Using Academic Analytics to Predict Dropout Risk in Engineering Courses

Jhonny Lima[1], Paulo Alves[2], Maria Pereira[2] and Simone Almeida[3]
[1]Polytechnic Institute of Bragança, Portugal
[2]CeDRI – Polytechnic Institute of Bragança, Portugal
[3]Federal University of Technology – Paraná, Ponta Grossa, Brazil
a38178@alunos.ipb.pt
palves@ipb.pt
mjoao@ipb.pt
simonea@utfpr.edu.br

Abstract: The increase of data generated and stored in the educational databases makes it possible to obtain essential information about the teaching and learning process. School dropout and performance problems continue to represent issues which challenge teachers, researchers and higher education institutions to seek solutions. Through the use of academic analytics techniques for data analysis, a sample of 1,844 students between graduates and dropouts on the period between 2007 and 2015 were used as the basis. The methodology followed is essentially quantitative and it allowed to compare student profiles and degrees based on scores, number of attempts and other performance indicators. The data set was processed using Excel software for statistical analysis and R software for data mining using the k-Means and C5.0 algorithms. The propose of a model based on decision trees has as main objectives the generation of standardized instructions, easy interpretation and allow the addition of several possible outcomes, contributing to the decision-making process. The results of this study resulted in contributions which enable higher education institutions to identify students with performance problems and those at risk of dropout and, therefore, allow teachers and course directors to adopt better strategies to increase success and reduce dropout.

Keywords: academic analytics, higher education, dropout, education, engineering

1. Introduction

Students dropout and performance problems have been some of the main themes discussed by higher education institutions around the world. These involve not only characteristics regarding the educational environment, but also characteristics inherent to the methodologies used, for example.

The enormous increase in the amount of educational data generated and stored in the databases of higher education institutions makes it possible to obtain valuable information about the teaching and learning process namely information from students who may be at dropout risk or who need specific activities to increase their success. In this way, it is possible to notice that data analysis is a key factor for understand the situation of each student and to choose the best approach to proceed accordingly.

Handling large data sets is often a slow and complex process. The data mining assists in the process of knowledge discovery through its various algorithms and tools that process this data in order to find important correlations between them. In this sense, the search of how to better understand the data generated by the students, how to predict their behavior, and how to improve teaching and learning, makes Academic Analytics an essential area nowadays.

Through the application of Academic Analytics techniques, this work aims to identify patterns related to the academic performance of the engineering students from a public Portuguese higher education institution. The objective is to identify the profile of the students who dropped out and, thus, to implement a model for the classification of dropout risk based on decision tree.

In order to achieve the research objective, the following goals were defined:

- Assess the evolution of dropout, and describe how it behaved over the nine years;
- Identify possible characteristics that can show that the age of access is a factor that can influence the dropout;
- Assess the similarities and differences between dropout and graduate students;

- Define the profile of the students that are at drop out risk.

The data supporting this research refers to the period from 2007 to 2015 and was based on a sampled of 1,844 students where 745 are students who dropped out and 1,099 graduate students. The main variables under study are: age of access, number of course units approved, number of attempts until succeed, and the mean of the classifications obtained in the course units approved.

This paper is organized into the following main topics: Theoretical Framework, Methodology, Results, Final Considerations, and References. Hereinafter is the development of these topics.

2. Theoretical framework

It is known that education is responsible for a large part of social development, that is, it offers sustainability for a society that wishes to evolve in an intellectual, economic, human and structural way (Prim & Fávero, 2013). The access of different publics to higher education has placed higher education institutions in the face of new challenges and responsibilities, in particular guaranteeing equal conditions of access and academic success.

The term school dropout, besides allowing different interpretations, can be applied in several contexts with slightly different meanings. In certain cases, it is considered as dropout the simple suspension of the relation between the student and the institution before the end of the process for the conclusion of the course (Rigo et al, 2014).

According to Quinn (2013), there are six key factors that lead students to drop out: sociocultural, structural, political, institutional, personal, and learning. The same author states that all these factors are interrelated, for example, personal factors, such as working during the studies, are determined by structural factors, such as poverty.

