A probabilistic analysis of student retention in a Federal University: A Case Study of a Computer Science Program

Marcelo S. Santos[1], **Daniela B. Claro**[1], **Veronica Maria Cadena Lima**[1]

[1]FORMAS - Formalismos e Aplicações Semanticas - Departamento de Ciência da Computação
Instituto de Matemática – Universidade Federal da Bahia (UFBA)
Av. Adhemar de Barros S/N – 40.170-110 – Salvador – BA – Brazil

marceloss@dcc.ufba.br,dclaro@ufba.br,cadena@ufba.br

Abstract. *Due to the increase in inflows mainly because of REUNI procedures and low completion rates observed at a Federal University, it is necessary to identify which courses may cause the students to remain in their programs longer than expected or even leaving the university before their conclusion. In this work, we suggest that the combinations of courses in a semester may not be appropriate thus causing a student retention. We present a case study on analyzing retention rules in a Computer Science program at the Federal University of Bahia. We have estimated the retention probability and our results could help universities to better understand their programs and students.*

1. Introduction

Brazilian universities have seen an increasing in the number of inflows, mainly because of REUNI procedures (Restructuring and Expansion of Federal Universities) [REUNI 2012]. In its design and implementation at the Federal University of Bahia, there was in 2012 an enrollment of over 32,000 students due to the new programs. According to [Filho et al. 2010], the success rates expected in the REUNI procedures were about 90% of the inflows. But success rates in undergraduate programs are still much lower than the desired.

According to [Barros and Mendonça 1998], despite the positive effects of student's failure in a course, it is possible to highlight three negative effects related to it: i) it obstructs the regular educational flow, raising the cost of a student from a University's perspective, ii) to the family's point of view, the student spends more than the average time of the program to have a professional position and as a consequence a financial return to aid their family, iii) student's failure has a negative effect on self-esteem and motivation, increasing the likelihood of future failures.

Due to REUNI and consequently the new programs in the Federal Universities, there is not enough students who have completed their programs, thus making it impossible to analyze the student as retained according to the MEC.

It is also important to note that if a student does not obtain success in a course that is prerequisite for another course on the next semester, he will be unable to take that course on the following semester. This research stated that students are retained when *for every recommended course in a given semester, if the student fails in at least one of the prerequisites of a recommended course which he wasn't enrolled yet, thus he is considered as retained.*

DOI: 10.5753/cbie.sbie.2015.1245

In this work, our goal is to refine the analysis of retention and academic success of students based on previous identified courses [Silva et al. 2013, Santos et al. 2014].

Specifically, this work discusses the following factors:

Considering a course D which has the prerequisite C and C has both prerequisites A and B, the following questions can be posed:

1. Given a student who is **able** to enroll in course C by the recommended semester (including when he failed a prerequisite course but over the time he become able again):
   (a) What is the probability a student could be enrolled in course C or D on the following 1,2,..., n semesters? A recommended semester course is one suggested by the Computer Science flowchart flowchart.
      i. if he was enrolled, what is the probability he can pass, fail or not enroll?
2. Given a student who is **unable** to enroll in course C by the recommended semester (due to not meeting the prerequisites A, B, ..., etc.):
   (a) What is the probability a student will be enrolled in a course that he has failed (prerequisites A and/or B)?
      i. if he was enrolled, what is the probability he can pass, fail or not enroll?
   (b) What is the probability a student can be able to enroll in course C or D on the following 1,2,..., n semesters?
      i. if he enrolls, what is the probability he can pass, fail or not enroll?
3. What is the probability a student is defined as not retained given he had passed in a specific course?

This work is structured as follows: Section 2 presents our related work; Section 3 depicts our background. Section 4 describes our methodology; Section 5 presents our questions which mean our experiments following by our results. Section 6 analyzes our results. Section 7 presents our Threats of Validity and finally section 8 depicts our conclusion and future work.

2. Related Work

Data Mining techniques applied to Educational data have already been used by several authors abroad the World [Romero and Ventura 2010, Baker et al. 2011]. Specifically in the context of retention and academic success, many authors used different techniques and approaches to analyze this problem [Nandeshwar et al. 2011].

