Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance

Sandra Merchán Rubiano[1] and Jorge Duarte García[1]

[1]Grupo investigación OSIRIS, Universidad El Bosque, Bogotá, Colombia
{smerchanr,jduarteg}@unbosque.edu.co

Abstract. This paper presents and analyzes the experience of applying certain data mining methods and techniques on 932 Systems Engineering students' data, from El Bosque University in Bogotá, Colombia; effort which has been pursued in order to construct a predictive model for students' academic performance. Previous works were reviewed, related with predictive model construction within academic environments using decision trees, artificial neural networks and other classification techniques. As an iterative discovery and learning process, the experience is analyzed according to the results obtained in each of the process' iterations. Each obtained result is evaluated regarding the results that are expected, the data's input and output characterization, what theory dictates and the pertinence of the model obtained in terms of prediction accuracy. Said pertinence is evaluated taking into account particular details about the population studied, and the specific needs manifested by the institution, such as the accompaniment of students along their learning process, and the taking of timely decisions in order to prevent academic risk and desertion. Lastly, some recommendations and thoughts are laid out for the future development of this work, and for other researchers working on similar studies.

Keywords: Data mining · Predictive modeling · Academic risk prevention · Academic performance · Educational data mining

1 Introduction

Student desertion -as a phenomenon which impacts heavily on educational processes- has been widely studied and modeled, thus identifying its possible causes in order to find strategies to treat it and prevent it. In Colombia, this phenomenon has been documented via the SPADIES [1], which is an information system developed by the National Education Ministry; it was created to gather and centralize all the necessary information to establish the factors which cause student desertion across the nation. Overall, the Country has progressed in the characterization and diagnosis of this phenomenon, and findings include the fact that the greatest rate of desertion takes place in the first semester of studies. Other interesting results reported by the Ministry states that 20 of the Country's departments (subdivisions with certain degree of autonomy) exhibit desertion rates exceeding 40 %; also the Country's annual desertion average rate is around 10 % for university students.

To counteract these results, the Ministry has modeled student desertion and has designed policies for its diagnosis, monitoring and prevention. This strategy open the way for subsequent publications [2] [3], in which academic risk has been recognized as one of the main factors that influences student desertion in higher education scenarios. Notwithstanding the above, in this context academic risk has not been studied as a phenomenon in itself, but all efforts have been rather directed towards preventing desertion as a direct consequence of this phenomenon. Previous studies have focused on preventing the student's desertion, underestimating the importance of generating preventive strategies that cope with its causes; particularly strategies that deal with academic risk as it is one of desertion's main causes.

Specific programs, projects and strategies have been implemented at El Bosque University in order to follow and accompany the students' learning process. One of the most relevant strategies, regarding academic risk tracking, has to do with the University's Online Academic Management System - SALA [4], where alerts are generated according to the grades the students obtain each period and their historical academic data. These

alerts classify students in three risk-related levels: low, medium and high risk, the later being a trigger for a student's accompaniment and intervention by the University's student support program (PAE).

Despite the importance of said strategies, the University lacks a mechanism which acts preventively; one able to identify the causes for academic risk, and act upon these causes before the risk has occurred. This would allow the institution to take timely decisions and design better strategies to prevent academic risk as a phenomenon, and as a consequence, diminish the students' probability of desertion.

Bearing in mind that the University requires a preventive and timely approach towards dealing with student desertion, this work proposes, applies and evaluates certain data mining methods and techniques, in order to take advantage of the information (which had served no similar purpose) that the institution gathers from its students. This practice is known as Educational Data Mining (EDM) [5], which derives from Data Mining, the semiautomatic process of discovering hidden patterns in data [6]. EDM has proven being a useful tool for making predictions in several scenarios [7, 8, 9, 10], and thus will be used to generate a predictive model based on students' academic and demographic data. This model should provide useful insights about the causes for academic risk, in order to empower the Institution in the taking of better decisions and the implementation of stronger strategies regarding student desertion.

