A predictive model for identifying students with dropout profiles in online courses

Marcelo A. Santana
Institute of Computing
Federal University of Alagoas
marcelo.almeida@nti.ufal.br

Evandro B. Costa
Institute of Computing
Federal University of Alagoas
evandro@ic.ufal.br

Baldoino F. S. Neto
Institute of Computing
Federal University of Alagoas
baldoino@ic.ufal.br

Italo C. L. Silva
Institute of Computing
Federal University of Alagoas
italocarlo@nti.ufal.br

Joilson B. A. Rego
Institute of Computing
Federal University of Alagoas
jotarego@gmail.com

ABSTRACT

Online education often deals with the problem related to the high students' dropout rate during a course in many areas. There is huge amount of historical data about students in online courses. Hence, a relevant problem on this context is to examine those data, aiming at finding effective mechanisms to understand student profiles, identifying those students with characteristics to drop out at early stage in the course. In this paper, we address this problem by proposing predictive models to provide educational managers with the duty to identify students whom are in the dropout bound. Four classification algorithms with different classification methods were used during the evaluation, in order to find the model with the highest accuracy in prediction the profile of dropouts students. Data for model generation were obtained from two data sources available from University. The results showed the model generated by using SVM algorithm as the most accurate among those selected, with 92.03% of accuracy.

Keywords

Dropout, Distance Learning, Educational Data Mining, Learning Management Systems

1. INTRODUCTION

Every year, the registration marks in E-learning modality has increased considerably, in 2013, 15.733 courses were offered, in E-learning or semi-presence modality. Furthermore, the institutions are very optimistic, 82% of researched places, believe that the amount of registration marks will have a considerable expansion in 2015 [1], showing the E-learning evolution and its importance as a tool for citizen's formation. The Learning Management Systems (LMS) [15] can be considered one of factors that has had an important role for popularization of this learning modality [1].

Despite the rapid growth of online courses, there has also been rising concern over a number of problems. One issue in particular that is difficult to ignore is that these online courses also have high dropout rates. Specifically, in Brazil, in 2013, according with the latest Censo, published by the E-learning Brazilian Association (ABED), the dropout average was about 19,06% [1].

Beyond the hard task on identifying the students who can have possible risk of dropping out, the same dropout also brings a huge damage to current financial and social resources. Thus, the society also loses when they are poorly managed, once the student fills the vacancy but he gives up the course before the end.

Online education often deals with the problem related to the high students' dropout rate during a course in many areas. There is huge amount of historical data about students in online courses. Hence, a relevant problem on this context is to examine those data, aiming at finding effective mechanisms to understand student profiles, identifying those students with characteristics to drop out at early stage in the course.

In this paper, we address this problem by proposing predictive models to provide educational managers with the duty of identifying students who are in the dropout bound. This predictive model took in consideration academic elements related with their performance at the initial disciplines of the course. Data from System Information course at Federal University of Alagoas (UFAL) were used to build this model, which uses a very known LMS, called Moodle.

A tool to support the pre-processing phase was used in order to prepare data for application of Data Mining algorithms. The Pentaho Data Integration [2] tool covers the extraction areas, transformation and data load (ETL), making easier the archive generation in the compatible format with the data mining software adopted, called WEKA[5].

Therefore, for what was exposed above, it justifies the needing of an investment to develop efficient prediction methods, assessment and follow up of the students with dropout risk,

allowing a future scheduling and adoption of proactive measures aiming the decrease of the stated condition.

The rest of the paper is organized as follows. Section 2 presents some related work. Section 3 Environment for Construction of predictive model. Afterwards, we present the experiment settings in Section 4, and in Section 5 we discuss the results of the experiment. Section 6 presents some concluding remarks and directions of future work.

2. RELATED WORK

Several studies have been conducted in order to find out the reasons of high dropout indices in online courses. Among them, Xenos [18] makes a review of the Open University students enrolled in a computing course. In this studies, five acceptable reasons, that might have caused the dropout, were identified: Professional (62,1%), Academic (46%), Family (17,8%), Health Issues (9,5%), Personal Issues (8,9%). According to Barroso and Falcão (2004) [6] the motivational conditions to the dropout are classified in three groups: i) Economic - Impossibility of remaining in the course because of socio-economics issues; ii) Vocational - The student is not identified with the chosen course. iii) Institutional - Failure on initial disciplines, previous shortcomings of earlier contents, inadequacy with the learning methods.