In the view of Benavente et al (1994), the profile of the student at risk of dropout usually shows a significant school delay, lack of school ambitions, lack of interest in school, subjects and classes. The student at risk of dropout is generally older than others in the same educational level, does not feel supported by the family, lives in an intellectually disadvantaged family environment and has, of course, an insufficient school performance.

As higher education institutions collect more and more data about their students, we enter in a new age of data use to improve student success, streamline processes, and utilize resources more efficiently. Once the data is analyzed, it is possible to obtain better student placement processes, more accurate enrollment forecasts and early warning systems that identify and help students at risk of failure or dropout (Matsebula & Mnkandla, 2017).

Academic analytics makes use of methods of statistical analysis, data mining, and predictive modeling to reveal and recognize hidden patterns in large educational databases. These standards allow us to better understand various educational aspects, such as student behavior and learning outcomes with better accuracy (Joshi et al, 2016). In other words, academic analytics is the application of business intelligence tools and strategies to guide decision-making practices in educational institutions. The purpose of an academic analytics program is to assist those in charge of strategic planning in a learning environment to effectively measure, collect, decipher, report and share data so that operational and student strengths and weaknesses can be identified.

Early identification of students at risk of dropout using data mining algorithms, have a very significant value in terms of preventive measures that the institutions can adopt to reduce dropout. That said, two of the key algorithms used in this task are k-Means and decision trees.

3. Methodology

This study of quantitative nature which aims to explore the academic data of students of engineering courses of Civil Engineering (CE), Renewable Energy Engineering (REE), Electrical and Computer Engineering (ECE), Informatics Engineering (IE), Mechanical Engineering (ME) and Chemical Engineering (CHE), from a public higher education institution in Portugal. From this, a model was created that allows applying academic analytics methodologies to identify the profile of students at risk of dropout.

The proposed analysis was reported for the period 2007 to 2015, taking into consideration the students already graduated and the students who dropped out. A dropout situation is characterized by a student who did not

renew his/her enrollment in the current school year and therefore did not complete the course. On the other hand, a graduate student is a student that got his degree diploma.

The data was treated anonymously and in accordance with the General Data Protection Regulation. Data Mining was performed using the k-Means and C5.0 (decision tree) algorithms.

From the data provided, it is possible to observe that between the years of 2007 and 2015, about 745 students dropped out, where 610 are male and 135 females. On the other hand, 1,099 students have completed the studies, of which 822 are male and 277 are female. Table 1 contains the information for each course.

Table 1: Characterisation of the courses according to sex and age of access

Course	Graduates			Dropouts		
	Male	Female	Age of Access (Mean)	Male	Female	Age of Access (Mean)
CE	183	91	23.14	154	52	24.5
REE	106	35	19.42	68	20	22.7
ECE	146	9	23.41	110	4	25.1
IE	160	34	20.89	143	21	21
ME	203	22	21.91	117	8	22.6
CHE	24	86	23.12	18	30	20.9

The data in Table 1 reveals that the age of access of students who dropped out tend to be higher than the age of access of the graduates, indicating this is a factor that is strictly related to dropout. In our data set, the age at which the student starts to study is highly variable and it has a strong influence in several performance indicators. In this way, it is possible to obtain answers to the following objective: identify possible characteristics that can show that the age of access to the studies is a factor that can influence the dropout.

4. Results

In order to characterize the courses regarding their dropout rate, Figure 1 was generated based on the total number of graduates and dropouts of each course. It is possible to notice that, in general, all courses have more than 50% of graduate students, however, they also have a high dropout rate, especially the Civil Engineering, Electrical and Computer Engineering, and Informatics Engineering courses which have more than 40% of dropouts over the nine years. Therefore, it is concluded that the courses with the largest and smallest number of dropouts are Informatics Engineering and Chemical Engineering, respectively.