Authors in [Nandeshwar et al. 2011] conducted a literature review and suggested using different learning methods. They found that student's family related variables, socio-economic status, high GPA (Grade Point Average) and exam grades had influenced over the student retention problem. Zhang et al. [Zhang et al. 2010] assemble a data-warehouse of three systems (library, online learning and academic), and used Naive Bayes in order to notify students on their potential chance to be retained.

Campello et al.[Campello and Lins 2008] analyzed the duration of 6 years student's socio-economic, enrollment exam and student transcripts. They used clustering algorithms to identify retained student profile at the Federal University of Pernambuco.

Authors in [Manhães et al. 2011] used ten different data mining algorithms to identify those who tend to evade the course of Engineering of the Polytechnic School of UFRJ. They used as variables the highest GPA and students' grades on a first semester course.

Our previous work [Silva et al. 2013], conducted an analysis regarding student retention on the Information Systems course at UFBA. Our work aimed at using association rules to identify which courses can contribute to student retention and on which courses a student tends to be approved or failed. With those information, the authors have suggested to reorganizing their curriculums in order to reduce student retention ratio.

After this, our next approach [Santos et al. 2014] aimed at analyzing the frequent and infrequent rules of all programs at the UFBA and creating a panorama of courses and programs that more retain per semester, thus retrieving such kind of information: which courses have the highest student retention ratio, which semesters have the highest retention rates, which courses most retain on a semester.

In this work we extend these ideas by using a Bayesian network to model the success, failure and consequently the retention of students in a Computer Science program in order to make inference about all students' performance in this program.

3. Background

According to Pearl [Pearl 2014], the Bayesian networks (BN) are representations through directed acyclic graphs that represent the probabilistic relationships between variables to make probabilistic inferences.

In a Bayesian network, each variable is represented by a node in graph and the edges represent the relationships between nodes. Thus, if a node X receives an arc of a node Y, it has a conditional probability defined as the probability P (Y|X). If no variable extend an arc to X, its probability is independent P(X) [Pearl 2014]. These probabilities inform the chances that a state or value of a variable can take. Each variable contains a set of finite states or values that can take over. The probabilities calculated for each node make up the table of conditional probabilities that the values of the node can assume, given the conditioned nodes value to it. Thus, through an BN it is possible to obtain any information relating to any variable in the network [Pearl 2014]. According to [Russell et al. 1995] BNs are a compact way of representing the joint probability distribution of a set of variables.

The definition structure of BN can be performed two-folds: 1) build the entire BN from the knowledge of an expert. However, depending on the domain being modeled, this can be difficult and time consuming; 2) based on a sample which defines the relation among the nodes. In this research, the construction of the BN was carried out by experts.

When there are no missing values in the dataset, and the structure has been specified by an expert, it is necessary to estimate the joint probability distribution of the BN. Thus, the probability should be estimated according to the observed frequencies in a dataset. In this work, the estimation algorithm used was the Maximum Likelihood [Anderson et al. 1986] since it does not consider any prior knowledge and their relative frequencies are on estimative values.

In BN, the term inference is commonly used to indicate the update probabilities throughout the network structure given a set of evidence. According to

[Russell et al. 1995], inference it is a distribution mechanism for calculating a posteriori probability for a set of variables, given a set of evidence, namely random variables with values instantiated.

There are two types of algorithms performing probabilistic inference: approximate and exact algorithms. The former are based on simulation methods to infer the likely resulting in a lower accuracy, but has a higher processing speed. The later produces more satisfactory results, but require a high computational effort when the network has a large number of variables. In this research, we used the clustering algorithm, an exact type of algorithm, to highlighting the accuracy of the data.

The clustering algorithm, also called Junction Tree [Nielsen and Jensen 2009], is the most usual and efficient algorithm of accurate inference. This algorithm divides the data into groups of items that have similar properties, obtaining an optimum result of elimination of variables and creating a structure to propagate multiplications of tables. This algorithm was used to infer the precise probabilities in the network.

Different from techniques that get logical results (true and false) as: decision trees, expert systems, clustering, etc. Bayesian networks bring probabilistic results of a given event occurs. This technique has the advantage among others in the context studied, given the uncertainty in the events, e.g., BN informs that there is a probability of 65 % of students passed and 35 % of them failed, since a technique that brings logical results informs that every student in a discipline X were passed, discarding the possibility of a student fail.

