Educational Data Mining with Focus on Dropout Rates

Rosangela Villwock, Andressa Appio and Aldioni Adaiani Andreta,

Western Paraná State University (UNIOESTE), Brazil

Summary
High dropout rates in higher education may be motivated by social, economic or personal factors and it is a reality in the Mathematics Major at *Universidade Estadual do Oeste do Paraná - Cascavel*. Thus, the aim of this study was to identify the factors that may have been influencing this reality through Data Mining (Classification). The study investigated the academic background and socioeconomic factors. Higher dropout rates were identified in the first year of this Major in which the course 'Differential and Integral Calculus I' was the one that most contributed for that. As for the socioeconomic factor, full-time working necessity was the one that was the most responsible for dropouts. Although trivial, the rules found endorse what by socializing with academics involved in the research was expected. It is hoped that this work may contribute to the elaboration of policies aimed at meeting the needs of the students, therefore ensuring their permanency at the University.

Key words:
Process of Knowledge Discovery in Databases; Classification; Socioeconomic Patterns; Patterns of Courses.

1. Introduction

Dropout in higher education is a global issue which affects the results of education systems. The losses provoked by students who start a Major but do not finish it point to social, academic and economic wastes. In the public sector, they represent public funds invested with no proper return [1].

The rupture of students from academic courses in public institutions may occur in indirect ways, due to the incompatibility between hours of study and hours of work for example [2, 3].

According to [4] the causes for dropout in higher education may come from a serious impediment or difficulty to a lack of motivation linked to the academic Major or to social individual context or due to the environment.

The problem of dropout in higher education is highly relevant and the permanency of students in Higher Education Institutions (HEI) depends on the pedagogical support offer by those, however, it is observed that many institutions are not prepared to offer this kind of support [5].

A research carried out on a group formed by all HEI in Brazil by [5] poses that between the years 2000 and 2005 the average number of drop-out was of 22% reaching 12% in public institutions and 26% in private institutions. According to the authors, the number of institutions that have regular institutional programs to prevent dropout with action planning, monitoring of results and collecting of data of successful experiences are very few.

A study of the researcher Oscar Hipólito of *Instituto Lobo* for the Development of Education, Science and Technology, based on the numbers of the census of the Ministry of Education point that in 2009 financial losses in higher education reached approximately R$ 9 billion. The calculation used is an estimative and tends to increase since education involves other costs rather than just that [6].

There are two sides to dropout. These sides reflect the student's own decision or a combination of social, economic and personal factors, as well as the precocious necessity to ingress the labor market, difficulties from adverse conditions of higher education curriculum, faculty and institution organization [7].

A broad research on dropout and its causes from student's point of view was carried out by [2]. The research took place in a Higher Education Institution (HEI) from 2000 to 2003. Some of the problems observed in the research were lack of vocational orientation, student's immaturity, successive failing grades, financial difficulties, lack of work perspective, lack of affective bounds in the university, forceful ingression in university (family imposition), unplanned marriage and pregnancy.

In one of his studies, [4] points out to a misconception in the young people thinking, that is, in not assessing the difference between the actions of ingressing a major course and actually take it, that is, conclude it. That happens because often they are not aware of the limitations offered by a university major.

Another study carried out by [8] on data from 61 Public Institutions of Higher Education (Federal and State Universities) which represented 77.2% of the public education sector at that time already indicated a high number of dropouts. Thus, one can come to knowledge that dropout in higher education is not a current issue, but instead, it has been discussed for many years.

Dropout phenomenon in higher education cannot be assessed in an isolate manner; it involves pedagogic, economic, administrative, psychological, social and political issues, among others [5]. According to [9] the study on dropout is related to the discussion on the quality of education. The author proposes the application of

Manuscript received March 5, 2015
Manuscript revised March 20, 2015

institutional assessment in order to reach advancements because this practice is closely associated to the course's curriculum and to the performance of students and professors.