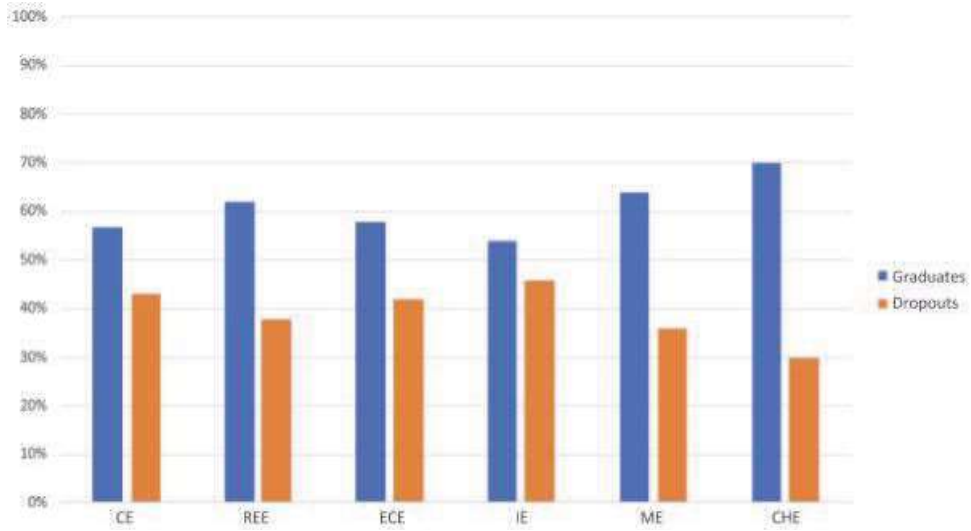

Figure 1: Percentage of graduates and dropouts per course

Figure 2 shows a more detailed analysis, based on the number of years registered of the dropout students over the nine years and per year, which shows that a high percentage of dropouts occurs before the students are enrolled for a year in their respective courses. On the other hand, it can be noted that a large number of dropout students even after they are registered at a much longer time than necessary for the conclusion of the course.

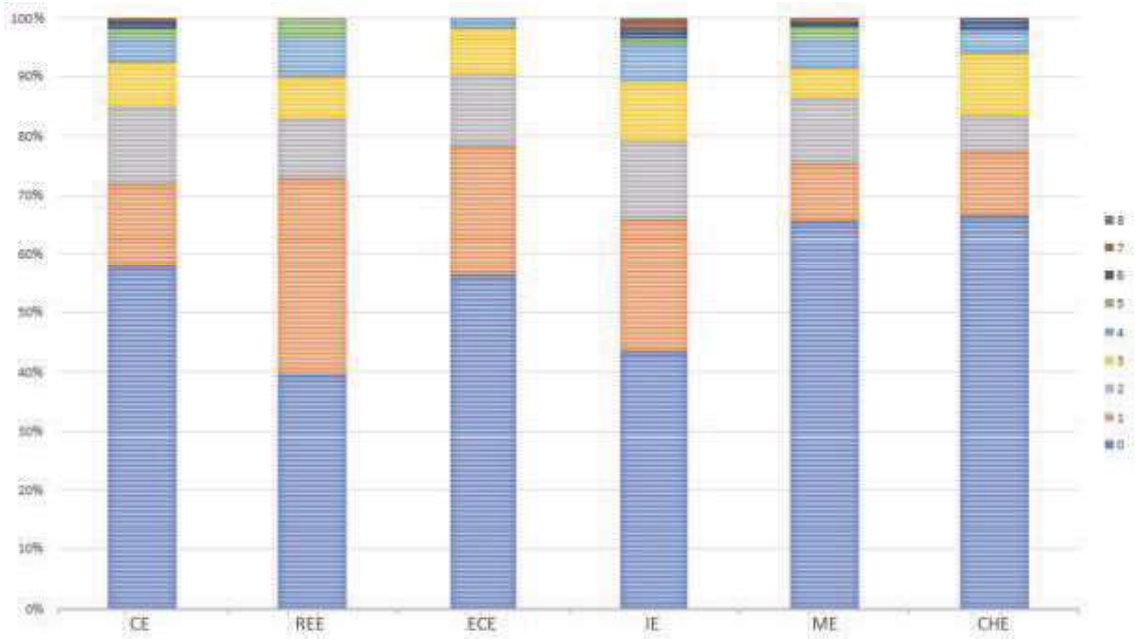

Figure 2: Percentage of dropouts per years registered

In order to make a comparative analysis between the dropouts of all courses and taking into consideration the mean of the number of attempts for approval in the subjects, the graphic of Figure 3 was created.

