**Data mining in academic databases to detect Behaviors of Students Related to Evasion and Disapproval School**

**Cunha, José Antônio**1 e **Cesar Analide Rodrigues2**

1Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN)

Natal-RN-Brasil

2Universidade do Minho. Departamento de Informática

Braga-Portugal-PT

**Abstract**: This article focuses on data mining in relational databases, in order to detect behaviors evasion and school failure, by mapping the factors that influence evasion. This work is relevant simply by the fact that evasion and school repetition, are big factors of concern to all who care about education in Brazil. At the end of this work, we intend to point out the need to implement solutions that enable access to results dynamically, thus allowing, that educators can in advance, diagnose the causes of the school evasion and enable conducting relevant educational activities. All this has the aim of reducing evasion and repetition school, in favor a more efficient teaching learning, in the Federal Institutes.

**Keywords**: School Evasion, School repetitions, Data Mining, Machine Learning, Teaching learning.

**1 Introduction**

The problem of evasion and school repetition in federal institutions, has generated some challenges to overcome. The high incidence related to these factors, it has been experienced in the practical experience of all educators that make education in these institutions. It is known in advance that, evasion and school repetition are associated with factors, such as 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 database in order to map the factors that are associated with evasion and school repetition.

I must point out that the data spectrum to be used for evaluation or analysis, not restricted only to the academic system database, more also the survey forms applied at the institute, to students and teachers at the end of each year school. 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 of acquired teachers through the school activities with the students. In the context of a multidisciplinary approach, 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. In addition to more, this information can be very important in building models that can serve as a basis for actions to be taken by the teaching staff and educational or managers in order to avoid or reduce evasion and school repetition.

This work aims to justify the need for a process to analyze the factors that are associated with evasion and school failure, using the available resources in the Data Mining. Justifies the use of Data Mining techniques, the fact that the academic system is a large database, making it difficult for humans to do analysis on it without using appropriate tools. Is also justified the use of data mining techniques, in the early identification and dynamic, of accurate information on evasion and school repetition, which can produce results that could be used to guide effective pedagogical actions, and that in some way, these results can make part of a broader process. Therefore, this work deals with the mapping of factors that may be associated with evasion and school failure, through machine learning and data mining techniques, in order to, allow proactive actions to stimulate the students, aiming the continuity of students in their courses, and thus power mitigate the risks related to evasion and school repetition.

This work is structured as follows. In the section 2 will be done a contextualization of the issues involved, with evasion and repetition school, for these are the challenges to be attacked in this work, and have implications for teaching and learning process in the institutes. In the section 3 we will discuss solutions to the problem of truancy using the Data Mining techniques, as well as the main aspects involved in this context. In the section 4 will be applied the Data Mining techniques on the Database of consolidated Academic System, to identify patterns of behavior that may indicate attributes associated with evasion and school repetition. At the end of the article to section 5 presents a conclusion of the study and discussion of future work.

**2 Evasion and Repetition school**

In the first instance, a search will be made on some historical aspects of truancy, to justify the importance of their study as well as the importance of carrying out actions to minimize its negative effect on teaching learning in the IFRN.

**2.1 Evasion school**

The problem of evasion school in Brazil is not a recent problem, but rather recidivist. It is one of the factors that concerns educators and responsible for public policy in our country. According to the Ministry of Education (MEC), the truancy reaches 6.9% in primary and 10% in high school (3.2 million children and young people, according data to 2005). More of 2.9 million students [9] who leave school a year and return in the next, thickening other disturbing index: the distortion of series-aged.

The Ministry of Education set up a working group to understand the causes and propose solutions to truancy in technical courses. According to decree published in the edition of 11.25.2013 the "Diário Oficial da União", signed by the Secretary of Vocational and Technological Education MEC, Marco Antonio de Oliveira, the group will have 120 days to complete the work [4].

That same year, ie in April 2013, the Court of Audit (TCU) conducted an audit in the federal network of vocational education, science and technology, which pointed out that school dropout rates have reached 24% of all students enrolled in the courses (PROEJA1), plus 19% in the subsequent average courses.

According to [9], many of evasion problems is due to "historic expansion," what is happening in vocational education network. For a century, we had 140 units in just over 10 years jumped to 440 campuses. It is a historic expansion of large scale and at high speed, which creates an imbalance, said Secretary of Education at the time.

As previously emphasized, truancy in Brazil is repeat offender and in 1995 was promoted by the MEC [8] a comprehensive study on the performance of Brazilian public universities with respect to graduation index, retention and dropout of students in its undergraduate courses.