The stages of the data mining process, aimed at constructing the predictive model for students' academic performance, are presented and analyzed in the first part of this paper. In the second part, applied techniques and methods, as well as the results are analyzed. The presented conclusions are drawn from these analyses.

2 Methodological Development

To build a predictive model for students' performance based on data mining, it's necessary to develop four important stages [6] [10] [11] . Firstly the data that will feed the mining process must be extracted and prepared. Secondly the data mining process itself must be implemented in an iterative fashion: each iteration consisting of data preprocessing, algorithm execution, results -rules- analysis, accuracy interpretation and testing [12]. Thirdly the predictive model formulation must be done by analyzing, selecting and defining the rule set which allows for proper academic performance prediction, regarding the institutional context and the proposed objectives that frame this work. Lastly the predictive model must be validated by applying it to different datasets with similar characteristics to the one utilized during the mining process, and obtaining favorable results. This work presents the development of the first two stages, analyzing the data mining process and its applicability in the building of predictive models regarding educational environments. Fig. 1 illustrates the method.

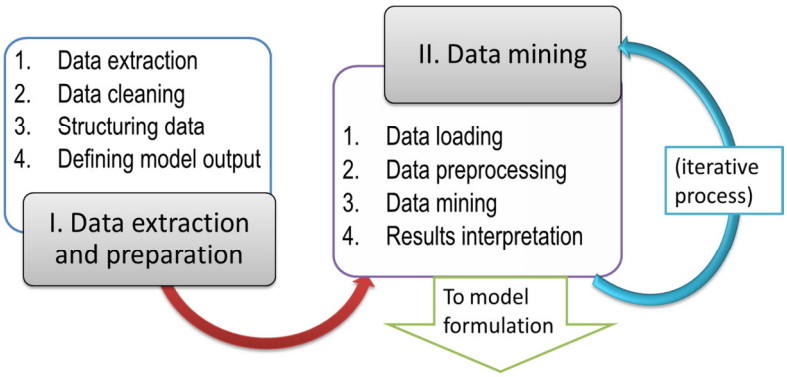

Fig.1. Methodological process aimed to construct the model

2.1 Data Extraction and Preparation

Data extraction. Being the primary objective to generate a predictive model -out of academic and demographic data- for the students of the Systems Engineering program, the students' information must necessarily fall into these two categories. Also, this data was selected in accordance with results presented in the SPADIES reports giving relevance to socioeconomic variables known to influence desertion rates.

Data mining requires a significant amount of data in order to provide meaningful results. For this study the anonymous records of 932 students were obtained from the University's IT department, which were delivered as four separate spreadsheet files. Performing a process similar to data warehousing, once the data received was loaded into a relational database engine (MySQL), it was normalized to the Third Normal Form [13]; this enabled properly query the information in a consistent and organized way. Further normalization was not considered necessary, due to the static nature of the database.

Data cleaning. Being the data then correctly structured, a process defined as *data cleaning* [6] was applied (through SQL queries) in order to check for consistency errors in the data and standardize attribute values. This was a necessary step because a large portion of the data obtained was manually typed into web forms by students at some point in their enrollment. This means typographic errors were present in the data, as well as diverse input formats for a selection of fields, such as home address and city.

Having substantially improved the dataset's quality by making it *cleaner* and more consistent, the data was randomly split into two sets of almost equal size: a set to be mined in order to generate the predictive model, and a control set to verify that the model generated works as expected. This partitioning of the dataset is done in order to assess the classifier model's error rate (product of the data mining process), using a dataset that played no part in the construction of the classifier [6]. This verification portion of the dataset will also allow to evaluate different learning schemes on a fresh dataset to determine which one performs better [6].

Structuring data. After defining which data was to be used for mining, the whole dataset was structured as a set of instances [6] Each instance represents a different student, and each one has values for all the attributes selected for the study. Said attributes are presented in Table 1.