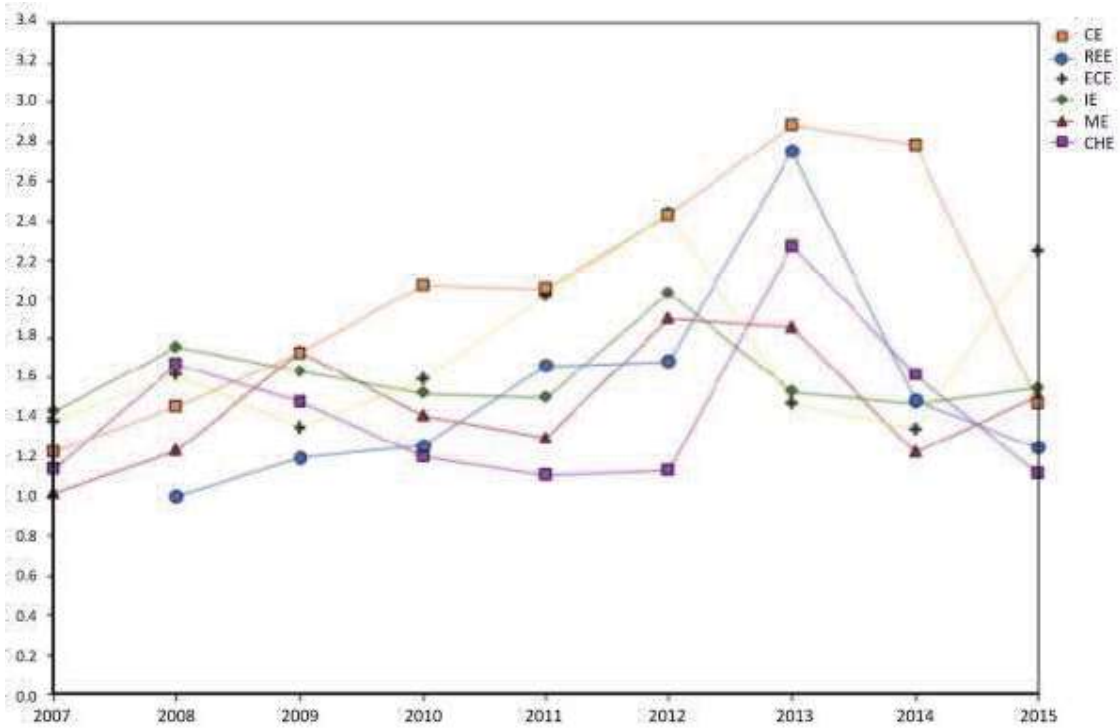

Figure 3: Comparative analysis of the mean of attempts for approval in the subjects of the dropouts

Figure 3 allows us to observe many variations in all courses, however, it is worth mentioning the Civil Engineering course where, in most years, the mean of attempts has been higher than the mean of all other courses. It should also be highlighted the years 2011 and 2012, where the mean of attempts has grown significantly. Likewise, the mean of attempts of many courses started to suffer a great fall between the years of 2012 and 2013.

As a consequence of previous analyzes, it can be concluded that the main characteristics that can help to distinguish between a student with tendencies to complete the course and a student who may be at risk of dropout are: the age of access, the number of approved subjects, the number of attempts for approval, and the

mean of grades. It should be noted that no relevant gender difference was identified, however, it is a factor that should be taken into account.

Based on these characteristics, the k-Means algorithm was used in order to identify the groups of students and the characteristics inherent to each one of them. By applying the Elbow Method, we determined that the ideal number of clusters for our data set is equal to 4. Thus, as a result of grouping according to the age of access, Table 2 presents the values of the centroids obtained.

Table 2: Centroids resulting from the k-Means algorithm

Cluster	Age of Access	Number of Subjects	Number of Attempts
1	36.61	7.25	7.68
2	20.27	3.61	4.21
3	25.62	23.46	35.97
4	21.34	10.15	15.55

The data in the Table 2 reveal that: students who start studying late try to attend a few subjects (most likely are student workers); the 25-year-old students have been studying longer and therefore have a greater number of attempts; most students between the age of 20 years old drop out before even starting to study; students between these last two profiles tried less.
Based on this, for the implementation of a classification model of school dropout, we used the decision tree C5.0. The construction of decision tree was based on the age of access, the mean of grades and the mean of number of attempts for approval in the subjects. It is possible to identify if a student has the characteristics of a graduate student or a dropout student and, from this, it verifies if the student is at risk of dropout.