Manhães et al.[14] present a novel architecture that uses EDM techniques to predict and identify those who are at dropout risk. The paper shows initial experimental results using real world data about of three undergraduate engineering courses of one the largest Brazilian public university. According to the experiments, the classifier Naive Bayes presented the highest true positive rate for all datasets used in the experiments.

A model for predicting students' performance levels is proposed by Erkan Er [9]. Three machine learning algorithms were employed: instance-based learning Classifier, Decision Tree and Naive Bayes. The overall goal of the study is to propose a method for accurate prediction of at-risk students in an online course. Specifically, data logs of LMS, called METU-Online, were used to identify at-risk students and successful students at various stages during the course. The experiment were realized in two phases: testing and training. These phases were conducted at three steps which correspond to different stages in a semester. At each step, the number of attributes in the dataset had been increased and all attributes were included at final stage. The important characteristic of the dataset was that it only contained time-varying attributes rather than time-invariant attributes such as gender or age. According to the author, these data did not have significant impact on overall results.

Dekker [8] in your paper presents a data mining case study demonstrating the effectiveness of several classification techniques and the cost-sensitive learning approach on the dataset from the Electrical Engineering department of Eindhoven University of Technology. Was compared two decision tree algorithms, a Bayesian classifier, a logistic model, a rule-based learner and the Random Forest. Was also considered the OneR classifier as a baseline and as an indicator of the predictive power of particular attributes. The experimental results show that rather simple classifiers give a useful result with accuracies between 75 and 80% that is hard to beat with other more sophisticated models. We demonstrated that cost-sensitive learning does help to bias classification errors towards preferring false positives to false negatives. We believe that the authors could get better results by making some adjustments to the parameters of the algorithms.

Jaroslav [7], aims to research to develop a method to classify students at risk of dropout throughout the course. Using personal data of students enriched with data related to social behaviours, Jaroklav uses dimensionality reduction techniques and various algorithms in order to find which of the best results managing to get the accuracy rates of up to 93.51%, however the best rates are presented at the end of the course. Whereas the goal is to identify early on dropout, the study would be more relevant if the best results were obtained results at the beginning of the course.

In summary, several studies investigating the application of EDM techniques to predict and identify students who are at risk dropout. However, those works share similarities: (i) identify and compare algorithm performance in order to find the most relevant EDM techniques to solve the problem or (ii) identify the relevant attributes associated with the problem. Some works use past time-invariant student records (demographic and pre-university student data). In this study, contribution to those presented in this section, makes the junction between two different systems, gathering a larger number of attributes, variables and time invariant. Besides being concerned with the identification and comparison of algorithms, identify the attributes of great relevance and solve the problem the predict in more antecedence the likely to dropout students.

3. ENVIRONMENT FOR CONSTRUCTION OF PREDICTIVE MODEL

This subsection presents an environment for construction for a predictive model for supporting educators in the task of identifying prospective students with dropout profiles in online courses. The environment is depicted in Figure 1.

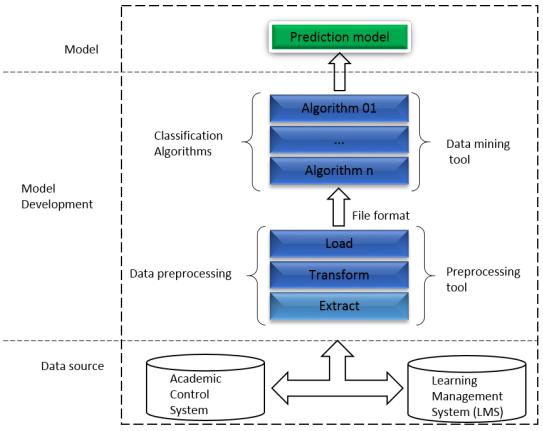

Figure 1: Environment for Construction of predictive model

The proposed environment in this work is composed by three layers: Data source, Model development and Model. The data sources are located in the first layer. Data about all

students enrolled at the University are stored in two data sources: The first one contains students' personal data, for example: age, gender, income, marital status and grades from the academic control system used by the University. Information related with frequency of access, participation, use of the tools available, and grades of students related the activities proposed within the environment are kept in second data source.