According to [4] causes for dropout may be related to the rigidity of the Major curriculum, that is, to the courses in itself – number of credits, course load and mainly student's lack of comprehension which leads to negative results in their assignments. Brazilian Ministry o Education - MEC- in the curricular guidance [10] mentions that the rigidity of the curriculum is the main cause for high dropout rates and low percentage of graduated students. As stated in the curricular guidance, higher education institutions should promote ways of learning that may contribute to the decrease of dropout rates as for an example the organization of the courses in modules. Furthermore, they should also implement Undergraduate Research Mentorship Programs in which the student would be able to develop his own creativity and critical analysis.

The curricular guidance [10] also states that undergraduate courses curriculum should become more flexible in order to facilitate student's life, especially of those who work, since the number of students from this group who leave university without graduating are quite high.

According to [4] evaluating student's behavior may help to explain the phenomenon of dropout and consequently contribute to the search for the solution for this problem. The author also affirms that dropout is a problem that is difficult to measure. Student's transcript held in the university database among other data concerning the students, curriculum and courses may reveal a behavior pattern in the academic life. Through the complete transcript, one may analyze how many times a student attended a specific class, which classes held the higher number of dropout and student's permanency in the institution before dropout , among others [4].

According to [11] many researches show that students drop out courses mostly not by choice but by finding difficulties in being productive. In a study carried out by the same author about some of the factors that contributed to dropout in UnB (University of Brasilia) during the period between 2002 and 2006 it was concluded that failing grades in some courses and lack of good productivity in general are the greatest responsible for dropouts.

A study in order to identify factors that influenced dropouts in undergraduate Majors in the Federal University of Rio de Janeiro was conducted by [12]. The authors used techniques of data mining and observed that the students that are actively enrolled in courses are the ones that show a regular behavior along the course, they enroll for a high number of classes, have better grades than the ones who drop out, but lower than that of the ones who conclude the course.

The aim of this study was to identify through the process of knowledge discovery in databases socioeconomic patterns and patterns of courses that may contribute to the student's decision of dropping out the Mathematics Major at *Universidade Estadual do Oeste do* Paraná – *UNIOESTE*.

This paper is organized as follows: section 2 presents the descriptions of the KDD Process, data mining task and the technique used in this study; section 3 shows the work methodology; section 4 presents the results and discussions followed by the conclusion.

2. KDD Process

The process of Knowledge Discovery in Database – KDD is a non-trivial discovery process of valid, new, useful and accessible patterns. The main advantage of the discovery process is that no hypotheses are needed and knowledge is extracted from the data without previous knowledge [13].

KDD is related to the broad process of discovering information in a database in which there is an emphasis on a high-level application of the particular Data Mining (DM) method. While the DM step is characterized by the extraction of patterns hidden in the data, the whole KDD process is broader and includes all the processing (data selection, pre-processing and transformation) that is needed for this to occur, making it possible to evaluate and interpret the results that were obtained after DM techniques were used [13].

The KDD process is a set of continuous activities that include five steps: Data Selection, Pre-processing, Formatting, Data Mining and Interpretation. The process starts by understanding the application domain and the targets that must be reached. Then, a selection can be drawn from these data so that one may work with the data that are of interest. The pre-processing step is the one in which missing or inconsistent data are analyzed and treated. During the formatting step data are prepared so Data Mining can be used, for instance, to map categorical data among numerical data or to use methods to reduce dimensions in the data [13].

Advancing along the process, there is the Data Mining step, the main one in KDD process, in which several methods can be used to extract information that are then presented to the last step, the interpretation, where knowledge is acquired. If results are not satisfactory, the whole process may be fed back, changing some of the information that may be reprocessed in the previous steps [13].

The main purpose of the KDD process is to obtain knowledge hidden in data that may be useful for decision-making by using methods, algorithms and techniques from different scientific areas. According to [14] these include Statistics, Artificial Intelligence, Machine Learning and Pattern Recognition.

According to [13] Data Mining tasks are predictive and descriptive. The predictive ones use some variables to forecast unknown or future values of other variables while the descriptive ones find patterns to describe the data. The main tasks of Data Mining are related to pattern Classification, Association and Clustering. In this work, the Data Mining task used is the Classification.

2.1 Classification

According to [4] Classification is a process of search of sets of models or functions which may identify and describe classes or concepts. A classification technique is a systematic approach to the construction of classification models through an input database [14]. Examples of this are the decision tree, rule-based classifiers, artificial neural network, supporting vector machine and simple Naive Bayes classifier, among others.