According to [6], school evasion 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 cause school evasion. Among them, teaching-learning misapplied by inadequate methodologies, ill-prepared teachers, social problems, neglect by the government, and so on. The debate over the source of the problem varies depending on the point of view of the debaters. Can from both of the role of the family as well as of the state and the school regarding the student's school career, and also of the ruling elites, be they economic, religious, or other species [6].

According to [3] in some cases it is regarded as evasion school the abandonment the course by the student, regardless of the amount of participations made; in other situations differs evasion according to average periods for course completion and annual periods [9]; You can also identify as truancy situations of the definitively abandonment after a certain contact the course. As you can see, there are many situations where you can declare that a student has abandoned the course.

According to [1] truancy, have emphasized the complex nature of relationships involved, such as personal characteristics, expectations and motivational events. These studies involve characteristics of the social condition of the student, such as gender, age, personal skills, previous school experiences, along with their expectations for personal development and career associated with motivation for academic achievement and its recognition.

Several factors are used to calculate indices for measuring truancy [1], such as the disciplines abandonment of a course, the waiver of a course, and enters another course within the same institution, which is different from situations in which the student turns off completely a particular institution. The MEC uses the equation (1) to calculate the evasion school index.

(1)

Where,

IMEC: indicator of evasion of the MEC

Mn: number of students with enrollment in the period

In: amount of entrants in the period

Rn: amount of re-enrolled in the period

Mn-1: amount of students registered in the previous period

Fn-1: amout of completed in the previous period

Equation (1) it is a description of the index used to calculate evasion by MEC capable of measuring evasion seen in a course between subsequent semesters (immediate evasion).

**2.2 School repetition in the Federal Education Network**

The term repetition mean censored, criticized, condemned. Now the meanings of expressions already revealed themselves their implications. However, a trivial excuse to justify the act of repetition is that the student spend another year in that series, seeing again the contents that can not assimilate, you will be more successful, 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 [7].

To [10], the repetition is today 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 find alternatives that work better, through preventive action, such as remedial classes throughout the year. In Finland, teachers are advised to devote more time to students who have more difficulties. Result: the repetition 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 [10], Brazil is one of the countries that most disapproves. 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 school repetition and evasion.

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 repetitions. And 15.7% give up, certainly after school failure experiences [10].

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

**3 Data Mining in Databases**

The constant advances in information technology have made possible the storage of large databases. Technologies such as the Internet, database management systems, with higher capacity data storage devices, lower cost and information systems in general are some of the examples that have enabled the proliferation of numerous bases of commercial data, administrative, government and scientific [5].

Therefore, the large amount of data generated, collected and stored, obtained by daily operations or scientific research, requires an automated process to discover patterns, exceptions, trends or correlations between them. Hence, there arises the need for Data Mining (English - Data Mining) [5].

In this scenario, the analysis of large amounts of data by man, is not viable without the aid of appropriate computer tools. Therefore, it is essential develop tools that can assist man, automatically and intelligently, the task to analyze, interpret and relate these data so that we can develop and select action strategies in each application context [5]

It occurred in schools and university, wide dissemination of the use of computerized systems, and thus grows every day the volume of data generated and stored in databases. It is noteworthy that this large volume of data has raised interest in their use, along with data mining techniques, in seeking answers to specific questions of education, related to learning processes, development of instructional materials, monitoring and forecasts, among others [2], from obtaining patterns important to support certain pedagogical practices.

The data mining is an area of research in expansion, the main approaches related are prediction, clustering, association rules, classification, discovery of models and processing of data for decision support, and so on. With this range of applicability, the data mining, can glimpse applications related to the diagnostic process of the evasion and grade repetition in IFRN and also the acquisition of general models that can assist the domain experts in carrying out actions in order to reduce the repetition rate and evasion in our institution.

In the academic system database, we have of the historical data of all educational lives of students that can be used to detect relationships between the attributes the evasion and school repetition and, thereby generating information that can be used in the prevention and broad and educational actions. Since the data from the surveys every six months applied in IFRN teachers and students, can be used to generate diagnostic and indication of immediate actions within a smaller scope of time, but equally important, dealing with situations of possible immediate evasion.