4. Methodology

Our methodology was divided into three steps: 1) data selection and transformation, 2) definition of a Bayesian network and 3) evaluation of our results.

4.1. Data Model

The dataset made available for this research includes all students of the Computer Science program who was admitted from 2004.1 to 2008.2. Only students with the same courses on their curriculum were considered, thus our dataset was limited to students from 2007.2 to 2008.1, it means, all of them following the same curriculum.

There are three entities in our dataset: students, courses and programs. A program offers required and optional courses in a given semester. Courses are identified by a code (i.e. MATA02) followed by a name (i.e. Calculus A), the duration and its prerequisite.

On every semester, a student is required to enroll on one or more courses. Each course has a recommended semester (1st, 2nd,...) which the student may or not enroll. The recommendations are based on various reasons, including the amount of workload and the degree of difficulty in course. Each curriculum has an expected time for graduation.

To conclude a program, students must complete a certain number of hours concerning both types of courses: required and optional. Usually the curriculum is modified over the years and, for each modification, a new version is created.

4.2. Transformation

Concerning the research objectives as well as the Bayesian network, some changes in the dataset were needed.

Relative Semester. From the semester the student enrolled in the program, we defined each new semester as their first, second, third, etc. semester, which can be understood as the semester in relation to the student's enrollment year (e.g. 2009.1, 2009.2, 2010.1, etc.). With this transformation we can now compare students who enrolled on different years, but are facing a similar flow of courses during their programs.

Repeated Course Enrollment Refinement. In this transformation, we now refine each course on multiple courses based on the number of repeated enrollment. For example, students can be enrolled in a particular course for their second third, or fourth time before they are required to leave the program. Doing this, we are then able to distinguish the associated probabilities even for the same course for varying amounts of repeated enrollment.

Semester of Enrollment. On both research questions 1 and 2, we consider how many semesters the student is enrolled for the course (1,2,...,n). Usually students do not enroll in all courses that are recommended for several reasons: the lack of prerequisites, a shock time between a course and another activity, a short time to conciliate the course with an extra course activities,etc. Thus, it is important to identify the semester a student attends a course in an specific time.

After making all the necessary changes in the dataset, the dataset initially used for the construction of Bayesian network is composed of 530 variables (courses taken in a defined semester and a defined time at least one student) and student 581.

4.3. Bayesian Network defined by an Expert

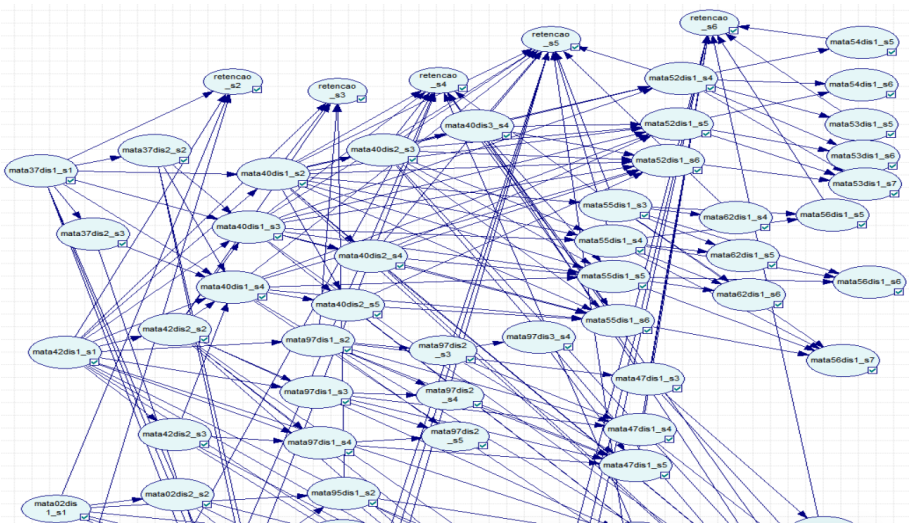

Figure 1. Bayesian Network sample set

Our Bayesian network was defined using all courses except the optional ones. However, the network generated was too complex to be analyzed by the conditional probabilities. Thus, the most important variables were selected to be used. We adopted two exclusion criteria: a) course variables which not have at least 20 students enrolled and b) courses which are not prerequisite to any course on further semesters. For each recommended course in a semester, we selected four courses which have the highest retain ratio based on the work proposed by [Santos et al. 2014].