Table 1. Defined instances and attributes.

Students' Academic Data Attributes	Students' Demographic Data Attributes
Average ICFES test score, ICFES[3] subject test scores (biology, math, philosophy, physics, history, chemistry, language, geography and English), students' high school, year and period of enrollment, whether the student has received academic incentives, class schedule type (day or night classes), First semester grades (6 subjects) and Second semester grades (6 subjects)	Age, gender, residence distance to University (calculated from home address), City of origin, Social Class (ranked 1-6 according to national legislation), Marital status, job status, whether the student registers a mother, mother's education, mother's job, mother's city of origin, whether the student registers a father, father's job father's education, father's city of origin, number of siblings.

[3] The ICFES test is a test given to graduating high school students to measure their performance along different knowledge areas, and if often required by Universities in Colombia as part of the admittance process. It is comparable to the SAT Test for college admissions in the USA.

Defining model output. Being these attributes the information that characterizes the population under study (as input variables), the output variable of our model would be the academic performance, as this is what the model aims to predict. It was necessary to define academic performance according its definition in the institutional context, that is, directly related to academic risk. In accordance to the University's PAE (students' support program) and SALA (institutional academic information system), academic risk appears when a student's grade weighted average[4] is below 3.3/5. So, guided by the above and by the average performance statistically calculated in the sampled data, the output variable academic performance is defined in three types as shown in Table 2.

Table 2. Definition of academic performance as output variable

Type of academic performance	Weighted average
Outstanding Performance	from 4.0 to 5.0
Average performance	from 3.3 to 3.9
Academic Risk	below 3.3

The final structure of all instances is presented in Table 3.

Table 3. Final structure of instances

Id	Demographic Data	Academic Data	Total grade average
	...	...	

Note that in this case, the academic and demographic data are represented by single columns for the sake of visualization.

Finally both data sets (mining and verification), which have been restructured as seen above, were exported as separate text files. Within each file, each instance corresponds to a line of text in which its attributes are separated by commas, and missing values are represented by the '?' character instead of a blank space (requirement for ARFF file conversion) [6]

2.2 Data mining

To undertake the task of data mining the *Waikato Environment for Knowledge Analysis* was chosen. The *WEKA* workbench is a data mining software platform developed at the University of Waikato in new Zealand [6]. It offers an extensive collection of state-of-the-art machine learning algorithms, data preprocessing and visualization tools, all wrapped up in a comprehensive and easy to use common interface. It is licensed and distributed under the GNU General Public License, it's available for free, and runs on all major computer operating systems (Windows, Linux and Mac OS).

As defined previously, data mining as an iterative process comprises a cycle of three main activities: data preprocessing, data mining and results interpretation. Due to *WEKA*'s resourcefulness the fulfillment of such tasks are achieved with great ease.

[4] Weighted average is Σ[(Sc*Sg)]/[Σ(Ct)], where Sc: a subject's amount of credits, Sg:a subject's grade, Ct:total credits taken by a student

Throughout each data mining iteration the following procedures were performed, and on each one, variations of each procedure were implemented -as permitted by the *WEKA Explorer* view-:

- **Data loading:** the data was loaded into *WEKA* so it could be mined; this was done by assembling an ARFF file using the text file containing the mining dataset.
- **Data preprocessing:** the dataset was preprocessed according to which data mining tasks were to be performed over it. *WEKA* allows the filtering and transformation of the loaded dataset in several ways both manually, or using several data mining and sorting algorithms; also, attributes from the loaded dataset can be selectively removed.
- **Data mining:** machine learning algorithms were executed over the dataset. Classification algorithms -which find classification rules that *classify* a particular instance according to the value of its attributes- were chosen to perform the task at hand.
- **Results interpretation:** the resulting output from these algorithms was recorded as individual text files; information present in these -such as classification rates, classification rules and prediction error measurements- was analyzed and interpreted according to bibliography and previous results if applicable. Also, if further iteration was considered necessary, plans for it were laid out in accordance of the results obtained.