Each patter contains a set of attributes and one of them is named class. The aim of Classification is to find a model for predicting the class as a function of other attributes [14]. According to [15] the aim of Classification is to build up through a classification database containing pre-classified objects (which classes are known) a model that is capable of classifying automatically new objects (which classes are unknown) according to their features.

Each technique implies a learning strategy to identify a model which may be appropriate to the relation between the features and the label of the classes of the input data. The model generated by the learning algorithm must adapt well to the input data and correctly predict the label of the classes of records that it has never seen before [14].

As mentioned by [14], a general approach to resolve the problems in classification consist in, first, constructing a classification model using a set of training that are made of records which labels are known and must be provided; second, in applying the classification model to the test that is made of records with labels of unknown classes (or known, however not given).

One of the ways to evaluate the performance of a classification model is based in counting the records of tests given correctly or incorrectly by the model. These counting is tabulated in a table known as confusion grid. This grid provides the necessary information to determine how well a classification model is executed [14].

Resuming these information in a single number would ease the comparison between the performance of different models. This could be done by using a performance improvement measuring form such as the precision, or, in an equivalent form, the performance of a model can be expressed by its error rates. Most of the classification algorithms look for models that reach a greater precision or, in the same way, lower rates of errors when applied to a set of tests [14].

Among the methods that are usually used in assessing the performance of a classifier there is the Cross validation, which is an alternative used when the dataset is small. In this approach the dataset is divided in K partitions. During each run one of the partitions is chosen for test while the others are used in training. This procedure is repeated k times in a way that each record is used a number of times for training and one single time for the test. The total error is obtained by the sum of the errors of every k execution [14].

In this work we opted to use the Decision Tree technique because it is a method of easy application and interpretation.

The decision tree method is a very adequate one when the aim is to generate rules that may be easily understood, explained and brought to natural language [16]. The decision tree has three types of nodes: a root node that does not have input edges and zero or more output edges; internal nodes, each one with exactly one input edge and two or more output edges; leaf nodes or terminals, each one with exactly one input edge and no output edge [14].

Each route in the tree (from root to leaf) correspond to a classification rule. In the decision tree each node must be associated to an attribute that is the most informative among the attributes not yet considered in the path from the root [16].

In order to generate a decision tree with a high range of prediction it is necessary to choose correctly the features that will be used as test in the clustering of cases. These features must generate a tree with a number of subsets that is as lower as possible in a way that each leaf on the tree may contain a significant number of cases [16].

For best clarifying the criteria that conduces to the choices of attributes it is necessary to study two concepts: **entropy** and **information gain**. Entropy is the means that indicate the homogeneity of the examples contained in a set of data. The lower the entropy, the more informative the attribute will be. The information gain is the measure that indicates how much a specific attribute will separate the learning examples according to their function/aim (classes) [16].

According to [16] a decision tree holds the function of partitioning recursively a training set until every subset obtained from the partition contains cases of a single class, by this reaching a model that will serve further classifications.

In this way one may possibly generate more complex trees which will end up losing their power of prediction. It will be therefore necessary to take some measures to turn more complex trees in less complex. Thus, after constructing the decision tree, pruning may be executed in order to reduce its size. Decision trees that are overly enlarged are susceptible to a phenomenon known as over fitting. Pruning helps to remove the branches from the initial tree in a way that the capacity of generalization of the decision tree is improved [14].

There are two possibilities for pruning a decision tree. The first one is to stop growing the tree early (pre-pruning) and the second one is to grow it completely and then prune it (post-pruning) [17]. According to [16], the second possibility is slower but safest.

A good model of classification must not only adapt to the data of training but also classify with precision the records never seen before. In other words, it must have a low quantity of errors of training and of errors of generalization. That is important because a model that is appropriate to the training data may have less error of generalization than a model of training with a high range of error [14].

The method J48 has as its base the C4.5 algorithm, which is one of the most traditional for the classification task. This was the method chosen for this study.

