Visual analysis to compare academic performances of quota and non-quota students from computer-related programs

Pedro B. Pio[1], **Igor C. Sodré**[1], **Vinicius R. P. Borges**[1]

[1]Universidade de Brasília - Departamento de Ciência da Computação
Brasília – DF – Brazil

pedro.pio@aluno.unb.br, csodreigor@gmail.com, viniciusrpb@unb.br

Abstract. *The implementation of affirmative actions in public universities is a topic of debate within the Brazilian society, specially regarding the academic performance of students that have been admitted through the quota system. This paper describes a visual analysis process to explore and compare the academic performances of quota and non-quota students from computer-related programs in a public Brazilian university. The results revealed that both failure and dropout rates for quota students are slightly higher than non-quota students in the first terms, but tends to present similar rates at the final terms.*

Resumo. *A implementação do sistema de cotas em universidades públicas é um tópico de debate na sociedade brasileira, especialmente no que concerne ao desempenho acadêmico dos estudantes que foram admitidos pelo sistema de cotas. Este artigo descreve um processo de análise visual para explorar e comparar o desempenho entre estudantes cotistas e não-cotistas em cursos relacionados com a área de Computação em uma universidade pública brasileira. Os resultados evidenciaram que as taxas de reprovação e evasão entre os estudantes de cotas é levemente superior em relação aos demais, no entanto estas taxas apresentaram valores similares nos períodos finais dos cursos.*

1. Introduction

The implementation of affirmative actions in university and educational institutes has been extensively discussed around the world. Crosby et al. [Crosby and Cordova 1996] stated concerning affirmative actions in education: "Whenever an organization expends energy to make sure that females and males, people of color and white people, or disabled and fully enabled students have the same chances as each other to be educated, then the organization has a policy of affirmative action in education".

In the last decade, Universidade de Brasília (UnB) implemented its first selection process based on affirmative actions, in which 20% of the approved candidates belonged to the racial quota [Baggi and Lopes 2011]. Eight years later, the Brazilian Government approved the federal law 12711, demanding that public universities should reserve 50% of all admissions for racial and social quotas. Since then, UnB has been admitting students of different social and financial conditions, thus generating important discussions concerning the academic performances and dropouts of quota students [Bailey and Peria 2010].

Over the years, the technological evolution of computer systems has enabled universities to collect and store large amount of students' records. The analysis of such diverse and complex data by human specialists has became difficult, once such data can

settle several questions concerning the academic performance of quota students. This challenging scenario demands the use of intelligent approaches in order to obtain relevant implicit knowledge within the data. For that purpose, data visualization appears as a powerful strategy by generating visual outputs of data, so that the implicit patterns and local structures of data are conveyed intuitively to users and domain specialists [Keim 2002].

The literature in educational data mining (EDM) has reported several researches for analyzing educational data [Romero and Ventura 2020]. [Valente and Berry 2017] explored the relationship between students' performances and the manner of their admittance concerning the results of the Brazilian National Survey of Student Performance (ENADE). The results reported that students admitted by affirmative actions presented similar performances to the others students in public universities. [dos Santos and Queiroz 2016] analysed the students admitted by affirmative actions in Univ. Federal da Bahia, which verified that in 61.1% cases of an engineering course, the quota students presented equal or better performances in relation to non-quota students. [Costa et al. 2017] applied some classification models to determine the failure probability of a student in different courses, obtaining an accuracy up to 83% in two courses.

The relevance of affirmative actions in universities motivated us to propose a method based on visual analysis for obtaining implicit knowledge from the academic datasets, but focusing on the academic performances of quota and non-quota students. The key idea is that the generated graphical representation can convey implicit information and patterns between quota and non-quota students from computer-related courses at UnB. Therefore, the obtained knowledge can guide specialists in tasks related to academic management or public and social policies.

This paper is organized as follows. Section 2 details the proposed method. Section 3 presents the results and the obtained knowledge from the visual analysis. Section 4 reports the final considerations and possibilities for future work.

2. Proposed Method

Figure 1 depicts the steps of the proposed method, which is based on the Cross-industry standard process for data mining (CRISP-DM) [Wirth and Hipp 2000]. The steps of the proposed method are described in the subsequent sections. The difference of the proposed method in relation to CRISP-DM is the employment of the "Data Visualization" step instead of "Data Mining", since our goal is to perform a visual data analysis.

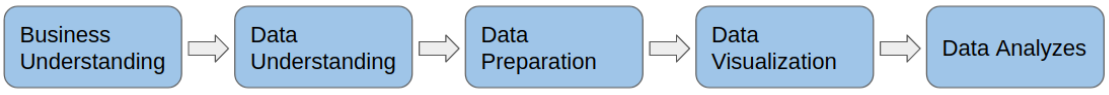


Figure 1. Flowchart of the proposed method.

