Data Mining in Academic Databases to Detect Behaviors of Students Related to School Dropout and Disapproval

José Antônio da Cunha[1], Elionai Moura[1] and Cesar Analide[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN)
Rua Dr. Nilo Bezerra Ramalho, 1692. 59015-300 Natal, RN, Brasil
{jose.cunha, elionai.moura}@ifrn.edu.br
[2]Universidade do Minho. Departamento de Informática
Braga, Portugal
analide@di.uminho.pt

Abstratct. This work focuses on data mining in relational databases, aiming to detect behaviors related to school dropouts and disapproval by mapping the factors that influence this dropout. This work is relevant by the fact that the dropout and school disapproval are big factors of concern to all who care about education in Brasil. At the end of it, we intend to point out the need to implement solutions that enable access to results dynamically, thus allowing educators can early diagnose the causes of school dropout and disapproval, and allow for relevant pedagogical actions. This way, we intend to reduce the school dropout and school disapproval, towards a more efficient teaching and learning process at brazilian federal education institution named Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte - IFRN.

Keywords: School Dropout, School Disapproval, Educational Data Mining, Machine Learning, Teaching and Learning.

1 Introduction

The problem of dropout and school failure in federal institutions has generated some challenges to overcome. The high incidence related to these factors, it has been lived in the practical experience of all educators that make education in these institutions. It is known in advance that school dropout and school disapproval are associated with factors such: areas of knowledge of students, educational levels and specific methodologies of teaching and learning. Therefore, it is intended in this work, apply data mining techniques in the academic data base in order to map the factors that are associated with dropout and school failure.

The data spectrum used for evaluation or analysis was not restricted to the academic system database, but also the survey forms applied at the institute, filled in by students and teachers at the end of each school year. This fact allows for a more comprehensive analysis because it involves multidisciplinary teams, with important additional aspects, such as the knowledge acquired by the teaching staff with the interactions with students and their parents, and knowledge acquired by teachers through the school activities with students.

In the context from multidisciplinarity, the information acquired, can represent both an element of support in the teaching-learning process as well as provide sources of information for the continuous monitoring of results obtained by data mining, both by the system specialist, as by the application domain experts. Furthermore, this information can be very important in building models that can be the basis for actions to be taken by the teaching staff and education managers in order to avoid or reduce dropout and school failure.

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013. This paper aims to map factors that may be associated with dropout and school failure through machine learning techniques and data mining, with the purpose, allowing proactive actions to stimulate the students, aiming the continuity of students in their respective courses, and thus it can mitigate the risks related to dropout and school failure.

Á. Rocha et al. (eds.), *New Advances in Information Systems and Technologies*,
Advances in Intelligent Systems and Computing 445,
DOI 10.1007/978-3-319-31307-8_19

2 School Dropout and Disapproval

The problem of school dropouts in Brazil is not a recent problem, but rather a repeat offender. It is one, of the factors, that concerns educators and public policy makers in our country. According to the Ministry of Education (MEC), school dropouts reaches 6.9% in primary and 10% in high school (3.2 million children and young people, according to data of 2005). There are over 2.9 million students [14] who leave school a year and return the next, thickening other disturbing content: the level-age distortion.

According to [11] [6], school dropout is what happens when a student fails to attend school and is characterized early school leaving, and is historically one of the topics that is part of the debates and analyzes of public education. Several factors can lead to school dropout. Among them, teaching-learning misapplied by inadequate methodologies, ill-prepared teachers, social problems, neglect by the government, and so on.

Historically, one of the first works to systematize the dropout of the problem in Brazil was conducted from a national commission established by Ministry of Education (MEC). The Special Commission for the Study of Dropout in Brazilian universities came within an institutional assessment thread context, defined by indicators of the Institutional Evaluation Program of Brazilian Universities (PAIUB), directed by different educational institutions, specifically the public [14].

Studies prior to this, particularly in the second half of the 80s, emphasized only statistical surveys and case studies in a fragmented way, carried out by the Ministry of Education initiative and public universities. However, these studies did not develop the problem in order to create institutional policies, estimates, administrative and pedagogical actions, that is, side dishes needed to minimize the results [14]. This initiative was a first joint effort of different public higher education institutions (HEI) to systematically organizing a study that defined a single methodology in order to identify causes and possible solutions to the problem. The ultimate goals of this committee was to clarify the concept of avoidance, examine the rates and causes of this phenomenon and standardize a methodology to be employed by the institutions.