After removal the variables deemed less important for inference, our Bayesian network was created with 77 nodes and 224 arches supported by Genie 2.0 tool. Only a sample of our Bayesian network (Figura 1) is presented due to page limit.

Our questions and probabilities are associated to the moment a student might enroll on a recommended course. Furthermore, this enrollment, if it occurs, can be for the 1st, 2nd,..., nth time. We can now define each node with the following states: pass, fail or not enroll.

The arcs of our network have been created based on the following principles:

1. if A is prerequisite of B, it creates an arc from A to B;
2. for the student who enrolls in course A in their first semester for the first time, it creates an arc with the following nodes possibilities:
   (a) enroll again in the second semester (A1_s1 $\rightarrow$ A2_s2), third (A1_s1 $\rightarrow$ A2_s3), if he has been failed;
   (b) enroll in B on the second semester which has a prerequisite A (A1_s1 $\rightarrow$ B1_s2), third semester (A1_s1 $\rightarrow$ B1_s3) if he has been passed;
3. prerequisite of a course in the second semester, it creates an arc from A to the second semester retention node (A1_s1 $\rightarrow$ retention_s2).

5. Questions and our Results

Some questions were highlighted in order to conduct our experiments and validate our results.

It is important to note that our probability values are only part of all the possibilities to infer from this Bayesian network.

5.1. Question 1

> 1. Given a student who is **able** to enroll in course C by the recommended semester. What is the probability a student could be enrolled in course C or D on the following 1,2,..., n semesters? If he was enrolled, what is the probability he can pass, fail or not enroll?

According to the curriculum of a student, there are courses which are prerequisites of a further course which has another prerequisite and so on. This leads to a chain of courses that a student might follow. A course is a direct prerequisite to another course when this course is a prerequisite of the latter semester course (A $\rightarrow$ B). A course can be an indirect prerequisite when there is a transitive prerequisite, (i.e A$\rightarrow$B and B$\rightarrow$C). Thus, it is important to understand the impact of student's retention on direct and indirect requirements of a course. The following probabilities were inferred by our Bayesian network:

1. Considering all students who passed in MATA37 and MATA42 course by the first time and the first semester, only 26% of them passed in MATA62 (Software Engineering I recommended in the fourth quarter) when attended by the first time on the fourth semester, 1.7% failed and 72% were not enrolled. Continuing analyzing the group of passed students in MATA37 and in MATA42, only 12% of them were enrolled in MATA62 by the fifth semester where 8.5% was passed and 3.5% failed. Considering those students who take the MATA62 course on the sixth semester (9.6%), only 5.8% passed and 3.8% failed;

2. Considering the students who passed in MATA42, MATA97 and MATA47 in the first semester for the first time, we have a percentage of 70.9% of students enrolled in MATA51 course for the fourth semester by the first time. From this 70.9%, 65.7% students were passed, 5.2% failed and 29.1% of the students were not enrolled. Continuing analyzing the group of passed students in MATA42, MATA97 and MATA47, 10% enrolled in MATA51 in the fifth semester for the first time, iwhere 1.5% passed and 8.5% failed;
3. The students passed in MATA37 and MATA42 in the first semester in the first time, MATA40 the first time in the second semester and MATA55 in the third in the first time, 97% are studying the MATA62 course in the fourth quarter in the first time. MATA37 and MATA42 are prerequisite of MATA40 which is MATA62 prerequisite;
4. For students passed in MATA37 and MATA42 for the first time in the first semester, 80.2% are studying MATA40;
5. The students passed in first time on MATA38 by the first semester, 18% of students are studying MATA48 (Computer Architecture), recommended course in the fourth semester which has MATA38 as a prerequisite.