After some tests, it was noted that the ideal division for the training and testing phase was 70% and 30%, respectively. The Figure 4 show the result of the training phase of the algorithm. It is observed that the size of the tree is 12 and, of the 1,159 cases, the algorithm correctly classified 1,009, having a margin of error of 12.9%. It is possible to notice that of the 335 cases of dropout, the algorithm correctly classified 203, whereas of the 824 cases of graduates were accurately classified 806.

```
Evaluation on training data (1159 cases):

     Decision Tree
   ----------------
   Size      Errors

     12  150(12.9%)   <<


    (a)   (b)    <-classified as
   ----  ----
    203   132    (a): class Dropout
     18   806    (b): class Graduate


Attribute usage:

100.00% MeanGrades
100.00% MeanAttempts
 31.06% AccessAge
```

Figure 4: Result of the training phase

At the conclusion of the training, the other 30% of the data set was applied in the testing phase of the algorithm. The results of the classification show that of the 223 cases of dropout, 150 were correctly classified, and of the 275 cases of graduates, were precisely classified 256. Therefore, it is concluded that the proposed model has an accuracy of 67.3% to identify characteristics related to dropout, while 93.1% for graduates.

5. Final considerations

School dropout and low performance are problems faced by many higher education institutions. The concern on how to identify students with learning difficulties and at risk of dropout represents a challenge for teachers, researchers and educational and training institutions. Therefore, a research work was carried out, by using academic analytics with the aim of analyzing the data from a sample of 1,844 undergraduate students regarding

their age of access, number of course units approved, number of attempts for approval, and the means of the marks obtained in the course units approved.

In light of the results obtained from the exploration of the referred data and the implementation of the model for prediction of dropout, we concluded that:

- Most dropouts occur before students are enrolled for 1 year in their courses;
- Graduates usually have more attempts than dropouts;
- In most years the number of graduates was higher than the number of dropouts, however, the total number of dropouts is significantly high;
- The model identifies more easily the profile of students with tendencies to complete the course.

In this study, just academic information was used because it is considered that only this kind of data is related with factors that can somehow be controlled by pedagogical strategies. This study aimed to confirm existing suspicions, and the numbers will be confirmed after updating the data set to the current date.

The development of a model based on decision tree allows, after being trained, to identify a student who has a profile that is at risk dropout more quickly, contributing significantly to higher education institutions in the decision-making process.

As far as the future works are concerned, we suggest adding qualitative data to data mining process and using learning analytics techniques to analyze the student's learning path using other sources of information such as classroom attendance and use of the virtual learning environment.

References

Benavente, A., Campiche, J., Seabra, T. and Sebastião, J. (1994) *Renunciar à Escola: O Abandono Escolar no Ensino Básico*, Fim de Século, Lisbon.

Joshi, M., Bhalchandra, P., Muley, A. and Wasnik, P. (2016) "Analyzing Students Performance Using Academic Analytics", *Proc. 2016 Int. Conf. ICT Business*, pp. 0-3.

Matsebula, F. and Mnkandla, E. (2017) "A Big Data Architecture for Learning Analytics in Higher Education", *2017 IEEE AFRICON*, Cape Town, pp. 951-956.

Prim, A. L. and Fávero, J. D. (2013) "Motivos da evasão escolar nos cursos de ensino superior de uma faculdade na cidade de Blumenau", *E-Tech Tecnol. Para Compet. Ind.*, Vol Special, pp. 53-72.

Quinn, J. (2013) "Drop-out and Completion in Higher Education in Europe: among students from under-represented groups", [online], European Commission by the Network of Experts on Social Aspects of Education and Training NESET, European Union, https://edudoc.ch/record/110174/files/dropout.pdf

Rigo, S. J., Cambruzzi, W., Barbosa, J. L. V. and Cazella, S. C. (2014) "Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios", *Rev. Bras. Informática na Educ.*, Vol 22, No. 2, pp. 215-224.