Data loading, preprocessing and mining. The product results from the data mining process are presented according to the steps previously described, along 4 executed iterations. Table 4, describes how the first two data mining stages were performed (data loading and preprocessing). Similarly, Table 5 exposes the results obtained by executing algorithms J48, PART and Ridor, regarding the amount of rules obtained and the precision of classifications achieved. These three algorithms were chosen due to their similarity in purpose (classification rule induction), but also for their particularly unique ways of achieving this task. Throughout all iterations 10-fold cross-validation was defined as a standard evaluation technique, to ensure the integrity of predictions and results [6]

The iterative process was considered fulfilled as results obtained from the fourth iteration exhibited sufficient and meaningful outcomes in terms of precision and accuracy -measured as classification rates-. Thus, a verification process involving the attempt to classify untrained data (control dataset) was performed in order to assess the data mining process. Aside from certain expected imprecisions in the validation results, it was considered as successful.

Table 4. Description of data loading and preprocessing

	Iteration #1	Iteration #2	Iteration #3	Iteration #4
Data loading	467 instances 42 Attributes	467 instances 42 Attributes	467 instances 42 Attributes	467 instances 42 Attributes
Data preprocessing	None	First year subject grades were removed **Resulting data:** 467 instances 31 Attributes	First year subject grades were removed 'Noisy' attributes were removed **Resulting data:** 467 Instances 24 Attributes	First year subject grades were removed 'Noisy' attributes were removed Filter applied: remove misclassified instances using a J48 decision tree **Resulting data:** 179 instances 24 Attributes

Table 5. Description of data mining process, according the algorithms executed

	Iteration #1	Iteration #2	Iteration #3	Iteration #4
J48	10 rules found 231 correctly classified instances 141 ignored instances	0 rules found 135 correctly classified instances 141 ignored instances	150 rules found 129 correctly classified instances 141 ignored instances	58 rules found 153 correctly classified instances 0 ignored instances
PART	4 rules found 204 correctly classified instances 141 ignored instances	4 rules found 137 correctly classified instances 141 ignored instances	56 rules found 115 correctly classified instances 141 ignored instances	16 rules found 149 correctly classified instances 0 ignored instances
Ridor	23 rules found 233 correctly classified instances 141 ignored instances	14 rules found 127 correctly classified instances 141 ignored instances	15 rules found 120 correctly classified instances 141 ignored instances	7 rules found 100 correctly classified instances 0 ignored instances

Table 6 shows the results obtained throughout each iteration, concerning classification accuracy for each of the output variables defined for academic performance.

Table 6. Accuracy rates obtained by each algorithm

	Iteration # 1	Iteration # 2	Iteration # 3	Iteration # 4
J48	Risk: 81% Average: 62.70% Outstanding: 83.30%	Risk: 27.8% Average: 42.3% Outstanding: 0 %	Risk: 43.2% Average: 42.7% Outstanding: 18.2 %	Risk: 86.8% Average: 85.3% Outstanding: 76.9 %
Part	Risk: 64,7% Average: 57.1% Outstanding: 78.8%	Risk: 16.7% Average: 42.6% Outstanding: 0 %	Risk: 34.8% Average: 40.3% Outstanding: 25.4%	Risk: 85.4% Average: 83.3% Outstanding: 66.7 %
Ridor	Risk: 72,2% Average: 71.8% Outstanding: 69.6%	Risk: 47.9% Average: 40.2% Outstanding: 22.7 %	Risk: 44.5% Average: 37.5% Outstanding: 22.1 %	Risk: 64.9% Average: 55.3% Outstanding: 25%

Results Interpretation

About each iteration. Given the results product from the first iteration, it was observed that the rules obtained were composed exclusively by first-year subject performance averages; such results seem logical because academic performance is a direct consequence of subject average grades. This also relates to the favorable precision results that were obtained. Nevertheless this knowledge is not useful regarding the study's objective, for it is redundant and somehow obvious; therefore such attributes cannot be included as input variables for the predictive model, and were removed in the next iteration.

