Compare students from Univeersity of Brasília by Gender Using t-SNE Techniques

Luiza Hansen[1], **Vinicius R. P. Borges**[1], **Aleteia Araujo**[1], **Maristela Holanda**[1]

[1]*Dept. of Computer Science - University of Brasilia*
Brasilia, Brazil

luizaahansen@gmail.com, viniciusrpb@unb.br, aleteia@unb.br, mholanda@unb.br

Abstract. *The presence of women in technology-related courses is declining every year, reaching in 2016 less than 20% of the total student body in the field of Computer Science in Brazil. This paper proposes to study visualization techniques to analyze and identify profile patterns of undergraduate students on courses in the computing field, comparing gender data (female and male). A visual data and a quantitative analysis were used with dimension reduction technique (t-SNE) considering the students situation at University of Brasília (active, drop out, graduated, death and others). The layouts revealed that the high number of students leaving without graduating is not a gender-related problem. Also the quantitative analysis shows that being in the group of quota students does not have a significant bearing on whether students leave with or without graduating.*

1. Introduction

Women in Computer Science is an important research topic, since only 18% of female students in Brazil finished programs in different computer courses in the year 2016 [SBC 2016] according the SBC (*Sociedade Brasileira de Computação*) which has analysed INEP (*Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*) data and data obtained from the Census of 2016 Higher Education. In UnB (*University of Brasília*) the Computer Science, Computer Engineering and Licenciate in Computer courses have a small percentage of women compared to men, less than 9% in 2014.

Ada Lovelace, ENIAC girls, sister Mary Kenneth Keller and Grace Hopper are great women who influenced the development of the technology used nowadays. Compared to other courses, Computer Science was in the early days a promising area for women, having more female participation than courses such as physics and medicine [Henn 2014]. However, since the 1990s, the participation of woman in computing programs has being decreasing.

Visualization techniques help to analyze and understand multidimensional data, through the generation of graphical representations and interaction mechanisms, which reveal characteristics or patterns of data. Thus, information visualization techniques can be used to: understand information quickly, making it possible to visualize large amount of data in a comprehensible and cohesive way; identify latest trends, facilitating the discovery of atypical values; identify relationships and patterns, which due to the graphic presentation, indicate coherence in large volumes of data; involve the user and transmit the message quickly, through the impact caused by diagrams, graphs and other visual representations [Estivalet 2000].

DOI: 10.5753/cbie.sbie.2019.1751

Data mining techniques have been used to discover patterns and implicit knowledge in educational databases, in other to develop methods that assists exploration of datasets collected in pedagogical environments [Baker et al. 2011]. Moreover, data mining have also been used in studies that analyse the profiles and characteristics of women in Exact courses, thus assisting education specialists. The use of visualization techniques can benefit these tasks, since the human visual system identifies and interprets relevant patterns in graphical representations more quickly than descriptive or exploratory analysis.

The goal of this article is to analyse and compare the data of male and female students on computer courses at UnB, using visual techniques based on t-SNE (*t-Distributed Stochastic Neighbor Embedding*) as explained in Section 3, employed to reduce the dimensionality of the database. Our research questions are: Is there difference between quota students an no-quota students about drop out rate? Is there difference between female and male student about drop out rate?

The remainder of this paper is structured as follows: Section 2 presents related works; Section 3 describes the methodology employed in this research; Section 4 presents the main outcomes; finally, Section 5 presents the conclusions and future work.

2. Related Work

Several studies addressing educational data mining and the gender issue in Computer Science have been presented in literature. This section describes previous research concerning the decrease in the number of women in Computing Majors, and also a few studies on educational data where the main goal is increasing students' performance and progress, understanding motivation, predicting which students will drop out, and what works to help the instructor understand the discussion topics and which students are participating. The works make use of visualization and/or prediction algorithms.

In [Dominguez et al. 2018] the SPEET (Student Profile for Enhancing Engineering Tutoring) project is presented, which seeks to determine and categorize the different engineering profiles in Europe to improve tutoring methods and help students achieve better results and graduate. In that work a prototype was implemented that uses coordinated histograms and interactive dimensionality reduction. The visualization of the coordinates helps the educators to understand the real influence of the nature (compulsory or elective) and the methodology of the courses (practical or theoretical), besides visualizing the mobility and the distribution of the notes. Also dimensionality reduction was used in pattern recognition of the data.