3. Material and Methods

Three databases were created for the development of this work. The first two databases contained the results for the courses that were taken by the students (1 for approval and 0 for failure). The data were provided by the Academic Registrar office for documentary research. The survey focused on freshmen students of 2003 in the Mathematics Major *UNIOESTE - Cascavel*. Researches on the records of the classes from 2003 to 2010 were carried out, considering that the maximum time to complete this Major is seven years and that the extension is allowed for an additional year.

No data from the course ‘Teaching Practices’ were used and since during the period of 2003 to 2006 there were changes in the Major curriculum some courses were not included in the research because they were either extinct or merged with others. The 2003 curriculum for the first two years of the Major was then composed of the following courses for the first year: Fundamentals of Elementary Mathematics, Analytic Geometry and Vectors, Differential and Integral Calculus I, Geometrical Drawing and Descriptive Geometry and Finance Mathematics. For the second year: Differential and Integral Calculus II, General and Experimental Physics, Geometry and Finite Mathematics.

Other information were added to the data, such as information regarding the credit hours transfer from previous courses for students transferred from other institutions and also regarding students who attended classes from other Majors or in the years before 2003 (in the case of students who had already joined the course before).

There were 50 freshman students in the Mathematics Major in 2003, however, one student was excluded from the research due to the fact that he had joined this Major at other times - there were many courses already taken in the past years and he was enrolled in very few courses after 2003.

The first database (base 1) had only the information regarding the courses taken in the first year of the Major. The database was formed by 49 patterns (students) and five variables. The correlation between variables was observed in each database. In this database there were no correlated variables.

The second database (base 2) had information on the courses taken in the first two years of the Major. The database was made of 27 patterns (students) since the students who dropped out the Major in the first year were removed. This database contains nine variables. Correlated variables were observed in this database: "result in the course of Differential and Integral Calculus I" and "result in Physics". The variable "result in Physics" was then removed from the base, leaving it with eight variables.

The third database (base 3) contains socioeconomic information such as family income, domestic budget, expenses with the university, city of residence, housing, among others. Data were collected through a questionnaire. The academics who took part in the survey were the ones enrolled for the courses in the Mathematics Major *UNIOESTE - Cascavel* in 2013 (except for those who started the Major in 2013) and those who dropped out in 2012.

In this study, Classification was used as the Data Mining task and the decision tree was used as the technique. We opted to use the software WEKA - Waikato Environment for Analysis Knowledge [18], which offers several features and is of easy use, in addition of being a public domain software. We used the J48 algorithm, an implementation of the C4.5 release 8 algorithm [19]. The parameters used in the algorithm were: Confidence Factor 0.25; number of folds 3; minimal number objects 2.

We used cross validation in order to choose the test set, by identifying the number of partitions. Cross-validation is used as an alternative when the data set is small (as it is in this work). The data set is divided into k partitions and during each run one of the partitions is selected for testing while others are used for training. This procedure is repeated k times so that each partition is used exactly once for testing. Cross validation was used with five partitions due to the size of the database used.

4. Results and Discussion

The high number of dropouts is also a reality in the *Universidade Estadual do Oeste do Paraná - UNIOESTE* (State University of West Paraná). The data provided by the undergraduate office show the reality of dropout in the Mathematics Major in *Cascavel* for the last five years.

Currently the Mathematics Major accepts 40 new students on an annual basis for evening classes. The number of students who graduate is very low, on the other hand, the number of students who dropout is very high. The graph in Figure 01 shows the number of dropouts from 2008 to 2012. The graph in Figure 02 shows the number of students who graduated in the same period.

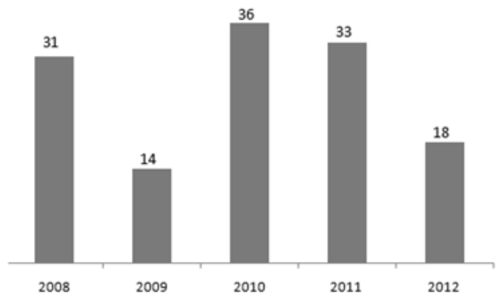

Fig. 1 Number of dropouts in the Mathematics Major in the last five years.