The development of the Commission's activities grasps also that, the predominant causes of dropout were with three rows. The 1st related to students, the 2nd related to courses and institutions and the 3rd, the more conjunctural order, called "socio-cultural and economic variables". This last is related to the labor market, social recognition of the chosen career, the quality of primary and secondary education, socio-economic context and government policies. [14]

In studies of the Special Commission for the Study of Dropout [18], we also find research on the performance of European universities and North American in a time series from 1960 to 1986. In this research, the best university system yields was found in Finland, Germany, Netherlands and Switzerland while the worst results occur in the United States, Austria, France and Spain. According to research, the United States dropout rates in the last 30 years are around 50%. A similar number is in France where rates in 1980 were 60 to 70% in some universities. In Austria, in turn, points to a 43% dropout rate, with only 13% of students complete their courses within the time limit [14].

2.1. School Disapproval at the Federal Institutions Network

The term failed means censored, criticized, condemned, as the word failure means disdain, criticism, contempt. Now the meanings of expressions already reveal by themselves the implications. However, a trivial excuse to justify the act of reproving is that the student spend another year in that series, seeing again the contents that could not assimilate, will be more successful, even in his academic life. This is a great fallacy, because the student who repeats a school year lose motivation, is the embarrassment of being again in that same school year, either by living with smaller colleagues with different interests [12].

To [10], the rejection is now widely questioned. After all, making students repeat the entire year to see the same content again is an outdated solution, dresser, expensive and inefficient. Countries with high-quality teaching and learning found alternatives that work better, through preventive action, such as booster classes throughout the year. In Finland, teachers are advised to devote more time to students who have more difficulties. Result: the failure rate is 2% and the primary education completion rate is 99.7%. In Hong

Kong, when a teacher has more than 3% of students with low performance, a committee will evaluate the teacher's work.

Data released in accordance with [17], Brazil is one of the countries most disapprove. In high school the rate reaches 13.1%. Are almost $ 3 billion / year spending beyond what is necessary, only in the final years of schooling. The worst is that, as shown in qualitative and quantitative research, there is great relationship between repetition and dropout. No wonder that the study recently published by the "Education for All" shows that only 54% of young Brazilians manage to graduate from high school up to 19 years. Of young people between 15 and 17, one in five still in elementary school, accumulating failures. And 15.7% dropped out, certainly after school failure experiences [17].

The fact that the school failure influence school dropout, justifies the importance of the study of these topics.

3 Data Mining in Databases

The constant technological advancement has enabled the rise of technologies such as the internet, social networking, mobile devices, virtual learning environments, sensors to collect different types of data, telecommunication systems and secondary memory for greater storage capacity. Those are some examples of features that are making possible the creation and growth of numerous databases of administrative, scientific, commercial, educational, governmental and social [9].

However, the amount of stored data is closely linked to the ability to extract knowledge of the highest level from them, that is, useful information that will serve to support decision-making, and / or operation and better understanding of the phenomenon that generate data [8] [9].

However, the large amount of data analysis by man is impossible without the aid of appropriate software tools. Thus, it becomes essential to use tools that help man the task to analyze, interpret, and relate these data, so you can prepare and select action strategies in each application area [8] [9].

Therefore, to meet this context, there is an area called Knowledge Discovery in Databases (KDD), which is attracting in recent years, considerable interest among the scientific and industrial communities [9].

3.1. Data, Information and Knowledge

Every moment in this work, we are talking about data, information and knowledge. Therefore, it is important to note the differences between data, information and knowledge [8].

- **Data**: they can be interpreted as elementary items, captured and stored by the information technology resources. They are strings of symbols and do not have semantics (ie, meaning). Its purpose express real-world facts in order to be treated in the computational context.
- **Information**: the information represents the processed data, with well defined meanings and contexts. For example, the monthly borrowing capacity is a calculated information from the income and monthly expense of each client. In this case, the debt indicates a percentage value as a financial client can contract loans in relation to their monthly income.
- **Knowledge**: knowledge corresponds to a standard or set of standards whose formulation may involve and link data and information. Knowledge can be represented in the form of a conditional rule (IF <condition> THEN <conclusion>). Another way to represent knowledge is through predictive trends.