In the second layer, the pre-processing [11] activity over the data is initiated. Sequential steps are executed in this layer in order to prepare them to data mining process. In the original data some information can not be properly represented in a expected format by data mining algorithm, data redundancy or even data with some kind of noise. These problems can produce misleading results or make the algorithm execution becomes computationally more expensive.

This layer is divided into the following stages: data extraction, data cleaning, data transformation, data selection and the choice of algorithm that best fits the model. Just below, will be displayed briefly each step of this layer.

Data extraction: The extraction phase establishes the connection with the data source and performs the extraction of the data.

Data cleaning: This routine tries to fill missing values, smooth out noise while identifying outliers, and correct data inconsistencies.

Data transformation: In this step, data are transformed and consolidated into appropriate forms for mining by performing summary or aggregation operations. Sometimes, data transformation and consolidation are performed before the data selection process, particularly in the case of data warehousing. Data reduction may also be performed to obtain a smaller representation of the original data without sacrificing its integrity.

Data selection: In this step, relevant data to the analysis task are retrieved from the database.

Choice of algorithm: An algorithm to respond with quality in terms of accuracy, which has students elusive profile, was considered the algorithm that best applies to the model.

Finally, the last layer is the presentation of the model. This layer is able to post-processing the result obtained in the lower layer and presenting it to the end-user of a most understandable way.

4. EXPERIMENT SETTINGS

The main objective of this present research is to build a predictive model for supporting educators in the hard task of identifying prospective students with dropout profiles in online courses, using Educational Data Mining (EDM) techniques [16]. This section is organized as follows: Section 4.1 describes the issue which drives our assessment. Section 4.2 shows which data were selected for to the data group utilized in the experiment and which algorithms were chosen for data mining execution. Section 4.3 indicates the employed tools during the execution of experiment. Finally, Section 4.4 shows every step in experiment execution, including data consolidation, data preprocessing and algorithms execution.

4.1 Planning

The research question that we would like to answer is:

RQ.Is our predictive model able to early identify the students with dropout risk?

In order to answer this question, EDM techniques with four different classification methods were used, aiming to get a predictive model which answers us with quality in precise ways which students have a dropout profile, taking in consideration only data about the initial disciplines of a specified course.

4.2 Subject Selection

4.2.1 Data Selection

The Federal University of Alagoas offers graduation courses, postgraduate courses and E-learning courses. In the on line courses, there are more than 1800 registered students[4].

An E-learning course is usually partitioned in semesters, where different disciplines are taught along these semesters. Each semester usually has five disciplines per semester, and each discipline has a duration between five to seven weeks. Anonymous data, from the Information Systems E-learning course, were selected from this environment, relative to first semester in 2013. Data of one discipline (Algorithm and Data Structure I), chosen based on its relevance, were analysed. Such discipline has about 162 students enrolled.

4.2.2 Machine Learning Algorithms Selection

In this work to predict student dropouts, four machine learning algorithms were used, using different classification methods. The methods used were: simple probabilistic classifier based on the application of Bayes' theorem, decision tree, support vector's machine and multilayer neural network.

These techniques have been successfully applied to solve various classification problems and function in two phases: (i) training and (ii) testing phase. During the training phase each technique is presented with a set of example data pairs (X, Y), where X represents the input and Y the respective output of each pair [13]. In this study, Y can receive one of the following values, "approved" or "reproved", that corresponds the student situation in discipline.

4.3 Instrumentation

The Pentaho Data Integration [2] tool was chosen to realize all preprocessing steps on selected data. Pentaho is a open-source software, developed in Java, which covers extraction areas, transform and load of the data [2], making easier the creation of an model able to : (i) extract information from data sources, (ii) attributes selection, (iii) data discretization and (iv) file generation in a compatible format with the data mining software.

For execution of selected classification algorithms (see Section 4.2.2), the data mining tool Weka was selected. Such algorithms are implemented on Weka software as NaiveBayes (NB), J48 (AD), SMO (SVM), MultilayerPerceptron (RN)

[17] respectively. Weka is a software of open code which contains a machine learning algorithms group to data's mining task [5].