5.2. Question 2

> 2. Given a student who is **unable** to enroll in course C by the recommended semester. What is the probability a student will be enrolled in a course that he has failed? If he was enrolled, what is the probability he can pass, fail or not enroll? What is the probability a student can be able to enroll in course C or D on the following 1,2,..., n semesters?

The following probabilities are inferred from our Bayesian network:

1. Considering all students who failed in MATA37 and MATA42 course by the first time and the first semester, only 1.2% of them passed in MATA62 when attended by the first time on the fifth semester, 0.03% failed and 98.6% were not enrolled. Continuing analyzing the group of failed students in MATA37 and in MATA42, only 2.7% of them were enrolled in MATA62 by the sixth semester where 2% was passed and 0.07% failed.
2. Considering all students who failed in MATA42 and MATA02 (Calculus A, recommended for the first semester) by the first time in the first semester, 20.4% of them were enrolled in MATA50 (Automata and Formal Language recommended for the third semester) when attended for the first time in the fourth semester only 10% passed, 10.4% failed and 79.6% were not enrolled. Continuing analyzing the group of failed students in MATA42 and MATA02, 13.4% of them were enrolled in MATA50 when attended in the fifth semester, 5.5% passed and 7.9% failed;
3. The students who enroll in MATA42 by the first time in the first semester, 44.5% of them passed, 48.8% failed and 6.5% were not enrolled. From the failure group, 48.8% of them was attending MATA42 by the second time in the second semester, where 27.4% passed and 21.4% failed. Continuing analyzing the group of failed in MATA42, 13.6% of them was attending MATA42 by the second time in the third semester, where 6.3% passed and 7.3% failed. For those who have passed in MATA42 when attended the second time in the second semester, 30.7% passed,

50% failed in MATA97 (which has MATA42 as a direct prerequisite) and 19.2% were not enrolled. About the approved group in MATA42 when attending the second time in the third semester, 15% is attending MATA97 where 5% passed and 10% failed;

5.3. Question 3

> 3. What is the probability a student is defined as not retained given he had passed in a specific course?

The following probabilities are inferred by our Bayesian network:

1. The probability of the student be retained in the third semester given to any event (pass, fail and not enroll) of the courses that imply a retention was 81.4%, 16.5% of not being retained and 2.1% of student does not enroll in any course. Given any event of the courses involving the retention and passed in the course MATA40 attended for the first time in the second semester implies 36.9% to be retained in the third semester and 63.1% not being retained. For the student passed in MATA97 the first time in the second semester implies 70.5% to be retained in the third semester and 29.5% not being retained. If the event of a passed in the last two courses mentioned for the first time in the second semester the student will have 35.9% chance of being retained in the third semester and 64.1% not being retained.

6. Analysis of our Results

The results presented in this work provided a better understanding of the behavior of students' enrollment in Computer Science program at the Federal University of Bahia. We do also analyze the possibility of success and retain ratio over all courses in Computer Science curriculum.

Evaluating our Bayesian network, it was possible to infer that some courses have a large positive and/or negative impact on other courses. This can make the students aware of courses that have a direct impact on retention and consequently on their conclusion within the expected on regular a time.

Initial courses such as (MATA37 and MATA42) have a strong impact on the enrollment of the student in later courses that have as an indirect prerequisite. Students who fail in MATA37 and MATA42 have a very low probability of attending MATA62 on the fifth or sixth semester (section 5.2, item 1). Therefore, most of the students who fail in initial disciplines such as MATA37 and MATA42 do not complete the course at the regular time. It is also important to note that for those students who have passed (section 5.1, item 1) the probability of attending MATA62 on the sixth semester is only 26%, that is, few students tend to complete the Computer Science program on the due time.

It was also noticed that students even fulfilling the requirements to attend some courses are not getting them. When this happens in a high probability as shown in our results (5.1, items 4 and 5), it is critical to be aware of what happened so as to step in.

Another point to be noted is that some courses help the students not to be retained. Some of the courses which contribute most to a no retention are MATA40 and MATA97.

MATA40 contributes to a no retention on the third semester and MATA97 on the second semester (section 5.3 item 1). Other courses contribute to a no retention, but these inferences were not described by page limit.