3.2. KDD Definitions

The term KDD was formalized in 1989 in reference to the broad concept of seeking knowledge from databases. One of the most popular definitions was proposed in 1996 by a group of researchers [5]. An adaptation of the original definition is shown below:

Definition 1: KDD is a non-trivial, interactive and iterative process to identify patterns understandable, valid, new and potentially useful from large data sets.

- The term *interactive*: It indicates the need for human action as responsible for process control. In fact, there are usually two human actors involved: the **data analyst** and **domain expert**.
- The term *interactive*: it suggests the possibility of full or partial repetition KDD, in the search for satisfactory results by successive refinements.
- The term *non-trivial*: warns of the complexity normally present in the execution of KDD processes.
- The term *identify patterns*: according to the definition, the purpose of performing the KDD process is to identify patterns. A standard is a knowledge representation in the syntactic rules in some formal language.
- The term *understandable*: one of the objectives of the KDD process is to produce knowledge that can be understood easily, thus allowing a clear understanding of the data that gave rise to this knowledge. One possible technique to accomplish this is to present the patterns in a graphical manner that facilitates their understanding.
- The term *valid pattern*: it indicates that knowledge should be true and appropriate to the context of the implementation of KDD.
- The term *new pattern*: a new standard to add new knowledge to previously existing knowledge in the application of KDD. The question of a standard found to be dependent on the new point of view in the scope of the KDD process or in the user's scope.
- The term *useful pattern*: a useful pattern is one that can be applied to provide benefits in the context of application KDD. Namely, the discovered patterns are useful only if they help to achieve the goal of domain expert.

The patterns extracted in the KDD process can be classified into two basic types: **descriptive** and **predictive** [8].

- **Predictive patterns**: they are constructed in order to solve a specific problem to predict the values of one or more attributes, depending on the values of other attributes.
- **Descriptive patterns**: the centerpiece of descriptive patterns is to present interesting information that a specialist application domain cannot yet know.

A standard describes facts (and trends) associated with a data set, with any degree of certainty. Therefore, the KDD process presupposes the existence of a data set. This may involve n attributes, thus representing a hyperspace (n-dimensional space). The greater the value of n and the number of registers available, the larger the dataset to be analyzed [8].

The representation of the degree of certainty with which the standards describe a collection of data is essential to determine how much a system or user can trust these patterns and make decisions from them. In general, the calculation of the degree of certainty of a standard involves several factors such as, for example, data integrity, the sample size used in the process, the existence of some knowledge on the field of application, among others [8].

Definition 2: KDD process consists of a sequence of complex interactions, which extends over a certain period of time, between a 'user' and a collection of data, possibly aided by a diverse set of computational tools [3].

In the definition 2, data analyst is always present and intimately involved with every step of the process. The term heterogeneous set of tools corresponds to the KDD system used by the analyst.

[3] and other authors claim that the interaction of the analyst with the data leads to the formulation of hypotheses about them. The **data analyst** view the data as a whole and decide where to explore based on what he sees in his own experience and knowledge provided by the domain expert. Recently, this type of professional is known as **data scientist**.

3.3. Related areas to KDD

This is a multidisciplinary area and there are already for a long time and originated from several research areas such as **Statistics**, **Machine Learning**, **Pattern Recognition**, **Computational Intelligence** and others.

- **Machine Learning:** one step in KDD process, the extraction of patterns (or data mining), uses machine learning methods (ML) to find regularities, patterns or concepts in data sets. Techniques developed in ML, as the rules of induction and decision trees, connectionist models and learning based on instances, form the core of the methods used in data mining.
- **Statistics:** statistics, together with the Machine Learning area, is considered ancestor of the KDD area. Pattern recognition techniques and exploratory analysis of data from the statistics are widely used in data mining algorithms. Data selection and sampling, pre-processing, data processing and evaluation of extracted patterns are just a few examples of methods widely used in statistics and which are applied during the process of KDD.
- **Database:** A database is an integrated collection of data, organized in a way to facilitate efficient storage, as well as its modification and recovery [4]. It is usually managed by a Database Management System (DBMS), which corresponds to a collection of procedures and mechanisms for recovery, storage and manipulation of databases.
- **Data Warehousing:** Data Warehousing is another area related to the KDD process, and refers to the process of collection and pre-processing of data stored in one or more operational databases in order to serve as a source for Decision Support Systems. As a result of this process we have a Data Warehouse, a collection of integrated data, consolidated and possibly organized in time (historical data).

