	Pha	Phy	Law
No progress	0.19	0.19	0.08	0.14	0.51	0.16
Progress	0.81	0.81	0.92	0.86	0.49	0.84

If we compare Tables 3 and 5, the percentage of students per classes is similar for training and test sets in all undergraduate courses.

4.4 Evaluation of EDM Experimental Results

In this section, we evaluate the experimental results of the Naïve Bayes algorithm used in the architecture. In turn, we defined the positive class as *no progress* and negative class as *progress*. In this context, several classification metrics can be used to indicate the performance of the classifier. For instance, (i) *accuracy* is one of most frequent value calculated to show the percentage of correctly classified instances; (ii) the statistical measures calculated from the confusion matrix: *true positive rate* (TP) is the proportion of positive cases (*no progress*) correctly classified and *true negative rate* (TN) measures the proportion of negative cases (*progress*) that are correctly classified; and (iii) *Kappa* (Cohen's Kappa) takes into account the similarities between the classes [32]. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. A Kappa value greater than zero indicates the classifier is doing better than chance.

Table 6 shows the accuracy of the Naïve Bayes EDM algorithms.

Table 6. Naïve Bayes accuracy for each test set

Y/S	CE	ME	PE	Pha	Phy	Law
1995-1	0.83	0.80	0.91	0.81	0.79	0.92
1997-1	0.90	0.84	0.95	0.80	0.70	0.90
1999-1	0.87	0.83	0.97	0.84	0.85	0.92
2001-1	0.88	0.76	0.90	0.86	0.76	0.88
2003-1	0.87	0.90	0.93	0.78	0.88	0.89
2005-1	0.87	0.88	1.00	0.86	0.80	0.89
2007-1	0.79	0.89	0.93	0.85	0.85	0.94
2009-1	0.90	0.92	1.00	0.93	0.78	0.87

In our datasets, most of the students belong to the *progress* class. However, our prior interest is to measure the performance of the classifier when it predicts the *no progress* class. The accuracy gives a measure for both classes of correct instances classified. The accuracy has some disadvantages to estimate the performance of the algorithms. Table 6 shows a high accuracy value for all the dataset; it does not give significant information to evaluate the model created by Naïve Bayes classifier to predict the performance of students of the minority class (*no progress*).

Tables 7 and 8 show the performance of the Naïve Bayes algorithm using measures calculated from the confusion matrix. Table 7 shows rates of true positive class (*no progress*), meaning that our approach scores highly in identifying imminent dropouts and giving feedback for students and educational managers just after the end of first academic semester.

Table 7. True positive rates (*no progress*)

Y/S	CE	ME	PE	Pha	Phy	Law
1995-1	83.33	83.33	66.67	68.75	75.00	88.89
1997-1	93.33	55.56	100.00	100.00	78.05	67.39
1999-1	63.64	66.67	100.00	72.73	90.91	60.87
2001-1	78.95	57.14	75.00	12.50	60.00	70.18
2003-1	83.33	66.67	100.00	42.86	85.29	70.49
2005-1	100.00	83.33	100.00	66.67	75.00	74.36
2007-1	72.73	66.67	66.67	83.33	78.38	80.33
2009-1	71.43	81.82	100.00	81.82	73.91	55.56
Average	**80.84**	**70.15**	**88.54**	**66.08**	**77.07**	**71.01**

Table 8 shows the rates for negative class (*progress*), based on the values of the rates, where we can assert that the architecture presents good results to identify the students with improved performance.

Comparing Tables 7 and 8 we can see that there are no great significant differences between the results for two classes of students. We pointed that the number of students in the *progress* class is higher than the *no progress* class (Table 4) and there are some students that drop out without a predictable reason. In this case, the error of the classifier could be minimized.

Table 8. True negative rates (*Progress*)

Y/S	CE	ME	PE	Pha	Phy	Law
1995-1	82.93	79.63	96.43	83.93	84.38	92.05
1997-1	88.68	90.24	94.74	77.78	63.04	94.14
1999-1	92.31	96.15	96.88	86.21	81.82	100.00
2001-1	93.33	83.87	91.49	95.24	100.00	93.02
2003-1	87.93	92.59	92.50	85.94	90.24	95.07
2005-1	85.48	88.68	100.00	89.23	86.21	91.88
2007-1	80.36	92.98	95.00	85.07	91.89	98.44
2009-1	92.45	94.00	100.00	94.12	85.71	92.48
Average	**87.93**	**89.77**	**95.88**	**87.19**	**85.41**	**94.64**

Table 9 shows Kappa values for each dataset. Kappa is another measure for evaluating the performance of the classifier. According to Table 9, all Kappa values are above 0, meaning that the prediction model is tuned.