Our Bayesian network can also contribute to the understanding of the probabilities on failure and success of certain courses. Thus a student can engage into courses that tend to collaborate with their conclusion of the program in a minimum time (sections 5.1 and 5.2). From the University perspective, it is important to understand what are the students' behaviors and thereby create retention policies to monitor and advise students during the semester.

Regarding the university point of view, a detailed probability analysis of retention can increase the number of graduates. As a consequence, we expect that the retention ratio decreases, thus achieving the level suggested by REUNI procedures.

7. Threats of Validity

There are a couple of threats to this work that are important to be highlighted: First, as described in Sections 4.1 selected the most current curriculum of the courses, which caused a reduction in the data set because some curriculums were changed during a recent period. Second, a heuristic retention was defined to characterize the state of retention student in each semester, but when there are not prerequisites in the semester cannot be classified as retained students. Other heuristics can be used for such cases: minimum workload each semester, minimum of approval for courses in a semester. Third, if a program has many courses with many prerequisites the nodes learning algorithm probability can not calculate the probabilities, forcing the expert to reduce network nodes number, making it impossible to map all courses of the program. Fourth, the analysis of intent of this study looked at data regarding student performance. However, socioeconomics and cultural data can be useful for a better understanding of retention at our University. Fifth, the lack of more specific course information limits the strength of conclusions. For instance, it is known and observed that students make choices based on which professor is offering the course. Such choices are not taken into account in our work due to privacy reasons.

8. Conclusion and Future Work

This work was proposed to create a Bayesian network for analyzing the academic retention of students of the Computer Science program.

Our results presented the probability of a student being able or unable to follow a given course in a specific semester given a result obtained in previous courses. Thus we can highlight that the retention in a course can have a high impact than in other course.

We observed that even if some students were able to take a given course in a recommended semester, they do not do it. These cases are complicated because it does not know is the reason the student has not taken the course he should attend, so the identification of these cases is very important to the reduction of retention.

These results aims to assist the college in Computer Science at UFBA to understand student behavior so that you can create retention policies.

As future work we intend to extend these experiments to other courses, such as Mathematical Sciences (Mathematics, Statistics, Information Systems, etc.) and continue

to use a Bayesian network to classify students as retained according to their academic transcript.

References

Anderson, J. R., Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (1986). *Machine learning: An artificial intelligence approach*, volume 2. Morgan Kaufmann.

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19(02):03.

Barros, R. P. and Mendonça, R. (1998). Consequências da repetência sobre o desempenho educacional. *Projeto de Educação Básica para o Nordeste*, Série Estudos(7).

Campello, A. d. V. C. and Lins, L. N. (2008). Metodologia de análise e tratamento da evasão e retenção em cursos de graduação de instituições federais de ensino superior. *XXVIII Encontro Nacional de Engenharia De Produção, RJ, 13p*.

Filho, N. d., Mesquita, F., Marinho, M., Lopes, A. A., Lins, E., and Ribeiro, N. e. a. (2010). Memorial da universidade. Technical report, Universidade Federal da Bahia.

Manhães, L. M. B., da Cruz, S. M. S., Macário Costa, R. J., Zavaleta, J., and Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 1.

Nandeshwar, A., Menzies, T., and Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, 38(12):14984–14996.

Nielsen, T. D. and Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

REUNI (2012). Reestruturação e expansão das universidades brasileiras.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics*, 40(6):601–618.

Russell, S., Norvig, P., and Intelligence, A. (1995). A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25.

Santos, M. S., de Santana, L. C., Pereira, Q. L., Silva, M., Claro, D. B., Lima, V. M., Vieira, V., Ribeiro, S., Telles, A. R., and Lopes, D. (2014). Mining retention rules from student transcripts: A case study of the programs at a federal university. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 25, pages 762–771.

Silva, C. V., Santos, M. S., Claro, D. B., Silva, V., Silva, M., Ribeiro, S., Telles, A. R., and Lopes, D. (2013). Mining retention rules from student transcripts: A case study of the information systems programme at a federal university. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 24.

Zhang, Y., Oussena, S., Clark, T., and Kim, H. (2010). Use data mining to improve student retention in higher education - a case study. In *ICEIS (1)*, pages 190–197.