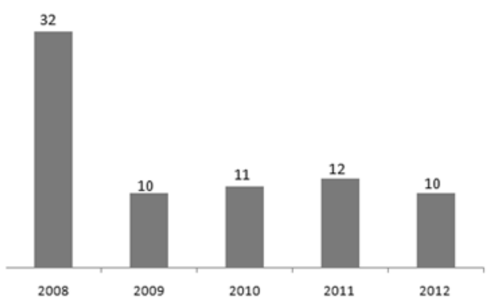

Fig. 2 Number of students who graduated in the Mathematics Major in the last five years.

4.1 Data Mining

Of the 49 students who entered university in the academic year of 2003 in the Mathematics Major, only one graduated in less than four years because he was able to transfer the course credits he took before. Only three graduated in four years; one completed in five years; seven completed in six years; two completed in seven years and one asked for an extra year to complete the course. In total there were 15 graduated students out of 49, that is, 30.6% of graduates.

Observing the collected data from the document research, it was possible to notice that most of the dropouts occurred in the first year of the Major. The number of students who dropped out in the first year, i.e. attended no course in the second year of the Major, was 22; that represents 44.9% of dropouts in the first year. In addition, five students dropped out the Major in the second year, five in the third year and two in the fourth year, that means that these last two had been enrolled for the courses in the fourth year but did not complete the Major.

The results obtained in the application of the Decision Tree technique for Classification is shown below. First, we show the results for database 1. The Decision Tree shown in Figure 03 has 91.84% accuracy. By the interpretation of the Decision Tree it is concluded that the student who has not been approved in the course of Differential and Integral Calculus I did not graduate. This rule was valid for 30 students. Moreover, the student who has been approved in the courses of Differential and Integral Calculus I and Financial Mathematics graduated. This rule was proved by 17 students, and was not valid for only two of them.

Considering only the courses taken in the first year, the course that contributed the most to dropouts was Differential and Integral Calculus I followed by Financial Mathematics.

The Decision Tree shown in figure 04 uses database 2 that is formed by 27 patterns and eight variables. Based on this Decision Tree, it is concluded that the student who was approved in Finite Mathematics concluded the Major requirements. This rule was applied to 16 students, and not valid for just one of them. Considering the subjects taken in the first two years, the course that contributed the most to dropouts was Finite Mathematics.

The Decision Tree showed in Figure 05 uses base 3. By interpreting the Decision Tree it is concluded that the work factor was the most important one for a student to keep attending the classes in the Major. Students working six hours or less do not dropout the Major courses. 28 students were applied to this rule and only two of them were not classified correctly, that is, worked six hours or less and still dropped out the courses.

The second most important variable was the marital status. The academic who works more than six hours a day and is married drops out the courses. This rule was applied to three students.

Finally, age also influenced in the decision to drop out the courses. If the student works more than six hours, is not married but is more than 24 years old, he will drop out the courses. Among the five students who were reached by this rule, only one was misclassified, i.e. did not drop out.

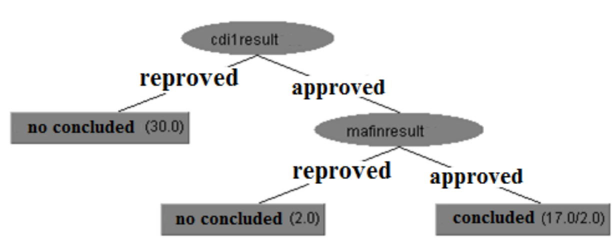

Fig. 3 Result of the Decision Tree for database 1.

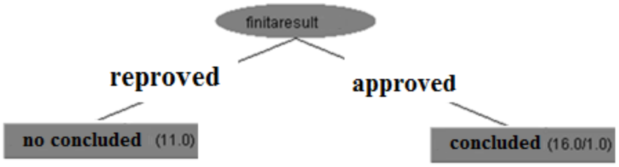

Fig. 4 Result of the Decision Tree for database 2.

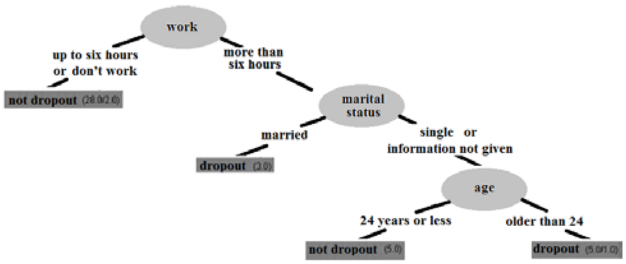

Fig. 5 Result of the Decision Tree for database 3.