Some features were taken in consideration for Weka [10] adoption, such as: ease of acquisition, facility and availability to directly download from the developer page with no operation cost; Attendance of several algorithms versions set in data mining and availability of statistical resources to compare results among algorithms.

4.4 Operation

The evaluation of experiment was executed on HP Probook 2.6 GHz Core-I5 with 8Gb of memory, running Windows 8.1.

4.4.1 Data's Preprocessing

Real-world data tend to be dirty, incomplete, and inconsistent. Data preprocessing techniques can improve data quality, thereby helping to improve the accuracy and efficiency of the subsequent mining process [11].

Currently, the data is spread in two main data sources: LMS Moodle, utilized by the University as assistance on E-learning teaching, including data which show the access frequency, student's participation using the available tools, as well as the student's success level related to proposed activities. Meanwhile, student's personal files as age, sex, marital status, salary and disciplines grades are kept in the Academic Control System (ACS), which is a Software designed to keep the academic control of the whole University [4].

Aiming to reunite a major data group and work only with relevant data to the research question that we want to answer, we decided to perform consolidation of these two data source in a unique major data source, keeping their integrity and ensuring that only relevant information will be used during data mining algorithms execution.

Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can help improve the accuracy and speed of the data mining process [11].

To maintain the integrity and reliability between data, a mandatory attribute, with unique value and present between in both data sources, was chosen. Thus, the CPF attribute was chosen to make data unification between the two selected data sources, once it permits the unique identification among selected students.

In order to facilitate algorithms execution and comprehension of results,predicting the dropout in an early stage of the study. In order to achieve a high rate of accuracy and minimum of false negatives, i.e. students that have not been recognized to be in danger of dropout. Some attributes were transformed, as we can seen below:

- The corresponding attributes related with discipline grades were discretized in a five-group-value (A,B,C,D e E), depending on the discipline's achieved grades. The student with a grade higher or equal 9, was allocated for "A" group. Those ones who had their grades between 8,99 and 7 were allocated for "B" group. the "C" students are those that had a grade between 6,99 and 5, and those who had grades under 5,99 stayed at "D" group and finally those that doesn't have a grade associated were allocated in "E" group.

- Every student was labelled as approved or reproved based on the situation informed by the academics registers. The final score of each discipline is composed by two tests, if the student did not succeed in obtaining the minimum average, he will be leaded to the final reassessment and final test.

- In the "City" attribute, some inconsistencies were found, where different data about the same city were registered in database. For instance, the instances of **Ouro Branco** and **Ouro Branco/AL** are related to same city. This problem was totally solved, with application of techniques for grouping attributes.

- The attribute "age" had to be calculated. For this, the student's birth date, registered in database, was taken in consideration.

When all the attributes were used the accuracy was low. That is why we utilized feature selection methods to reduce the dimensionality of the student data extracted from dataset. We improved the pre-processing method the data.

In order to preserve reliability of attributes for classification after the reduction. We use InfoGainAttributeEval algorithm that builds a rank of the best attributes considering the extent of information gain based on the concept of entropy.

After this procedure, we reduced the set of attributes from 17 to 13 most relevant. The list of the refined set of attributes in relevance ordercan be found in Table 1.

Table 1: Selected Attributes

Attributes	Description
AB1	First Evaluation Grade
Blog	Post count and blog view
Forum	Post count and forum views
Access	Access Count in LMS
Assign	Sent files count e viewed
City	City
Message	Count of sent messages
Wiki	Post count and wiki view
Glossary	Post count and glossary view
Civil status	Civil status
Gender	Gender
Salary	Salary
Status	Status on discipline

Taking in consideration that the main objective is to predict student's final situation with the earlier advance as possible inside the given discipline, to this study we will only use data until the moment of the first test.

The Figure 2 presents all the executed stages, during the preprocessing phase, in order to generate a compatible file with the mining software.

4.4.2 Algorithms Execution

The k-fold method was applied to make a assessment the model generalization capacity, with k=10 (10-fold cross validation). The cross validation method, consists in splitting of the model in k subgroups mutually exclusive and with the same size, from these subgroups, one subgroup is selected for test and the remaining k-1's are utilized for training. The average error rate of each training subgroup can be used as an estimate of the classifier's error rate. When Weka implements the cross validation, it trains the classifier k times to calculate the average error rate and finally, leads the build classifier back utilizing the model as a training group. Thus, the average error rate provides a better solution in terms of classifier's error accuracy reliability [12].