[Baker et al. 2010] presents several methods within educational data mining that normally fall into five categories: prediction, clustering, relationship mining, discovery with a model, and distillation of data for human judgment, explaining these categories and the differences between them. They also explain that in the distillation of data, the data are arranged so as to enable humans to identify easily implicit patterns.

[Stout et al. 2016] and [Cheryan et al. 2013] provide studies about stereotyping in Computer majors, also stating that there is a higher ratio of men than women in this field in the United States (US). Likewise, [Mercier et al. 2006] present surveys, drawings, and interviews which were used to examine the perception of US middle school students

about characteristics of experienced computer users. These results showed the cultural stereotypes of a computer user: 89% were male and 94% wore glasses.

[Papastergiou 2008] used descriptive statistics, principal component analysis and analysis of variance to investigate 358 Greek high school students' intentions and motivation for pursuing academic studies in CS. This study looked into several factors, such as the influence of family and academic environment on their career choices, their perception of a professional career in CS, and their self-efficacy beliefs regarding computers. The analysis showed that a lack of exposure and use of a computer at home and in school from early stages in the students' lives seems to be the main factor in discouraging them from studying CS, in particular when considering the data for women.

[Hansen et al. 2018] aimed to identify relevant patterns using a visual technique named Andrews Plot considering three profiles of women: active students, drop out students and graduated students, concluding that those who left without graduating have similar features with those who graduated. Holanda et al. [Holanda et al. 2017] also studied female profiles in UnB, presenting a summary of the profiles and analyzing the three indicators: entering the courses; student dismissals; and academic achievement. In academic performance, male and female studets have a close average of grades, but looking at all the course subjects female students have, on average, a slightly better performance than male students.

[Borges et al. 2018] described a method for students' performance prediction using principal components analysis (PCA) and classifiers Support Vector Machines and Naive Bayes. Experiments were performed to compare the prediction was between the high (original) and the reduced dimensional datasets. The students' data reveal some personal and academic variables that might influence their performance, such as failures, period grades, mother's education and alcohol consumption. Also, results indicated that the dataset with reduced dimensionality retained the most relevant information and relations among attributes.

Given this alarming decline and the widening disparity between male and female representation in the field of Computer Science, this paper aims to investigate female students in the Computing Major at UnB in Brazil. It used visualization techniques to analyze and identify profile patterns in female undergraduate students.

3. Methodology

Considering the goals of this research, the methodology proposed is based on the Knowledge Discovery in Databases (KDD) to identifying patterns in student data. The original methodology involves the following steps: selection, preprocessing, transformation, data mining and interpretation/evaluation, as shown in Figure 1. In this paper, the data mining step was replaced by data visualization. These steps are performed interactively and iteratively, as they involve the cooperation of the data analyst and the fact that this process is not applied sequentially, requiring parameter selection constantly [Corrêa and Sferra 2003].

The selection phase is a way to include the main university computing courses, and define that Computer Science, Computer Licentiate, Mechatronics Engineering, Communication Network Engineering, Software Engineering and Computer Engineering would be analysed.

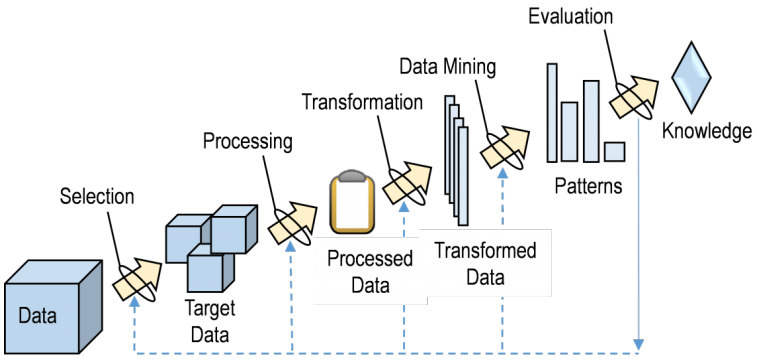


Figure 1. Steps of the proposed methodology.

The cleaning phase, also known as the preprocessing phase, includes searching for data discrepancy. Thus, examination ensured there was no lack of information, redundancy or discrepancy and noise [Navega 2002]. In the preprocessing step, few corrections of code mismatch were made considering that the data was in Portuguese. Furthermore, it was necessary to merge "Engenharia Mecatrônica" (Mechatronics Engineering) and "Engenharia de Controle e Automação" (Control and Automation Engineering), since they are considered the same course. Moreover, under the category "drop out", the following situations were aggregated: new entrance examination, the different forms of disconnection, change of course, transfer, entrance examination for another qualification, shift change and three consecutive reprimands in the same compulsory discipline, reducing the scope of variables from the course leaving form.