Table 9. Kappa values

Y/S	CE	ME	PE	Pha	Phy	Law
1995-1	0.58	0.36	0.68	0.48	0.57	0.64
1997-1	0.73	0.46	0.72	0.41	0.41	0.62
1999-1	0.56	0.65	0.91	0.50	0.71	0.49
2001-1	0.74	0.42	0.50	0.10	0.55	0.64
2003-1	0.61	0.52	0.54	0.28	0.76	0.69
2005-1	0.60	0.64	1.00	0.47	0.61	0.60
2007-1	0.41	0.57	0.53	0.41	0.70	0.83
2009-1	0.57	0.73	1.00	0.68	0.56	0.47

5. CONCLUSION

High dropout rates have many undesired consequences not only for the students but also for society. Students differ from one another in many ways, including the type of instruction to which they respond and the nature of knowledge they pursue. The main contribution of this work is the evaluation of an architecture based on classifiers and the data model which was developed to be coupled to a legacy AMSS. The architecture provides additional support for educational managers to monitor the progress of the students in each academic semester and to predict the ones who are at risk of dropping out. Our proposal to use only time-varying student data was effective enough to predict student performance. As far as we are concerned our architecture is one of the first to present the benefit of using sole student data stored in the AMSS, no external data being necessary to point students that are at risk of dropping out.

Our previous works [30,31] analyzed several classifier algorithms. However, in this work we focused on measuring the performance of the Naïve Bayes classifier. The classification accuracies found in the undergraduate courses were very similar, the accuracy ranging between 0.70 and 1.00. Such values are considered good for real data. The Naïve Bayes classifier had high values for a true positive rate for all 6 undergraduate courses analyzed. This means that, in the majority of the undergraduate courses, the combination of the Naïve Bayes classifier and student data model can (correctly) predict over 70% students who will not be approved (no progress) in any course in the second semester, meaning a high probability of dropout from the educational system. The true negative rate for the class (*progress*) is above 85% for the classifier, and in practice our architecture was able to identify and predict almost all students who will have a good performance and probably will not leave University. Finally, an important goal of this paper was to show that the set of attributes used to predict the performance in each academic semester is effective, concise and can easily be extracted from the existing AMSS.

6. REFERENCES

[1] INEP. 2011. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Resumo Técnico do Censo da Educação Superior. http://portal.inep.gov.br/web/censo-da-educacao-superior/resumos-tecnicos.

[2] MEC. 1997. Ministério da Educação e Cultura. *Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas.* http://www.udesc.br/arquivos/id_submenu/102/diplomacao.pdf.

[3] QS World University Rankings. 2013. http://www.topuniversities.com/qs-world-university-rankings.

[4] OECD. 2012. Brazil, in *Education at a Glance 2012: OECD Indicators*, OECD Publishing. http://dx.doi.org/10.1787/eag-2012-42-en.

[5] Barroso, M. F. and Falcão, E. B. M. 2004. *Evasão Universitária: O Caso do Instituto de Física da UFRJ*. IX Encontro Nacional de Pesquisa em Ensino de Física.

[6] Soares, I. S. 2006. *Evasão, retenção e orientação acadêmica: UFRJ - Engenharia de Produção – Estudo de Caso*. In: Anais do XXXIV COBENGE - Congresso Brasileiro de Ensino de Engenharia. Ed. Universidade de Passo Fundo, Passo Fundo - RS.

[7] Silva Filho, R.L.L., Motejunas, P.R., Hipólito, O., Lobo and M. B. C. M. 2007. *A Evasão no Ensino Superior Brasileiro*. Cadernos de Pesquisa. v. 37, n. 132, p. 641-659, Sept/Dec. São Paulo: Fundação Carlos Chagas.

[8] Andriola, W. 2009. *Fatores Associados à Evasão Discente na Universidade Federal do Ceará (UFC) de Acordo com as Opiniões de Docentes e de Coordenadores de Cursos*. REICE: Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 7(4), 342-355.

[9] Dias, E. C. M., Theóphilo, C. R. and Lopes, M. A. S. 2010. *Evasão no ensino superior: estudo dos fatores causadores da evasão no curso de Ciências Contábeis da Universidade Estadual de Montes Claros – UNIMONTES – MG*. In: Anais do Congresso USP de Iniciação Científica em Contabilidade. São Paulo: Êxito Editora.