In order to get the best results of the algorithms without losing generalization, some parameters of SVM algorithms were adjusted.

The first parameter was set the parameter "C". This parameter is for the soft margin cost function, which controls the influence of each individual support vector; this process involves trading error penalty for stability [3].

The default kernel used by Weka tool is the polynomial we changed to the Gaussian setting the parameters Gamma. Gamma is the free parameter of the Gaussian radial basis function [3].

After several adjustments to the values of the two parameters mentioned above, which showed the best results in term of accuracy and lower false positive rate, was C = 9.0 and Gamma = 0.06 parameter.

For comparison of results related to selected algorithms, we used Weka Experiment Environment (WEE). The WEE allows the selection of one or more algorithms available in the tool as well as analyse the results, in order to identify, if a classifier is, statistically, better than the other. In this experiment, the cross validation method, with the parameter "k=10" [5], is used in order to calculate the difference on the results in each one of the algorithms related to a chosen standard algorithm (baseline).

5. RESULTS AND DISCUSSIONS

In this section, the results of the experiment, described in Section 4, are analyzed.

The WEE tool calculated the average accuracy of each classifier. Table 2 shows the result of each algorithms execution. The accuracy represents the percentage of the test group instance which are correctly classified by the model built during training phases. If the built model has a high accuracy, the classifier is treated as efficient and can be put into production [11].

Comparing the results among the four algorithms, we can verify that the accuracy oscillates around 85.5 to 92.03%. Furthermore, a classifier which has a high error rate to false positives is not suitable to our solution. In this case, we have considered the algorithm which has the lower false positive rates.

Table 2: Accuracy and rates

Classifiers	NB	AD	SVM	RN
Accuracy	85.50	86.46	92.03	90.86
True Positives	0.76	0.77	0.88	0.85
False Negatives	0.24	0.23	0.12	0.15
True Negatives	0.89	0.91	0.94	0.93
False Positives	0.11	0.09	0.06	0.07

As we can see on table 2 the algorithm SVM presented a low false positive rate and better accuracy. Therefore, only the best algorithm was considered to our solution. The Naive Bayes classifier had the worst result in terms of accuracy and a high false positive rate. The other ones had an error average of 8%, and then, we end up with 8% of the students with dropout risk not so correctly classified.

5.1 Research Question

As can be seen in table 2, in our experiment, the SVM algorithm obtained 92% of accuracy. According to Han J. *et al.* [11] if the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known. Thus, the results are pointing to the viability of model able to early identify a possible student's dropout, based on their failures in the initial disciplines.

5.2 Statistical Significance Comparision

We often need compare different learning schemes on the same problem to see which is the better one to use. This is a job for a statistical device known as the t-test, or Student's t-test. A more sensitive version of the t-test known as a paired t-test it was used. [17]. Using this value and desired significance level (5%), consequently one can say that these classifiers with a certain degree of confidence (100 - significance level) are significantly different or not. By using the t-test paired in the four algorithms, performed via Weka analysis tool, observed that the SVM algorithm is significantly respectful of others.

5.3 Threats to validity

The experiment has taken in consideration data from the Information System course and the Data Structure Algorithm discipline. However, the aforementioned discipline was chosen, based on its importance in the context of Information System course.

6. CONCLUSION AND FUTURE WORK

Understand the reasons behind the dropout in E-learning education and identify in which aspects can be improved is a challenge to the E-learning. One factor, which has been pointed as influencer of students' dropout, is the academic element related with their performance at the initial disciplines of the course.

This research has addressed dropout problem by proposing predictive models to provide educational managers with the duty to identify students whom are in the dropout bound.