A relational database was built in the transformation step as shown in Figure 2. The information category was split into three tables: "Student" table has information about enrolled students, like gender, race, school attended, etc; "Subject" table has information like name and number of credits in each subject; Finally, the "Student_Subject" table has information about the student score in that subject, the average in the semester, and other attributes relating students and subject. The analysed data in this first phase are related to the "Student" table, although the "Student_Subject" table is also important and should be considered for further analysis.

In statistics, dimensionality is the amount of attributes of a dataset. In this work, all data instances are characterized by 15 attributes. To simplify the understanding of the data, numerically or visually, while maintaining the integrity, a technique for dimensionality reduction was used. That is, reduce the number of random variables under consideration by obtaining a set of principal variables [Maaten and Hinton 2008]

At the visualization phase, t-Distributed Stochastic Neighbor Embedding (t-SNE) was used because it has the advantage of giving a non-linear multidimensional visualization and representing similar data points close together. This metric uses a measure of similarity, the Euclidean distance, to " learn " discrepancies between pairs of data instances. This way, the structure and data patterns are preserved. The t-SNE Algorithm is a derivative from *Stochastic Neighbor Embedding* (SNE) which converts Euclidean distance of high dimension between similar data in conditional probability. Different from SNE, t-SNE uses distribution t-Student, instead of Gaussiana, to compute similarity [Maaten and Hinton 2008].

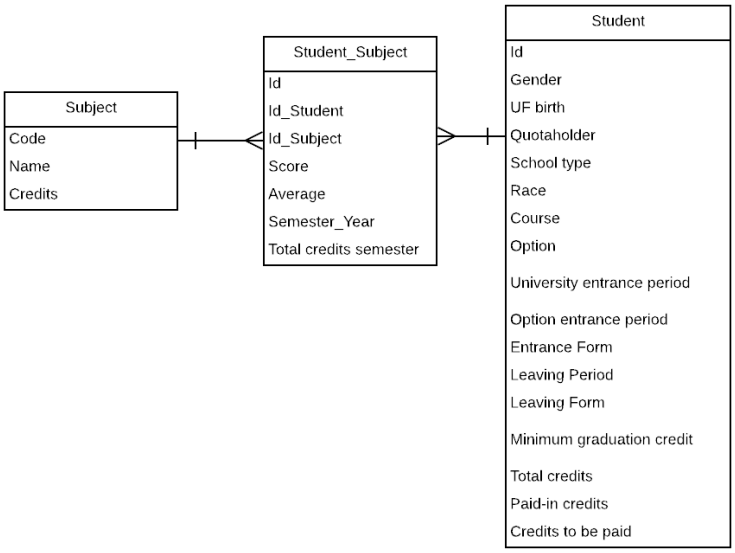


Figure 2. Database model.

4. Results

In Section 4.1 a quantitative study of data is carried out, seeking to infer knowledge from patterns. Also, in Section 4.2 a study is presented based on t-SNE using male and female databases.

4.1. Data Analysis

The dataset of students from Computer Licentiate, Computer Science, Mechatronics Engineering, Communication Network Engineering, Software Engineering and Computer Engineering has 6284 different students, where the leaving form is: “Active” students who are still on the course; “Drop Out”, students who left the course without graduating; “Graduated”, the students that graduated; “Death”, students who died during the course; and “Others”.

The database was split by gender for analysis and comparison. In the women’s database, there is a total of 789 students, being 117 from Computer Licentiate, 205 from Computer Science, 118 from Mechatronics Engineering, 196 from Communication Network Engineering, 90 from Software Engineering and just 63 from Computer Engineering. They represent a total of less than 8% of the data. The men’s dataset is composed of more than 87% of students in the courses, being 1017 from Computer Licentiate, 1517 from Computer Science, 830 from Mechatronics Engineering, 993 from Communication Network Engineering, 580 from Software Engineering and just 558 from Computer Engineering. Table 1 shows the values for each leaving form, comparing both datasets, women and men.

The first female students to enter the option in the first semester of 1991 were analysed. In this period 6 women entered the Computer Science course, 3 through the entrance exam and 3 through transfer. Curiously, of the 6 students, only 2 graduated in the first semester of 1996, the others left the option without finishing it. In the case of men, 21 students entered the Computer Science course in the first semester of 1991,

Table 1. Comparison between men and women of the leaving form.