**4 Data Mining application in IFRN Database**

Apply data mining techniques in database available to detect which attributes that are most influencing truancy and thus draw a profile of the factors that imply the truancy. 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 of all, a historical overview was done about truancy in the 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 evasion school rate was 20.26%, in 2001 was 43.23%, in 2005 was 34.63%. It is observed that the evasion percentage at campus Natal-Central is always above 15% and in some years more elastic. Without a doubt, it is a high index, and worrying, and it deserves a detailed study of the same. However, it has other information that deserves to be observed. It is the cancellation of enrollment in courses of federal schools. Figure 2 shows the cancellation rate of enrollment in campus Natal-Central, between 2000 and 2013.

**Figure 1**. Graph showing the Evasion school in Campus Natal-Central of 2000-2013. Source: Author

**Figure 2** Cancellation of enrollment Campus Christmas-Center between 2000 and 2013. **Source**: Author

Figure 2 shows the total number of dropout students and the total number of students who canceled the registration for the years between 2000 and 2013. Analyzing the year 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. Therefore, the registration cancellation index must also be taken into account in the evaluation process of teaching learning.

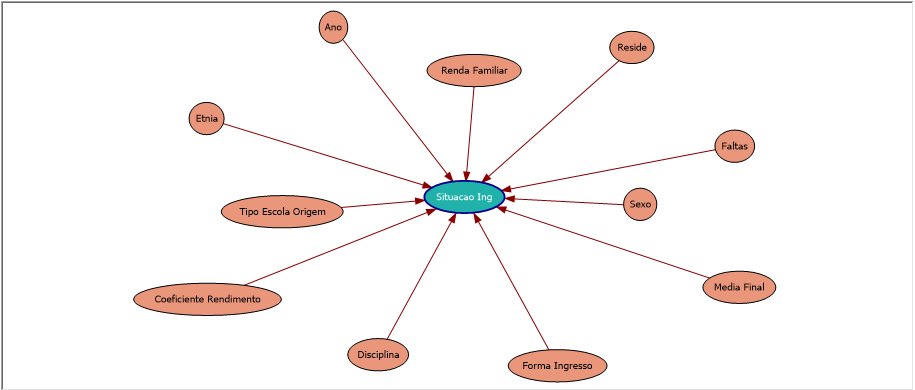
Table 1 shows the percentage of repetition by discipline in 2010. A filter of the data was done to show only percentages of repetition from 40% to 60%. So you see that many disciplines have repetition rate is very high. And as was said earlier, the repetition rate has implications for the truancy because it can often 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 repetition rates and truancy in IFRN, campus Natal-Central. Therefore it will be applied to the same data set, some data mining algorithms, in order to find some relation to the attributes of the database, which can be used to draw a profile of repetition situations and avoidance of our students.

Figure 3 shows a network obtained by the application of decision tree algorithm, using the tool Analysis Services. 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.

**Table 1** Index of Reptition in Courses in the IFRN (Campus Natal-Central) for the year 2010.

**Figure 3** Network Decision tree showing the relationships between the attributes. Source: Author



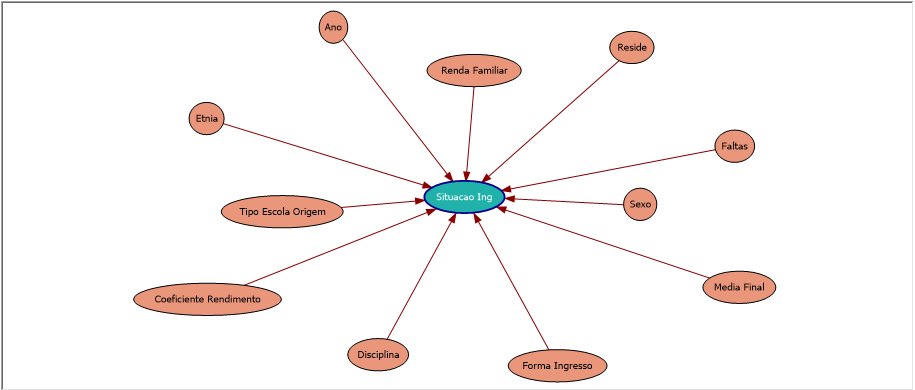
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Disciplina | Ano | Reprovado | Alunos | % |
| Prática como Componente Curricular | 2010 | 17 | 34 | 50 |
| Língua Estrangeira - Inglês | 2010 | 22 | 44 | 50 |
| Psicologia do Trabalho | 2010 | 63 | 127 | 49,61 |
| Algoritmos e Programação Orientada a Objetos | 2010 | 58 | 118 | 49,15 |
| Equações Diferenciais | 2010 | 20 | 41 | 48,78 |
| Conservação de Energia | 2010 | 17 | 35 | 48,57 |
| Autoria Web | 2010 | 228 | 473 | 48,2 |
| Técnicas de Laboratórios de Alimentos | 2010 | 66 | 138 | 47,83 |
| Biologia Celular | 2010 | 18 | 38 | 47,37 |
| Informática I | 2010 | 61 | 129 | 47,29 |
| Cálculo Diferencial e Integral II | 2010 | 80 | 170 | 47,06 |
| Biologia Ambiental | 2010 | 23 | 49 | 46,94 |
| Arquitetura TCP/IP | 2010 | 37 | 79 | 46,84 |
| Química Geral e Experimental I | 2010 | 79 | 170 | 46,47 |
| Sistemas Elétricos | 2010 | 26 | 56 | 46,43 |
| Mecânica dos Solos | 2010 | 117 | 252 | 46,43 |
| Biologia | 2010 | 45 | 97 | 46,39 |
| Estrutura de Dados | 2010 | 24 | 52 | 46,15 |
| Administração de Sistemas Abertos | 2010 | 29 | 63 | 46,03 |
| Elementos de Física | 2010 | 89 | 195 | 45,64 |
| Óptica | 2010 | 26 | 57 | 45,61 |
| Eletricidade | 2010 | 232 | 509 | 45,58 |
| Eletrônica Digital | 2010 | 81 | 180 | 45 |