In the **Business Understanding**, we define the hypothesis research and the knowledge discovery process to be employed. As the focus is to identify relevant and meaningful knowledge concerning the academic performances of students from affirmative actions, we formulate the following question Q: ***Is it possible to compare quota and non-quota students concerning their academic performances and dropout rates?***

In **Data Understanding** is important to take decisions regarding the specific techniques to be employed further. After the dataset is loaded, a preliminary data analysis is

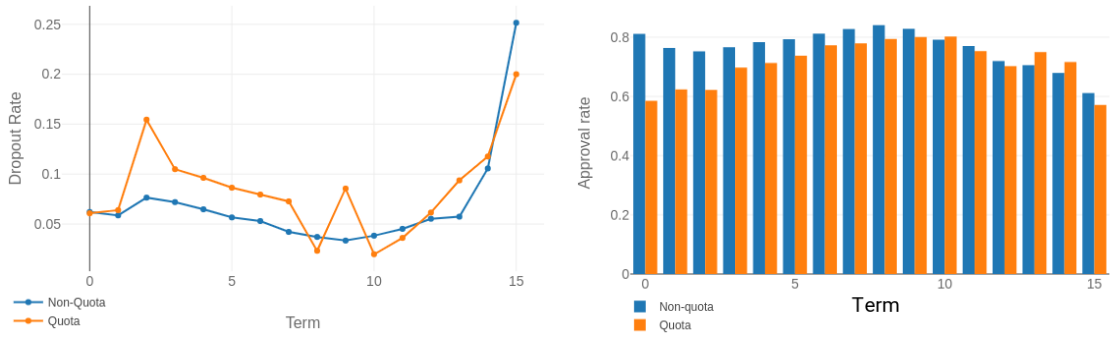
conducted using appropriate tools of Pandas library[1]. The raw data was provided by UnB and contains 7683 records, each one describing a course enrolled by a single student. For confidentiality purposes, we received the dataset presenting the personal identification of all students fully encrypted. Among these records, 1067 were identified as belonging to quota students. Each record is denoted by 25 attributes and characterizes the academic performance of a student in a single course, along with personal and admission factors.

The **data preparation** consists of applying traditional preprocessing techniques in the raw data. First, irrelevant and redundant attributes are removed. After that, categorical attributes are transformed to numerical ones, such as "date of birth" and "year and term of the course". In these cases, the new attributes "age" and "current term" are created respectively. Finally, the "course exit's forms" were joined to compose a new categorical attribute, comprising three different values (active, graduate, abandonment), thus obtaining dataset D_1 consisted of 18 attributes: Student Id; gender; quota status; prior school type; race; university admission term; course admission term; course output term; course output form; year and term when the course was taken; mean of student term; total of approved credits; total of credits approved during the term; course name; course grade; student age; semester and if the student left the university. Another dataset D_2 was derived by representing each student into a dataset row and its respective cells were created by counting and registering the quantity of times the student enrolled in each course.

The **data visualization** comprises the visual analysis process, which is conducted by generating graphical representations (also called layouts) of the underlying data in order to obtain knowledge from the research goal. For that purpose, in such process, we incorporated three visualization techniques: line chart, bar chart and heatmap. Specifically, the target attributes are those related to the students performances and other academical factors which are interesting to analyze: average of times the student groups enrolled in a course; the correlation between number of times a course is coursed along with the students' exit form; the dropout rates; and approval rates.

In the **data analysis**, the user interprets the generated layouts, thus obtaining implicit and useful information from the data that can validate the hypothesis researches.

3. Experimental results



(a) (b)

Figure 2. (a) Dropout rate comparison between quota and non-quota students; (b) approval rate comparison between quota and non-quota students.