Leaving Form	Female values	Male values
Active	307 (39%)	2101 (38%)
Graduated	203 (25%)	1218 (22%)
Drop out	279 (36%)	2167 (39%)
Death	0	7 (<1%)
Others	0	2 (<1%)

only 4 entered by transfer and the others by entrance exam. Of the 21 students, only 7 graduated, the other 2/3 left the course without completing it.

The dataset seeking a relation between the students who left UnB without completing the course with the students who entered by quota was also analysed. At UnB, the racial quotas were first applied in 2004, where 20% of vacancies was reserved for black people, and indigenous according to the demand at UnB. Also, in 2012, a federal law number 12.711/2012 was introduced requiring universities to reserve 50% of vacancies for public school students.

The first female student to enrol by quota was in the first semester of 2004. Before that period, 33 female students had graduated and 19 left without completing the course. The data for the period from 2004 until now were analyzed, noting that, from the students who left without completing the course, 202 were not quota holders and only 58 were. These 58 women represented only 22% of the all students who did not graduate. The first male student to enter by quota was in the second half of 2004, one semester after the first woman enter by quota. Previously 192 students had graduated and 153 had left the course without graduating. After 2004, 1624 non-quota students left without graduating, compared to 390 student quota holders, representing 19% of students in this category.

Since there are quotas for public schools and for black people at UnB, an analysis was made for these variables. Table 2 presents the number of students who left the course without completing it, related to the type of school (public, private or not informed) showing that only 27% of the students were from public school. Taking into account the total number of female students who left UnB without graduating, less than 7% are black and 18% are brown. In the male database, it is possible to verify that the students who left without graduating, only 26% are from public schools and 20% consider themselves brown and 7% consider themselves black.

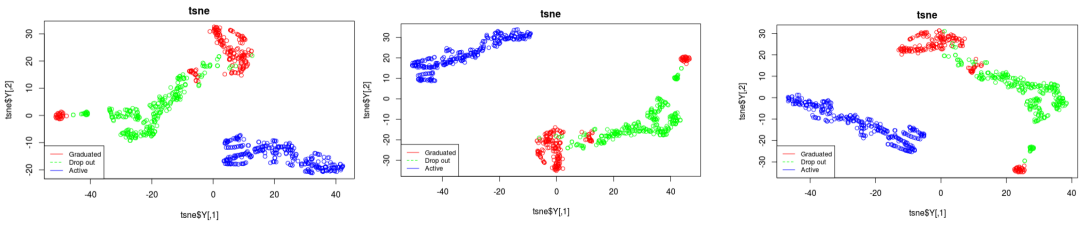
Table 2. School type comparison between men and women who left without graduating.

School Type	Dropout number of female students	Dropout number of male students
Private	134 (52%)	918 (46%)
Public	71 (27%)	531 (26%)
Not informed	55 (21%)	565 (28%)

4.2. t-SNE

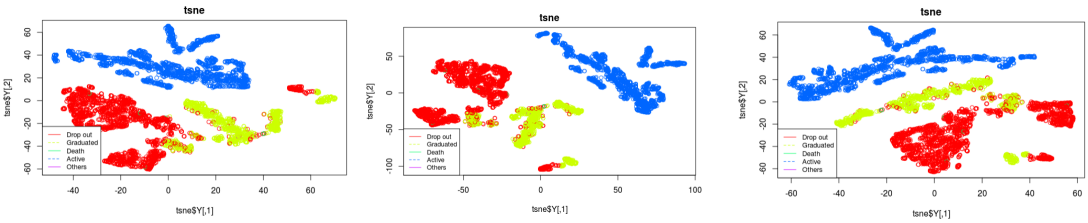
In order to generate intuitive layout using visualizations based on t-SNE, it is important to obtain appropriate parameters. The maximum number of interactions represents the total

interactions that must be performed to achieve the final result. If it is interrupted before reaching stability, the layout shows pinched shapes that are not representative. Figure 3 demonstrates the variation of the maximum number of iterations in the female students' database, seeking stability and an appropriate value. Figure 4 the effect of varying the maximum number of iterations shows for the male students' database.