[10] Lobo, M. B. C. M. 2011. *Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções*. Instituto Lobo/Lobo & Associados Consultoria.

[11] Márquez-Vera, C., Cano, A., Romero, C. and Ventura, S. 2013. *Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data*. Applied Intelligence, 1-16.

[12] Baker, R. S. J. D. and Yacef, K. 2009. *The state of educational data mining in 2009: A review and future visions*. Journal of Educational Data Mining 1.1, 3-17.

[13] Baker, R., Isotani, S. and Carvalho, A. 2011. *Mineração de Dados Educacionais: Oportunidades para o Brasil*. Revista Brasileira de Informática na Educação, 19(02), 03.

[14] Romero, C. and Ventura, S. 2010. Educational Data Mining: A Review of the State of the Art, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on , vol.40, no.6, 601-618,

[15] Romero, C., and Ventura, S. 2013. *Data Mining in Education*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, In Press. Volume 3, Issue 1, 12-27.

[16] Tinto, V. 1975. *Dropout from higher education: A theoretical synthesis of recent research*. Review of Educational Research, 45, 89-125.

[17] Tinto, V. 1993. *Leaving college: Rethinking the causes and cures of student attrition*. Chicago: University of Chicago Press.

[18] Terenzini, P. T. and Pascarella, E. T. 1977. *Voluntary freshman attrition and patterns of social and academic integration in a university: A test of a conceptual model*. Research in Higher Education, 6, 25-43.

[19] Gosman, E. J., Dandridge, B. A., Nettles, M. T., and Thoeny, A. R. 1983. *Predicting student progression: The influence of race and other student and institutional characteristics on college student performan*ce. Research in Higher Education, 18, 209-236.

[20] Pascarella, E. T. and Terenzini, P. T. 1991. *How college affects students*. San Francisco: Jossey-Bass.

[21] Astin, A. W. 1994. *Minorities in American higher education: Recent trends, current prospects and recommendations*. San Francisco: Jossey-Bass.

[22] Lykourentzou, I. Giannoukos, I., Nikolopoulos, V., Mpardis, G., and Loumos, V. 2009. *Dropout prediction in e-learning courses through the combination of machine learning techniques*. Computers & Education, Volume 53, Issue 3, (November, 2009), 950-965.

[23] Huang, S. 2011. *Predictive Modeling and Analysis of Student Academic Performance in an Engineering Dynamics Course*. Ph.D. Thesis, Utah State University, Logan, Utah, USA.

[24] Hamalainen, W. and Vinni, M. 2006. Comparison of machine learning methods for intelligent tutoring systems. in Proc. Int. Conf. Intell. Tutoring Syst., Taiwan, 525-534.

[25] Minaei-Bidgoli, B., Kortemeyer, G., and Punch, W. 2004. Association analysis for an online education system. Information Reuse and Integration. IRI 2004. Proceedings of the 2004 IEEE International Conference. (November, 2004), 504,509, 8-10.

[26] Kotsiantis, S., Pierrakeas, C., and Pintelas, P. 2003. *Preventing student dropout in distance learning using machine learning techniques*. KES, eds. V. Palade, R. Howlett & L. Jain, Springer, 2003. volume 2774 of Lecture Notes in Computer Science, pp. 267–274. 1087-6545.

[27] Dekker, G., Pechenizkiy, M., and Vleeshouwers, J. 2009. *Predicting Students Drop Out: A Case Study. In Proceedings of the International Conference on Educational Data Mining*. Cordoba, Spain, 41-50.

[28] Pal, S. 2012. *Mining educational data to reduce dropout rates of engineering students*. International Journal of Information Engineering and Electronic Business (IJIEEB), 4(2), 1.

[29] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

[30] Manhães, L.M.B., Cruz, S.M.S., and Zimbrão, G. 2014. The Impact of High Dropout Rates in a Large Public Federal Brazilian University: A Quantitative Approach Using Educational Data Mining. In: CSEDU, 2014, Barcelona, Spain, 6th International Conference on Computer Supported Education, 2014, 124-129.

[31] Manhães, L.M.B., Cruz, S.M.S., and Zimbrão, G. 2014. WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM. In: SAC 2014, 2014, Gyeongju, Korea. ILLE - Symposium of Applied Computing, 2014.

[32] Cohen, J. 1960. *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, 20 (1), 37-46.