When the results for the second iteration were obtained, several issues arose. Firstly no rules were obtained by the J48 decision tree classifier. Secondly the decision list generated by PART relied heavily on the fact that a student registers an empty high school name to generate rules -which is rather suspicious-. Lastly RIDOR performed surprisingly well in contrast with the other two algorithms for it included several relevant attributes to generate rules. Sadly it did not perform as well in terms of precision, for it exhibits an inadmissible amount of error -as did the other two-. It was observed that the classification rate diminished considerably in contrast with the last iteration, however this makes sense after removing redundant attributes: the algorithms had less available correlated data to induce rules from. Bearing in mind these observations, certain attributes were chosen to be removed for the next iteration due to their 'noisiness'.

For the third iteration the following attributes were discarded from the dataset because of their large dispersion -which causes 'noise' in the data-: students' high school, year and period of enrollment, whether the student has received academic incentives, mother's job, mother's city of origin, father's job and father's city of origin. As a result of this, the amount of outputted rules rises considerably. Even though rules are composed of relevant attributes, precision is still very low. This is interpreted as an indicator for rules that do not perform too well on the training dataset, which might be caused by instances that are not providing useful data to the algorithms, that in turn allow for proper predictions to the made. Also, it was found that the 141 instances that had been ignored throughout this and previous iterations correspond to instances without defined values for its class attribute. Concerning these observations, it was suspected that a significant portion of the 326 instances that are being utilized to train the algorithms were interfering with proper rule induction mechanisms inside the algorithms. Therefore more elaborate preprocessing was encouraged for the next iteration.

Concerning the fourth and final iteration, a preprocessing filter (*removeMisclassified*) was applied in order to remove misclassified instances product of a classification's algorithm execution. The objective of such task is to enable a better rule induction by using only instances that provide useful data for performance prediction. The chosen algorithm to perform this task was the J48 decision tree learner, because of it's good average performance in past iterations and other tests performed on the data. Having done this the number of training instances was greatly reduced (in around 60%), however the algorithms were able extract more valuable and meaningful knowledge from this smaller dataset, while maintaining favorable and tolerable precision levels. By further analyzing the rules product from this iteration, it was noted that they were more generalized than previous exemplars, but managed to describe the dataset in a more accurate fashion. It was also noted that classification rates improved for J48 and PART classifiers. Due to this performance increases and precision prevalence, a verification exercise was considered pertinent to assure these results apply when 'fresh' data is evaluated.

During the proposed verification exercise, the model obtained from iteration number 4 for the J48 algorithm was loaded into *WEKA*, as was the control dataset (465 instances). Then the control data was adjusted in the same way that the training data was preprocessed in iterations number 2 and 3 (so that the control data's structure matches the training data's). By executing the loaded model over the new data, the control dataset was evaluated using previously defined classification rules, in order to obtain new classification predictions and precision measurements; these new results reveal a slightly lower classification rate for the new data, as well as a small increment in error. In spite of this, the results obtained can be considered as profitable because our previously trained algorithm (J48) managed to correctly classify around 78% of the new instances evaluated.

Classification Rules Summary. The rules discovered by the J48 decision tree learner can be summarized as follows. Students whose social class is 2 or undefined are classified within academic risk, while students whose social class is higher than 3 are classified depending on other variables such as marital status. Regarding the latter, whereas students who are 'single', depends on additional attributes such as mother's education and student's gender to be classified, students who are currently 'in union' -but not married- are classified within academic risk by default.