(a) 5000 interactions **(b) 70000 interactions** **(c) 10000 interactions**

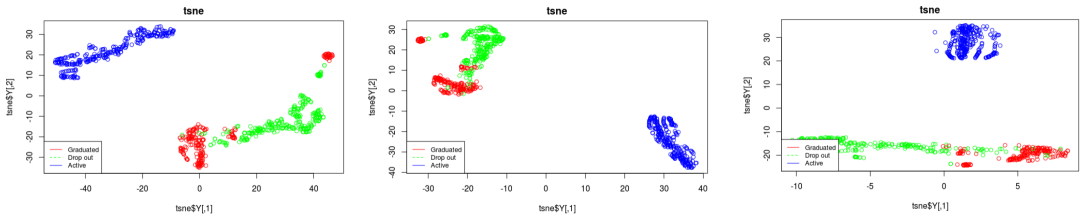
Figure 3. The difference in the interaction number in women's database.



(a) 5000 interactions **(b) 70000 interactions** **(c) 10000 interactions**

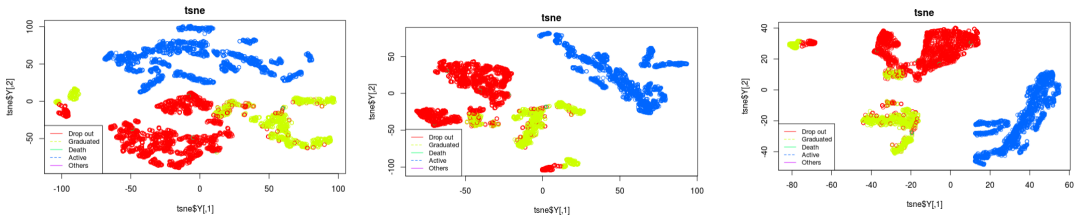
Figure 4. The difference in the interactions number in male's database.

The method introduced by van der Maaten and Hinton in 2008 has another variable parameter, the perplexity, that balances the attention between the local and global aspects of the data, it is a hunch about the number of neighbors of each point. The perplexity values were changed from 5 to 50, as suggested by van der Maaten & Hinton. Although it is important to note that the perplexity must be smaller than the number of points, as the women's database has 789 points and the male database has 5495 points, there is a safe interval. Values above those stipulated can lead to unexpected behavior. Figure 5 demonstrates an experiment in which by varying the perplexity of the women's data set is varied and Figure 6 demonstrates execution for the same perplexity values for the men's data set.



(a) perplexity 30 **(b) perplexity 50** **(c) perplexity 100**

Figure 5. The difference of the perplexity values in women's database.



(a) perplexity 30 **(b) perplexity 50** **(c) perplexity 100**

Figure 6. The difference of the perplexity in male's database.

As noted in Figure 5c, higher perplexity values than recommended generate behavior and unexpected layouts. According to the generated images, the following hyperparameters were chosen: for the women's database the perplexity will be 30 and the maximum number of iterations 70000, generating the layout of Figure 7. For the male database, for better visual results, the following hyperparameters were chosen: perplexity 50 and maximum number of iterations 70000. Figure 8 displays the generated layout.

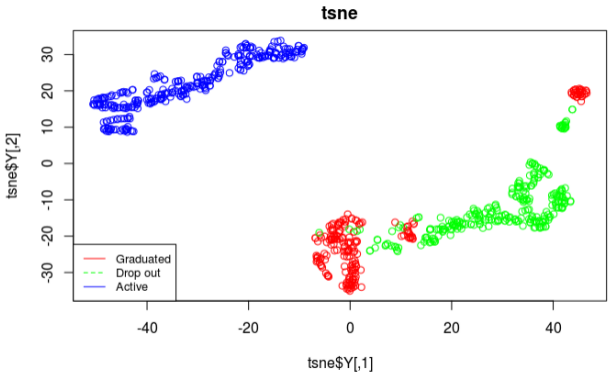


Figure 7. Layout of women's database with appropriate hyperparameters.

Analyzing the generated layout in Figure 7 it is noticeable that the active female students have different characteristics from the other two groups, those who left without finishing the course and those who graduated. It is possible to notice that there is some overlap in the last two groups, but those who graduated are more concentrated in the lower region, except for a small separated group.

Analyzing the layout, it is possible to notice that only 3 colors stand out, the ones that dropped out, the graduated and the active ones. Both male students who have died and those who have left for other reasons have values much lower than the 3 highlighted groups, as shown in Table 1. In addition, it verifies that the active students have characteristics which are distinct from the groups "Graduated" and "Drop Out", these two have overlapping points, however it is possible to cluster in different areas of predominance.