5. Conclusion

Dropout in higher education is a well discussed subject in the last years in Brazil and at the *Universidade Estadual do Oeste do Paraná - UNIOESTE*, it is not different. Only in the Mathematics Major (*Cascavel*) there were 75 dropouts in the last five years. Thus, this study aimed to investigate academic, social and economic aspects of the students in Mathematics Major at *UNIOESTE - Cascavel* that may have influenced the decision to drop out the course.
This study was carried out using data from students who entered in the Mathematics Major in 2003, based on documentary research, from which data about the courses were collected. We also used data collected via questionnaire (addressing socioeconomic issues) applied to academics who were enrolled in the courses in 2012 (who were still enrolled in 2013 or who dropped out the courses in 2012).
The Data Mining technique was used in the search for useful information in the databases created. The task of data mining was the Classification. The patterns (students) were labeled as dropout and not dropout. The variables are the factors that could influence the students in making the decision to drop out the course. The role of the Classification is to find rules that predict the classes according to the attributes. In this work, to predict the decision to drop out the course due to variables related to socioeconomic factors and to the courses taken. For this work, WEKA software was used, applying the J48 algorithm in order to obtain Decision Trees.
It was possible to observe the courses that contributed to dropouts in the Major in different years. Considering only the subjects taken in the first year, the course that most contributed to dropouts was Differential and Integral Calculus I. Considering the courses taken in the first two years, the one that contributed the most to dropouts was Finite Mathematics.
By analyzing the results related to socioeconomic factors it was concluded that the work factor was the one that most contributed to the decision to drop out the courses. It is believed that this happens due to the fact that the student who works has little time to devote to extracurricular study. Other significant factors were marital status and age. Although trivial, the rules found endorse what by socializing with academics involved in the research was expected. It is hoped that this work may contribute to the elaboration of policies aimed at meeting the needs of the students, therefore ensuring their permanency at the University.
Among the expected actions there is the implementation of a University Restaurant, which would help to decrease the expenses the students have. Furthermore, an increased supply of scholarships (Scientific Initiation, Initiation to Teaching, academic monitoring, am ong others), giving opportunities to the students to have an income in order to remain without the need to work and still cooperating to its formation. These are just some suggestions for future measures, since fighting dropout is still a long process.
As a suggestion for future works we can mention the identification of the courses that most fail students, the identification of factors that may influence the high rate of failures in these courses, and the study of the profile of students who graduate the Mathematics Major.

References

[1] SILVA FILHO, R. L. L. E; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. DE C. M. A EVASÃO NO ENSINO SUPERIOR BRASILEIRO. Cadernos de Pesquisa, v. 37, n. 132,p. 641 – 659, 2007.
[2] GAIOSO, N. P. de L. O fenômeno da evasão escolar na educação superior no Brasil. 2005. 75 f. Dissertação (Mestrado em Educação) – Programa de Pós-Graduação em Educação da Universidade Católica de Brasília, Brasília, 2005.
[3] GISI, M. L. A Educação Superior no Brasil e o caráter de desigualdade do acesso e da permanência. Diálogo Educacional, Curitiba, v. 6, n. 17, p. 97-112, 2006.
[4] SOUZA, S. L. Evasão no Ensino Superior: Um Estudo Utilizando a Mineração de Dados como Ferramenta de Gestão do Conhecimento em um Banco de Dados Referente à Graduação de Engenharia. Dissertação de Mestrado. Coppe – UFRJ, Rio de Janeiro, 2008.
[5] BAGGI, C. A. dos S.; LOPES, D. A. Evasão e avaliação institucional no Ensino Superior: uma discussão bibliográfica. Avaliação (*Campinas*) [online], vol.16, n.2, P. 355-374, 2011.
[6] NOGUEIRA, F. País perde R$ 9 bilhões com evasão no ensino superior, diz pesquisador. [atualizada em 07 fev. 2011; acesso em 07 out 2012]. Disponível

em: http://g1.globo.com/vestibular-e-educacao/noticia/2011/02/pais-perde-r-9-bilhoes-com-evasao-no-ensino-superior-diz-pesquisador.html.