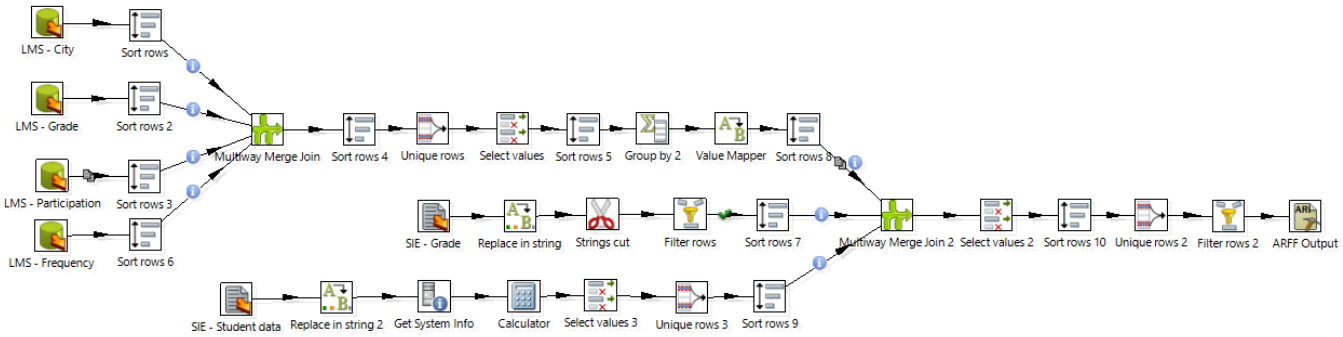

Figure 2: Steps Data Preprocessing

The adopted approach allowed us to perform predictions at an initial discipline phase. The preliminaries results has shown that prediction model to identify students with dropout profiles is feasible. These predictions can be very useful to educators, supporting them in developing special activities for these potential students, during the teaching-learning process.

As an immediate future work, some outstanding points still should be regarded to the study's improvement, as apply the same model in different institution databases with different teaching methods and courses, including new factors related to dropout as: professional, vocational and family data, execute some settings in algorithms' parameters in order to have the best achievements. Furthermore, a integrated software to LMS, to provide this feedback to educators, will be developed using this built model.

7. REFERENCES

[1] Abed - E-learning Brazilian Association. http://www.abed.org.br/. Accessed December 2014.

[2] Pentaho - Pentaho Data Integration. http://www.pentaho.com/. Accessed January 2015.

[3] SVM - support vector machines (svms). http://www.svms.org/parameters/. Accessed December 2014.

[4] UFAL - Federal University of Alagoas. http://www.ufal.edu.br/. Accessed January 2015.

[5] Weka - the University of Waikato. http://www.cs.waikato.ac.nz/ml/weka/. Accessed January 2015.

[6] M. F. Barroso and E. B. Falcao. University dropout: the case of ufrj physics institute. *IX National Meeting of Research in Physics Teaching*, 2004.

[7] J. Bayer, H. Bydzovská, J. Géryk, T. Obsívac, and L. Popelínsky. Predicting drop-out from social behaviour of students. In A. H. M. Y. Kalina Yacef, Osmar Zaiane and J. Stamper, editors, *Proceedings of the 5th International Conference on Educational Data Mining - EDM 2012*, pages 103–109, Greece, 2012.

[8] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers. Predicting students drop out: A case study. In T. Barnes, M. C. Desmarais, C. Romero, and S. Ventura, editors, *EDM*, pages 41–50, 2009.

[9] E. Er. Identifying at-risk students using machine learning techniques: A case study with is 100. In *International Journal of Machine Learning and Computing*, pages 476–481, Singapore, 2012. IACSIT Press.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[11] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[12] S. B. Kotsiantis, C. Pierrakeas, and P. E. Pintelas. Preventing student dropout in distance learning using machine learning techniques. In V. Palade, R. J. Howlett, and L. C. Jain, editors, *KES*, volume 2774 of *Lecture Notes in Computer Science*, pages 267–274. Springer, 2003.

[13] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Comput. Educ.*, 53(3):950–965, Nov. 2009.

[14] L. M. B. Manhães, S. M. S. da Cruz, and G. Zimbrão. Wave: An architecture for predicting dropout in undergraduate courses using edm. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, pages 243–247, New York, NY, USA, 2014. ACM.

[15] M. Pretorius and J. van Biljon. Learning management systems: Ict skills, usability and learnability. *Interactive Technology and Smart Education*, 7(1):30–43, 2010.

[16] C. Romero and S. Ventura. Educational data mining: A review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618, Nov 2010.

[17] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[18] M. Xenos, C. Pierrakeas, and P. Pintelas. A survey on student dropout rates and dropout causes concerning the students in the course of informatics of the Hellenic Open University. *Computers Education*, 39(4):361 – 377, 2002.