The Figure 4 shows all the attributes that influence evasion school. So we can draw a profile for truancy, analyzing each of these attributes. The attribute "Source Type of 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 form" indicates how the student entered the course (ENEM, examination, 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.

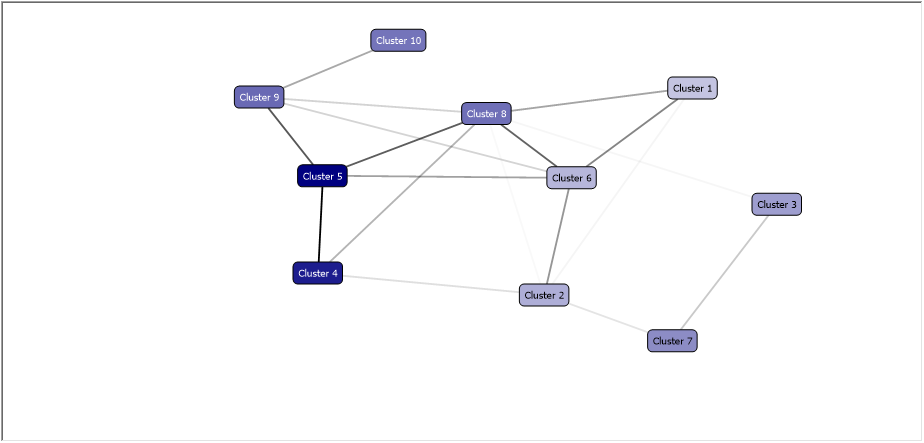
The Figure 4 shows a graph generated by the clustering algorithm using the tool Analysis Services. In setting the cluster algorithm was selected the "Evasion" situation, the cluster with higher dropout cases, appears on the chart with the intense blue color and clusters with fewer cases of evasion with the lighter color.

**Figure 3** Network Decision tree showing the relationships between attributes. Source: Author



As the cluster 5 is what has the darker blue color, it means that it is having the largest number of truancy and Cluster 1 has the lowest number of cases of evasion school. Cluster 1 presents the most approved.

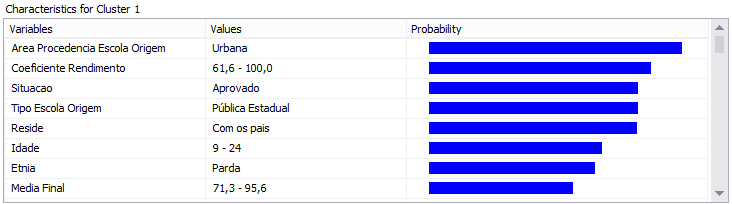
Figure 4 cluster graph generated by Analysis Service. Source: Author



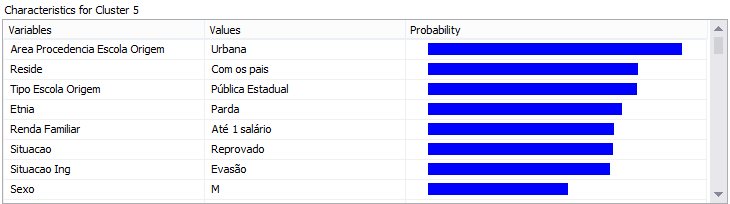
Let's analyse the cluster 1 and 5 to see which were the input attributes that influence the composition of the same. Figure 5 shows the cluster characteristics 1, which is concentrated the largest number of approved. Figure 6 shows the characteristics of cluster 5, which is concentrated the largest number of school evasion.