[7] BRAGA, M.M; PEIXOTO, L.C.D.M; BOGUTCHI, F.T. A evasão no Ensino Superior Brasileiro: o caso da UFMG. [atualizada em 25 maio 2003; acesso em 28 out. 2012]. Disponível em: http://educa.fcc.org.br/pdf/aval/v08n03/v08n03a08.pdf

[8] BRASIL, Comissão Especial de Estudos sobre a evasão nas Universidades Públicas Brasileiras. Brasília: ANDIFES/ABRUEM/SESU/MEC, 1997a. [Acesso em 28 out. 2013]. Disponível em: <www.mec.br/sesu/diretriz.htm>.

[9] PEREIRA, F. C. B. Determinantes da evasão de alunos e os custos ocultos para as instituições de ensino superior: uma aplicação na universidade do extremo sul catarinense. 172 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-graduação em Engenharia de Produção. Universidade Federal de Santa Catarina, Florianópolis, 2003.

[10] BRASIL, Diretrizes Curriculares para os Cursos de Graduação Superior (1997b). [Acesso em 28 out. 2013]. Disponível em: <www.mec.br/sesu/diretriz.htm>.

[11] BORGES, P. MEC e universidades estudam planos para combater a evasão. [atualizada em 11 fev. 2012; acesso 28 out. 2012]. Disponível em: <http://ultimosegundo.ig.com.br/educacao/mec-e-universidades-estudam-planos-para-combater-evasao/n1597622390779.html >.

[12] MANHÃES, B. M. L; CRUZ, S. M. S; COSTA, M. J. R; ZAVALETA, J; ZEMBRÃO, G. Identificação dos Fatores que Influenciaram a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: Uma Abordagem Quantitativa. In: VIII Simpósio Brasileiro de Sistemas de Informação (SBSI), Rio de Janeiro, 2012.

[13] FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHRUSAMY, R. Advances in knowledge Discovery & Data Mining. California: AAAI/MIT, 1996.

[14] TAN, P – N; STEINBACH, M.; KUMAR, V. Introdução ao DATA MINING Mineração de Dados. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.

[15] GONÇALVES, C.E. Data Mining com a ferramenta Weka. Fórum de Software Livre de Duque de Caxias – 2011.

[16] LEMOS, E. P. "Análise de crédito bancário com o uso de DATA MINING: REDES NEURAIS E ÁRVORES DE DECISÃO." Dissertação de Mestrado. UFPR, Curitiba, 2003.

[17] CARVALHO, D.R. Um método híbrido árvore de decisão / algoritmo genético para data mining, 2002. Tese de doutorado – PUCPR, Curitiba.

[18] WITTEN, I. H.; FRANK, E. Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers. San Francisco, California, 2000.

[19] OLIVEIRA, F. L. de; PADILHA, T. P. P.; PREVIERO, C. A.; MELO, A. C. G. de. Utilização de Algoritmos Simbólicos para a Identificação do Número de Caroços do Fruto Pequi. In: Encoinfo 2002, Palmas – TO, 2002. [acesso em 04 dez 2012]. Disponível em: http://arquivo.ulbra-to.br/ensino/43020/artigos/anais2002/Encoinfo2002/Algoritmos_Simbolicos.pdf.

Rosangela Villwock completed her undergraduate studies in Mathematics at Western Paraná State University, at Cascavel city, State of Paraná, in Brazil, in 1997. She got her Master's degree in this same university in 2003 and her Ph.D.' degree in Numerical Methods in Engineering at Federal University of Paraná, at Curitiba city, State of Paraná, in Brazil, in 2009. She is an Adjunct Professor at the Western Paraná State University, since 1999. Her recent interest includes Data Mining and Applied Mathematics.

Andressa Appio completed her undergraduate studies in Mathematics at Western Paraná State University, at Cascavel city, State of Paraná, in Brazil, in 2012. Her recent interest includes Data Mining, more specifically, Classification.

Aldioni Adaiani Andreta undergraduate in Mathematics at Famper – Ampere College, at Ampere city, State of Paraná, in Brazil. Her recent interest includes Data Mining, more specifically, Classification.