3.4. KDD Activities

Activities in the KDD area can be organized into three main groups: activities related to technological development, KDD process execution activities and activities involving the application of results obtained in the process of KDD [8].

- **Technological Development:** covers all design initiatives, development, improvement and optimization algorithms, tools and assistive technologies that can be used in the search for new knowledge in large databases.
- **KDD Execution**: refers to the activities related to the effective pursuit of knowledge in databases.
- **Application of Results:** has been achieved models of useful knowledge from data set, activities are focused on the application of the results in the context in which it was carried out the process of KDD.

3.5. Clustering Algorithm and Analisys Services

The Microsoft Clustering algorithm is a segmentation algorithm provided by Analysis Services software. The algorithm uses iterative techniques to group instances in a set of data clusters that contain similar features.

The Microsoft Clustering algorithm, provides two methods for creating clusters and assigning data points to clusters. The K-means algorithm, a "hard clustering" method. This means that one data point can only belong to a cluster and that one probability is calculated for associating each data point that cluster. And the method of Expectancy Maximization (EM), a flexible clustering method. This means that one data point always belong to multiple clusters, and a probability is calculated for each combination of data point and cluster.

You can choose the algorithm to be used by setting the parameter CLUSTERING_METHOD. The cluster standard method is the evolutionary EM.

In EM cluster, the algorithm iteratively refine an initial clustering model to fit the data and determine the probability of a data point exists in the cluster. The algorithm terminates the process when the probability model fits the data. The function used to determine the fit is the probability of log data according to the model. If empty clusters are generated during the process or the combination of one or more clusters is below a certain threshold, the clusters with low populations will be propagated again in new points and the EM algorithm will run again.

4 Data Mining Application in IFRN Database

Are applied data mining techniques on the basis of data available, to detect which the attributes that are most influencing school dropout and thus draw a profile of the factors that imply the school dropout. It is known that some factors that influence school dropout, are external to the school environment, such as relationships with parents, dysfunctional families, and so on, plus the profiled here can be used along with other factors in order to have a more precise analysis of the problem in question.

First, we did a historical overview about school dropout in IFRN, 2000 to 2013. Figure 1 shows the percentage of dropouts. The graph in Figure 1 shows that for the campus Natal-Central, in 2000 the dropout rate was 20.26%, in 2001 was 43.23%, in 2005 was 34.63%. It is observed that the dropout percentage at Natal-Central campus is always above 15% and in some years more elastic. Without a doubt, it is a high rate, and worrying, and it deserves a detailed study of it. However, it has other data that also deserves to be noted. It is the cancellation of enrollment in courses of federal education network.

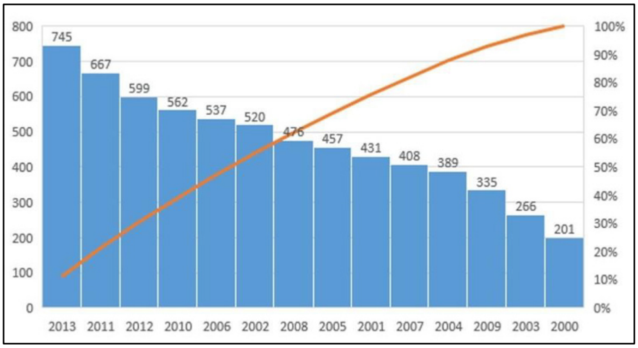

Fig. 1. Graph showing the percentage of school dropout per year at Natal-Central campus of IFRN in years 2000-2013.

Figure 2 shows the evolution of cancellation in enrollment and dropout on campus Natal-Central, between the years 2000 and 2013. In Figure 2, if we analyze the situation for 2007, we have the total of students who canceled their enrollment in courses was around 240, and the total number of students who dropped out of courses was around 650 students. If we add the two factors we have 900 students who dropped out of their courses in that year of 2007. Therefore, the registration cancellation rate must also be taken into account in the assessment of the teaching and learning.