[1]Python Data Analysis Library `https://pandas.pydata.org/`

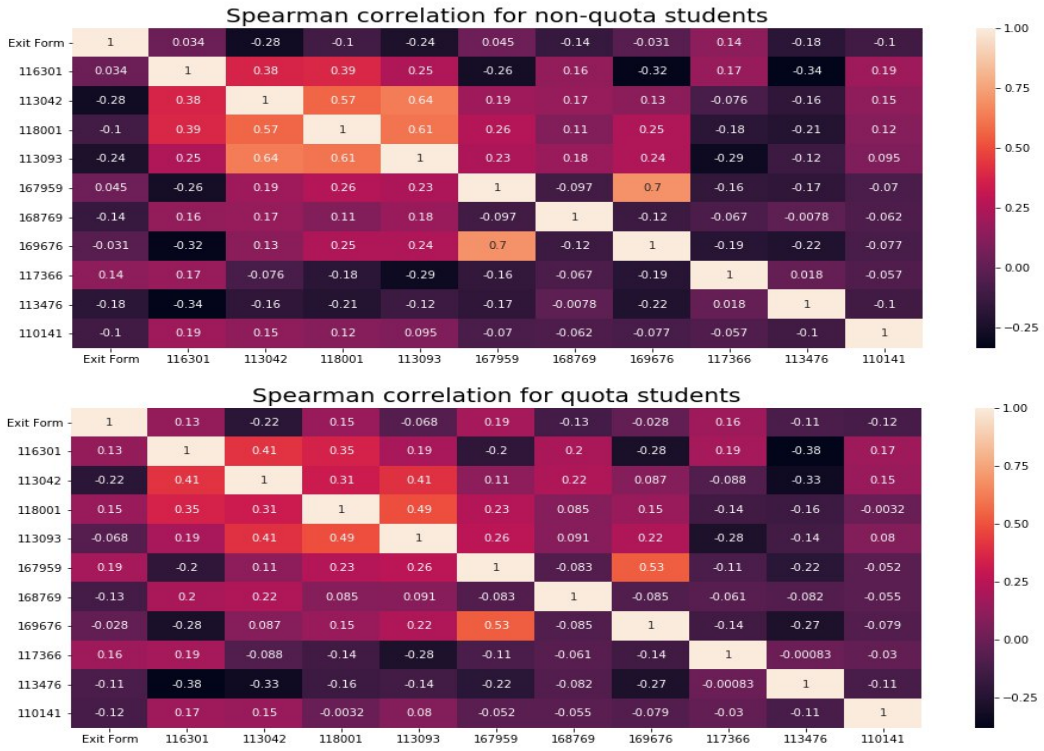


Figure 3. Correlation between quantity of enrollment of the main courses (denotes by the numerical identifiers) and exit form.

In this section, we present the generated layouts of the proposed method in order to answer question Q. Generated from D_1, Figure 2(a) displays the relation between dropout rates of all students and the terms. The chart was built by dividing the total number of dropouts that were currently in that term by the total number of students also in that term. As shown by the chart, the overall dropout rates for quota students is higher than for non-quota students, while it is noticeable that both dropout rates decrease as the terms progress. The differences between such dropout rates can be possibly justified by financial problems, lack of interest in the course and lack of time to enroll classes.

Also created from D_1, Figure 2 (b) shows the average approval rate of quota and non-quota students for each term. Each bar height were computed by dividing the quantity of courses that got a passing grade or higher by the total number of courses taken that semester. It can be noted that the approval rates of quota students are lower in the first terms, in which the difference among the quota students decreases during the next terms.

The pattern of the bars changes noticeably after the third term, which can be associated to the dropout rates regarding the students with lower performances that do not finish the course and leading to progressively increasing the approval rates of the final terms. Moreover, the first terms of related programs present courses that demands high theoretical background, which may not be previous acquired by the enrolled students.

Analyzing D_2, we selected the most frequent enrolled courses in the dataset, we computed the Spearman correlation between those courses and the exit form of the students. Figure 3 presents the obtained correlation matrix as a heatmap, in which lower correlations indicate that students who enroll more often in the program courses are asso-

ciated to lower chances of dropout, while higher correlations are related to higher chances of dropout. The analysis of the heatmap shows that there are no strong correlations (higher than 0.5 or lower than -0.5) between those courses and the exit form, so there is not a relation of specific courses that motivate the dropout of undergraduate students. The heatmap depicts correlations between some courses, such as Physics and Linear Algebra, and, Computer Networks and Introduction to Computer Engineering, for both cases.

4. Conclusion

This paper proposed a visual data analysis process for comparing the academic performances between quota and non-quota of undergraduate students from computer-related programs. The proposed method is a modification of CRISP-DM, in which visualization is employed instead of data mining, thus visualizing students' records, so that we can verify their progressions during the terms and identify aspects that can lead to their dropouts.

The layouts revealed the courses in which both quota and non-quota students present higher failure rates. The heatmap of the Spearman correlation showed that those courses are not relevant concerning the dropout events. This study can support university managers and specialists in education to make decisions by considering the students' difficulties and strategies for preparing them to the first terms, which present more dropouts.

Future work will concentrate efforts on applying this study to more recent data and including interactive resources on the visualization strategies.

References

Baggi, C. A. d. S. and Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *R. da Avaliação da Educação Superior*, 16(2).

Bailey, S. R. and Peria, M. (2010). Racial quotas and the culture war in brazilian academia. *Sociology Compass*, 4(8):592–604.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., and Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73:247–256.

Crosby, F. J. and Cordova, D. I. (1996). Words worth of wisdom: Toward an understanding of affirmative action. *Journal of Social issues*, 52(4):33–49.

dos Santos, J. T. and Queiroz, D. M. (2016). The impact of the "quota system" in the federal university of bahia (2004-2012). *Creative Education*, 7(17):2678.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1355.

Valente, R. R. and Berry, B. J. (2017). Performance of students admitted through affirmative action in brazil. *Latin American Research Review*, 52:18–34.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29–39. Springer-Verlag London, UK.