Attention is drawn to the surfacing of attributes such as individual ICFES scores, parent's education, number of siblings and whether the student registers a mother within the rule set, as influencing factors for academic performance. More interestingly it was found that if a student is not married and his/her social class is 4 it is immediately classified within risk. Also for social classes 5 and 6 (higher classes) most rules defaulted a classification within outstanding performance.
Decision lists generated by the PART algorithm seem to agree with the J48 decision tree, in that students whose social class is 2 are almost defaultly classified within academic risk. Nonetheless students are classified within academic risk if their social class is 3 and they are single or if their social status is 3 and their age is above 24. Concerning a student's social class being 3, greater variations arise depending on the his/her marital status, parents' education, gender, age and certain individual ICFES scores. Furthermore students whose social class is 4 and have no siblings are immediately classified within outstanding performance.

Regarding the Ripple-down rule learner -Ridor-, it was discovered that if a student's social class is 3 most of the times it is classified within average performance. However if the student's father's education is undefined and his/her's english ICFES score is below 50, it is classified within academic risk. Surprisingly for social classes 1, 4 and 5 the defaulting classification is within outstanding performance, but it lies within average performance if the student's age is below 26.5; if a student's age is below the latter and his/her mother's education is undefined the student is classified within academic risk. The default classification for a student's social class being 2 is within average performance.

Rules interpretation. Concerning the classification rules described previously, certain particular features about the dataset's attributes emerge. For instance, social class seems to be a highly decisive variable regarding characterization of rules that accurately describe the data utilized in this study. It recurrently appears as the first classifying attribute of tree-generated resulting rule sets. Also its defaulting classification for a value of 2 coincides in both the J48 and PART algorithms; this makes sense because PART obtains rules from partial decision trees built using J48. In the case of marital status, it seems to be the second most important classifying attribute; most of the resulting rules revolve around its value 'single' and frequently encompass other attributes such as parent's education. Classifications within academic risk are present for all values of this attribute except for 'married' students. Lastly, attributes such as age, gender and selected individual ICFES scores (like biology, philosophy and languages) seem to play a relevant role in conjunction with social class and marital status. Relating to Ridor's results, students were classified within risk when their age was below 26 alongside with other attribute values such as mother's education and ICFES scores.

The pertinence of the rules obtained, in relationship with the social and academic context of the studied population is adequate. Particularly the students' socioeconomical stratum has been perceived as a highly influential variable in their academic performance, as their obtained ICFES scores. Nonetheless, other variables appear as influential, such as age and gender towards which more attention should be drawn.

The obtained results gave way to some important recommendations. Several lessons were learned regarding the data used in this study; primarily the data procured for this kind of work must be as clean and consistent as possible. In case that extensive data preparation is required, no stone must be left unturned; failing in correctly standardizing attribute values can heavily impact data mining precision. Similarly, the input variables defined must be meaningful in order to predict class attributes; irrelevant or obvious correlation between these might result in useless predictions. Concerning the methodological development, time allocated for data preparation and preprocessing must not be underestimated; it could potentially consume great amounts of time [6].

3 Conclusions

In accordance with the strategies implemented during the data mining process, it was found that careful data preprocessing can and will have drastic impact over the data mining results; such decisions -like attribute removal, or data discretization - should not be taken lightly. Furthermore, the careful assessment and removal misclassified and ignored instances can profoundly improve success rates when seeking to generate a predictive model.

The rule sets product of the data mining process are found useful to characterize the students' academic performance, for they allow to identify which demographic and academic attributes hold influence over the studied phenomenon. Additionally, having identified these attributes, it is possible to determine if a given student will incur in academic risk along his or her first year of studies. All of the above-mentioned sets base for students' accompaniment and academic risk prevention strategies generation. Bearing all results and conclusions, the formulation for a predictive model based on an iterative data mining process gains feasibility [10].

Regarding the algorithms' performance, not only must their classification accuracy and rule quantity be evaluated, but also - and specially- their quality and pertinence [6]. By the way in which Ridor structures its results and performs generalizations regarding them, this algorithm -in general- did not allow to obtain rules that involved valuable attributes for the consecution of the study's objective; for those rules that might have been relevant, their precision values were significantly low.