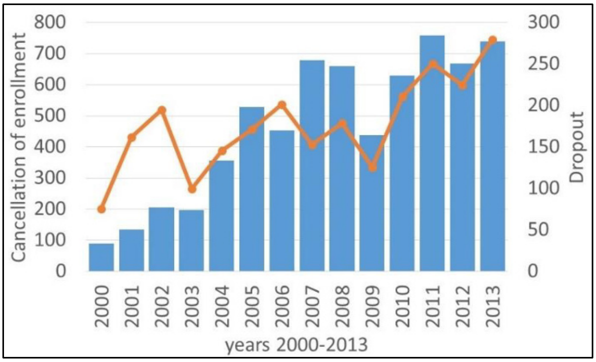

Fig. 2. Cancellation of enrollments at Natal-Central campus of IFRN in years 2000-2013.

Table 1 shows the failure percentage per subject in 2010. It was done in a filter data to show only percentage of failure from 40% to 60%. Then see that many disciplines have very high failure rate. And as was said earlier, the failure rate has implications for the dropout because often may discourage the student to continue on the course.

Based on the data shown in the graphs of Figure 1 and 2, we have the proof of the high failure rates and dropout in IFRN, campus Natal-Central. Therefore, it will be applied to the same data set, some of Data Mining algorithms in order to find something to do in the attributes of the database, you can trace a profile of failure situations and dropout of our students.

Table 1. A sample of the failure rate in some disciplines in IFRN for 2010.

Discipline	**Year**	**Retention**	**Students**	**%**
Practice as Curricular Component	2010	17	34	50
Foreign Language - English	2010	22	44	50
Work Psychology	2010	63	127	49,61
Algorithms and Object Oriented Programming	2010	58	118	49,15
Differential Equations	2010	20	41	48,78
Conservation of energy	2010	17	35	48,57
Web Authoring	2010	228	473	48,2
Techniques of Food Laboratories	2010	66	138	47,83
Cell Biology	2010	18	38	47,37
Informatics I	2010	61	129	47,29
Differential and Integral Calculus II	2010	80	170	47,06
Environmental Biology	2010	23	49	46,94
TCP/IP Architecture	2010	37	79	46,84
General Chemistry and Experimental I	2010	79	170	46,47
Electrical Systems	2010	26	56	46,43
Soil Mechanics	2010	117	252	46,43
Biology	2010	45	97	46,39
Data structure	2010	24	52	46,15
Open Systems Administration	2010	29	63	46,03
Elements of Physics	2010	89	195	45,64
Optical	2010	26	57	45,61
Electricity	2010	232	509	45,58
Digital electronics	2010	81	180	45

Figure 3 shows a network obtained by the application of decision tree algorithm, using the tool Analysis Services [16] [10]. For this network training was provided as predictive attribute the situation of the student and the other attributes were defined as input attributes to the algorithm. Also in Figure 3, we can see the attributes that influence school dropout. So we can draw a profile for school dropout, analyzing each of these attributes.

The attribute "type of home school", can take the walloons private or public and philanthropic school. The attribute "income" is the family income of the student, the attribute "efficiency coefficient" measures the performance of the student in the course, and attributes "media" and "faults" represent the academic performance of students. The attribute "entry way" indicates how the student entered the course (ESMS, take selection, transfer, and so on).

We will use the cluster algorithm for grouping students with similar characteristics in the same group, and then analyze each cluster to identify the degree of influence of each input attribute shown in Figure 3 in relation to the predictive attribute "Status = Dropout".

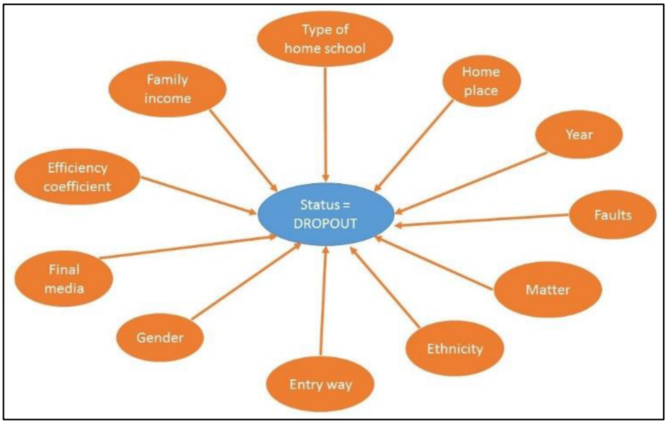

Fig. 3. Network Decision Tree showing the relationships between attributes. Tool used to create the chart was Microsoft Analysis Services [16] [10].