The cluster 1 shows that the profile of successful students are those with final average above 70, yield coefficient above 60, coming from public school, live with their parents and brown ethnicity. Observing the cluster features 5, realizes that in cluster formation 5, the family income (up to 1 salary) and the situation (deprecated), appear as factors influencing truancy. Justifying thus the presence of these attributes in the attributes list that influence school evasion, as shown in Figure 3.

**Figure 5** Cluster Features 1. Source: Author.



**Figure 6** cluster features 5. Source: Author.



**5 Conclusions**

Based on the obtained results, show that the dropout rate at Natal-Central campus is quite high, above 16% for the years after 2010. One other thing, very important, that until then, no one had observed, is the index of canceled enrollment in courses. If we add the evasion index with the index of canceled registrations, we will have an index above 25%. If we observed that the IFRN has around 5,000 students, so 25% of 5,000 = 1,250 students who dropped out or canceled their registrations annually. And how much is all this represents financially for IFRN? Remember that the IFRN consists of 21 campuses? Logically that the campus Natal-Central, where the data for this study were used, is what contains the greatest number of students and consequently that the number of dropout students on other campuses will be less. However, we have to analyze dropout rates and cancellation of enrollment in other units. If we extrapolate this calculation for the entire federal system, the thing is gigantic, because the federal network consists of 38 federal institutes spread across the country and we have now more than 450 campuses.

Analyzing the graphics obtained by the Data Mining algorithms, at first, one can trace a profile for truancy as being of pupils from state schools, with family income up to one minimum wage, living with parents consequently, they are unemployed or are minors of age, of cor brown race, with very low yield coefficient. It is known in practice that students that entering in the IFRN, comes of the public schools, and arrive with knowledge far short of the desired in basic subjects such as mathematics and Portuguese, which are key to having a good performance in our course. These students face many difficulties to do disciplines that contain logic, abstraction, and advanced math, such as the technical disciplines of technology area.

Based on this profile, one can suggest that the IFRN, Natal-Central campus, adopt some preventive measures to minimize both evasion, as school repetition. Among them, we can mention:

• The result of the analysis should be shared with all faculty of the IFRN, so that all are aware of the real 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 this the discrepancy in these disciplines and thus make plans and goals to create tutoring 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 given continued in the analysis of the academic system data and certainly more knowledge will emerge, and these will be passed at managers of the IFRN, so that action can be taken that will reduce the problem of truancy in our school.

**6 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. 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. Porto Alegre: V. 4 Nº 2, Dezembro (2006).

4. Globo. <http://g1.globo.com/educacao/noticia/2013/11/mec-cria-grupo-para-estudar-evasao-escolar-nos-institutos-federais.html>, [Exame da OAB](http://g1.globo.com/educacao/oab/) , 25/11/2013 08h35 - Atualizado em 25/11/2013 08h35 (2003)

5. Han, Jiawei; Kamber, Micheline. Data Mining: Concepts and Techniques. Second Edition. Elsevier. San Francisco, CA (2006).

6. Karinny, Shintani, Leonardo Armond, Vassily Rolim. [Dificuldades Escolares](http://www.profala.com/arteducesp19.htm). Visitado em 21-5-2014 (2014).

**7. Lilian L. Pereira.** Matéria publicada na Revista Direcional Educador - [*Edição 80 - setembro/2011*](http://www.direcionaleducador.com.br/edicao-80-set/11/edicao-80-set/11) (2011)

8. LOBO, R. L.; HIPÓLITO, O. e LOBO, M.B., Estudo: Evasão no Ensino Superior: Causas e Remédios. Jornal Folha de S.Paulo. (15/janeiro/2007)

9. MEC. Reuni – Restruturação e Expansão das Universidades Federais. Diretrizes Gerais (2007)

10. Ramal, ADREA. Reprovar não é a solução, mas aprovar que 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>, Terça-feira, 09/12/2014, às 08:26 (2014)