It should be noted that the obtained results are strongly related to the institution´s characteristics and the objectives of the model to be built. To make useful this study in other institutions is necessary to consider other comparison and evaluation techniques of the algorithms, like statistical tests, applied to large volumes of data[14]

It is necessary to propose and perform studies that consider the academic risk phenomenon as part of academic desertion. Latter phenomenon is frequently studied but separated from risk factors, using techniques and methods such as the analyzed in this research.

Based on the results obtained, the next step to be developed in the study, it is the one which frames this work, is to make further use of the rules obtained by the data mining process, by executing them over several other data sets that belong to the same educational institution. This should be done always minding the pertinence, precision and applicability of said rules in order to formulate a predictive model, which can be validated and proven useful within the institutional context.

Having proved this model as useful, it should be implemented as a software tool, which aids in the decision-making processes, and help to generate stronger strategies for academic accompaniment and risk prevention. Lastly, it is recommended to the academic community to extend the theoretical framework of the Educational Data Mining, adding more studies that identify and analyze the most appropriate techniques according to the diversity of phenomena presented in the post university context [15][16].

4 References

[1] Ministerio de Educación Nacional ¿Qué es el SPADIES? - Sistemas información. Ministerio de Educacion Nacional - Sistemas de Información.
http://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-254648.html.

[2] Guzman R. C, Duran M. D, Franco G. J et al.: Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención. Imprenta Nacional de Colombia, Bogotá (2009)

[3] Velez W C, Burgos M. G, Angulo M et al.: Deserción estudiantil en la educación superior colombiana. Elementos para su diagnóstico y tratamiento, Bogotá (2008)
[4] Universidad El Bosque Sistema de Gestión académica en Línea. SALA. Sistema de Gestión Académica en Línea. SALA. http://artemisa.unbosque.edu.co/serviciosacademicos/consulta/facultades/consultafacultadesv2.htm.
[5] Baker R Educational Data Mining.: Predict the future, change the future iTunes U Video. (2012)
[6] Witten IH, Frank E, Hall MA.: DATA MINING: Practical Machine Learning Tools and Techniques, Third edn. Morgan Kaufmann, Burlington, MA, (2011)
[7] Brijesh Kumar Bhardwaj, Saurabh Pal.: Data Mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security 9(4), 136 (2011)
[8] Bhise RB, Thorat SS, Supekar AK.: Importance of Data Mining in Higher Education System. IOSRJournal of Humanities And Social Science 6(6), 18-21 (2013)
[9] Sen B, Ucar E .: Evaluating the achievements of computer engineering department of distance education students with data mining methods. Procedia Technology 1:262-267. doi:10.1016/j.protcy.2012.02.053. (2012)
[10] Vera C.M, Morales C.R, Soto S.V .: Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. Revista Iberoamericana de Tecnologías del/da Aprendizaje/Aprendizagem 7,3 (2012)
[11] Karina Gibert, Miquel Sànchez-Marrè, Víctor Codina .: Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation. In: 2010 International Congress on Environmental Modelling and Software Modelling for Environment's Sake Fifth Biennial Meeting, David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova, Ottawa, Canada (2010)
[12] Piatetsky-Shapiro G, Workshop.: Knowledge discovery in databases. AAAI Press Menlo Park, Cal, (1993)
[13] Watson TJ .: Further normalization of the data base relational model. IBM Research Center (1971)
[14] Davidson I, Tayi G .: Data preparation using data quality matrices for classification mining. European Journal of Operational Research 197(2):764-772. doi:10.1016/j.ejor.2008.07.019. (2009)
[15] Stenvall J, Syväjärvi A .: Data Mining in Public and Private Sectors, 1 edn. Information Science Reference (Isr), US. (2010)
[16] Manpreet Singh Bhullar, Amritpal Kaur.: Use of Data Mining in Education Sector. Lecture Notes in Engineering and Computer Science 2200(1):513-516. (2012).