Figure 4 shows the graph generated by the cluster algorithm of Analysis Services tool. Cluster chart shows that the highest concentration of cases of dropout, are precisely those with the most intense blue color. For this ran was selected in the cluster algorithm configuration, the situation of "Dropout" and the clusters with fewer cases of dropout, are those with a less intense color.

Thus, the cluster 5, the darker blue color, means that it is having the largest number of dropout and the cluster 1 has the lowest number of cases of school dropout. Cluster 1 has the highest number of approved.

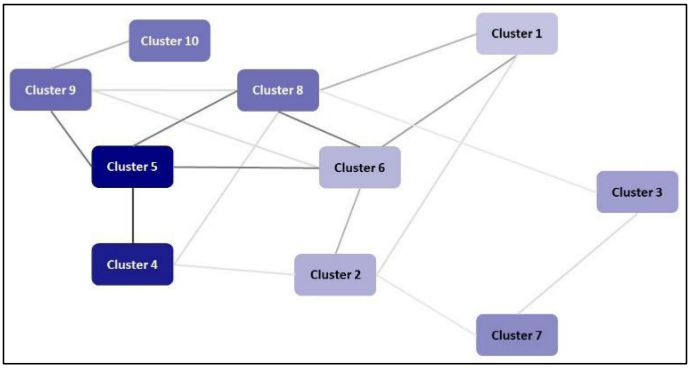

Fig. 4. Cluster chart. Tool used to create the chart was Microsoft Analysis Services [16] [10].

Let's look at the cluster 1 and 5 to see which were the input attributes that influence the composition of them. Figure 5 shows the cluster 1 characteristics, which is concentrated the largest number of approved. Figure 6 shows the characteristics of cluster 5, which is concentrated the largest number of dropouts.

The cluster 1 shows that the profile of successful students are those with final average above 70, GPA above 60, coming from public school, live with their parents and brown ethnicity. Observing the cluster 5 characteristics: it is clear that in cluster formation 5, family income (up to 1 salary) and the situation (deprecated), appear as factors influencing school dropout. Justifying thus the presence of these attributes in relation attributes that influence school dropout, as shown in Figure 3.

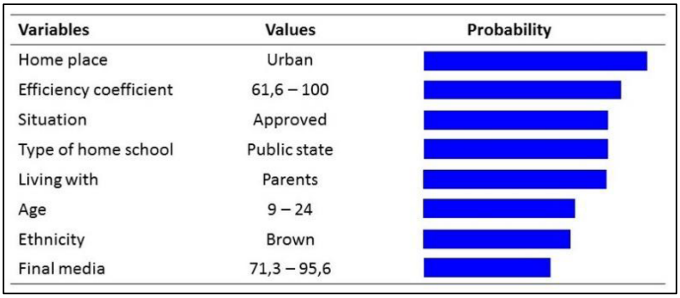

Fig. 5 Cluster 1 characteristcs. Tool used to create the chart was Microsoft Analysis Services [16] [10].

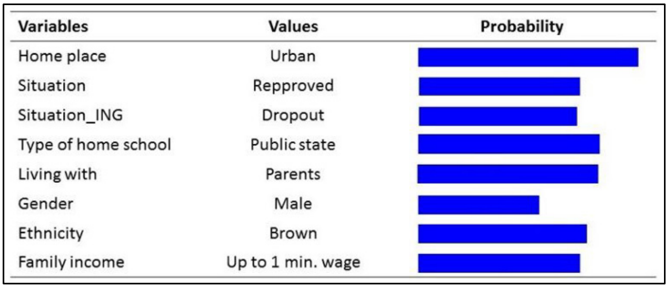

Fig. 6 Cluster 5 characteristcs. Tool used to create the chart was Microsoft Analysis Services [16] [10].

5 Results and Conclusion

Based on the obtained results show that the dropout rate at Natal-Central campus is quite high, above 16% for the years after 2010. Another finding, very important, that until then, no one had noticed, is the index registrations canceled the courses. If we add the dropout index with the index of canceled registrations, we will have an index above 25%. If we observe that, the IFRN has around 5,000 students, 25% of 5,000 = 1,250 students who dropped out or canceled their registrations annually.

Analyzing the graphics obtained by the data mining algorithms, at first, one can trace a profile for school dropout as being: students from state schools, with family income up to one minimum wage, living with parents consequently, they are unemployed or are minors, of mixed race, with the final and very low yield coefficient. It is known in practice that students entering the public schools come IFRN, arrive with knowledge far from desired in basic subjects such as mathematics and portuguese, which are fundamental to have a good performance in those courses. These students face many difficulties to subjects that contain logic, advanced abstraction and or mathematics such as the technical disciplines of technological area.

Based on this layout profile, one can suggest that the IFRN, Natal-Central campus, adopt some preventive measures to minimize both tax dropout, as school failure. Among them, one can cite:

- The result of the analysis should be shared with all staff of the IFRN, so everyone has knowledge of the actual situation;
- Propose the development of outreach projects, to work with the new students the basic knowledge of portuguese and mathematics;
- Making an analysis of the data of the selection tests in order to predict the actual situation of students in the target disciplines (mathematics and portuguese), to have real numbers that lag in these disciplines and thus make plans and goals to create booster classes in matters in which the incoming students have more difficulties.

These are just some of the goals that will be proposed this preliminary study, however, will be continued in the analysis of the academic system data and certainly more knowledge will emerge, and these managers IFRN will be passed, so that action can be taken that, will reduce the problem of dropout in our school.

References

1. Adachi, A.A.C.T, Evasão e evadidos nos cursos de graduação da Universidade Federal de Minas Gerais, Dissertação de Mestrado. Faculdade de Educação, UFMG (2009)
2. Baker, R. S. J., CARVALHO, A., M., J., ISOTANI, S., Mineração de Dados Educacionais: Oportunidades para o Brasil, Revista Brasileira de Informática na Educação, Volume 19, Número 2 (2011)
3. Brachaman, R. J.; Anand, T. Processo f Knowledge discovry in Database. KDD Process for Rxtracting Useful Knowledge from Volumes of Data (1996)
4. Date, C. J. Introdução aos sistemas de bancos de dados. Campus (2003)
5. Fayyad, U. M.; Piatetsky-Shapiro, G.; smyth, P. From Data Mining to Knowledge Discovery: An Overview. Knowledge Discovery and Data Mining, Menlo Park: AAAI Press (1996)
6. Favero, R. V. M., Dialogar ou evadir: Eis a questão!: Um estudo sobre a permanência e a evasão na Educação a Distância, no Estado do Rio Grande do Sul. CINTED-UFRGS. Novas Tecnologias na Educação, V. 4 Nº 2, Dezembro, Porto Alegre (2006)
7. Globo Comunicação e Participações. http://g1.globo.com/educacao/noticia/2013/11/mec-cria-grupo-para-estudar-evasao-escolar-nos-institutos-federais.html (2013)
8. Goldschmidt, Ronaldo. Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações/ Ronaldo Goldschmidt, Eduardo Bezerra. 2.ed, Elsevier, Rio de Janeiro (2015)
9. Han, Jiawei; Kamber, Micheline. Data Mining: Concepts and Techniques. Second Edition. Elsevier. San Francisco, CA (2006)
10. Harinath, S.; Pihlgren, R.; Lee G., D.; Simon, J.; Bruckner, R. M. Professional Microsoft SQL Server 20012 Analysis Services with MDX and DAX. Wrox, (2012)
11. Karinny, Shintani, Leonardo Armond, Vassily Rolim. Dificuldades Escolares http://www.profala.com / arteducesp19.htm (2014).
12. Lilian L. Pereira. http://www.direcionaleducador.com.br/edicao-80-set/11/reprovacao-escolar-uma-questao-muitas-facetas (2011)
13. Lobo, R. L., Hipólito, O. e Lobo, M.B., Estudo: Evasão no Ensino Superior: Causas e Remédios. Jornal Folha de S.Paulo. São Paulo (2007)
14. MEC/Setec. Documento Orientador para a Superação da Evasão e da Retenção na Rede Federal de Educação Profissional, Científica e Tecnológica (2014)
15. Piatetsky-Shapiro, G. e Data-Mining Industry Coming of Age. IEEE Inteligent System (1999)
16. Rainardi, V. Building a Data Warehouse: With Rxamples in SQL Server. Apress (2008)
17. Ramal, Adrea. Reprovar não é a solução, mas aprovar quem não aprendeu é pior. http://g1.globo.com/educacao/blog/andrea-ramal/post/reprovar-nao-e-solucao-mas-aprovar-quem-nao-aprendeu-e-pior.html (2014)
18. BRASIL. Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas. Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras. Brasília (1996)