5. Conclusion

This paper presented a visual and a statistical analysis of undergraduate students in technology, in courses at UnB separated/categorized by gender. It was necessary to collect the data, clean them and apply techniques that can extract important information to

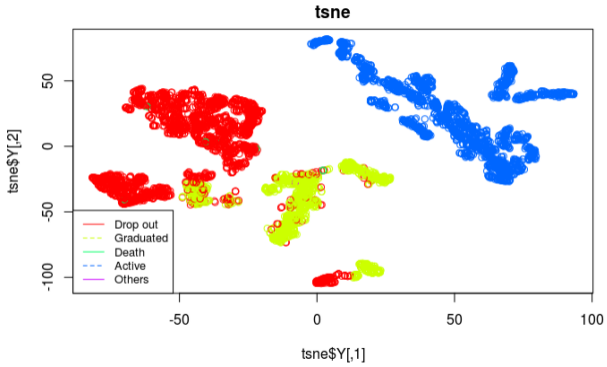


Figure 8. Layout of male's database with appropriate hyperparameters.

develop a greater understanding of the main differences between women and men on these courses and if there are any relevant differences or similarities between both genders.

The results of statistical analysis highlight that the high number of students leaving without graduating is not only a problem for the woman, though these have lower representation. There is a larger number of active students on the courses, but only 25% and 22% of female and male students graduated, respectively. Another analysis made was the relation between quota students and leaving form, it was possible to notice that there is a small number of students that fits in the quota profile that leaves the course without graduating, concluding that these informations are not related.

Using the t-SNE technique, the best parameters for male and female students were chosen, different databases require different parameters. In the end the following definitions were used: female database - 70000 for max iterations and 30 for perplexity; and male database - 70000 for max iterations and 50 for perplexity. The visualization using t-SNE emphasized the correlation between variables, and grouped the students by leaving form (active, graduated, drop out, death and others). This way it was possible to determine some relevant factors, noticing that there are certain similarities in the data of both genders, where students in the "Drop Out" and "Graduated" categories are overlapping and the "Active" students are separated from the others.

In order to continue this work, the following studies are suggested: understand the differences and similarities between male and female students' reasons, dimensionality reduction evaluation technique to ensure that the best parameters were chosen; the study of other non-linear techniques, with a process of exploratory view to extract implicit and relevant information. Further areas for investigation include the use of the data of each student related to the subjects to have a more accurate analysis, evaluating not just the profile, but also the behavior in the university, for example, the grade in each subject and average of the period; and, finally, the use of visualization as a support to data mining might predict and analyse the performance of the computer students, distinguishing between graduated and non-graduated.

References

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7(3):112–118.

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, 19(02):03.

Borges, V. R. P., Esteves, S., de Nardi Araújo, P., de Oliveira, L. C., and Holanda, M. (2018). Using principal component analysis to support students' performance prediction and data analysis. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1383.

Cheryan, S., Plaut, V. C., Handron, C., and Hudson, L. (2013). The stereotypical computer scientist: Gendered media representations as a barrier to inclusion for women. *Sex Roles*, 69(1):58–71.

Corrêa, Â. M. J. and Sferra, H. (2003). Conceitos e aplicações de data mining. *Revista de Ciência & Tecnologia*, 11:19–34.

Dominguez, M., Vilanova, R., Prada, M., Vicario, J., Barbu, M., Pereira, M. J., Podpora, M., Spagnolini, U., Alves, P., and Paganoni, A. (2018). Speet: visual data analysis of engineering students performance from academic data. *LASI 2018-Learning Analytics Summer Institutes-Universidad de Leon*.

Estivalet, L. F. (2000). O uso de ícones na visualização de informações.

Hansen, L. A., Chagas, L. M., Borges, V. R., Holanda, M., et al. (2018). Análise visual de dados educacionais: um estudo de gênero nos cursos de computação da universidade de brasília. In *12º Women in Information Technology (WIT 2018)*, volume 12. SBC.

Henn, S. (2014). When women stopped coding. *NPR Planet Money*, 21.

Holanda, M., Dantas, M., Couto, G., Correa, J. M., de Araújo, A. P. F., and Walter, M. E. T. (2017). Perfil das alunas no departamento de computação da universidade de brasília.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605.

Mercier, E. M., Barron, B., and O'Connor, K. M. (2006). Images of self and others as computer users: the role of gender and experience. *Journal of Computer Assisted Learning*, 22(5):335–348.

Navega, S. (2002). Princípios essenciais do data mining. *Anais do Infoimagem*.

Papastergiou, M. (2008). Are Computer Science and Information Technology still masculine fields? High school studentsâ perceptions and career choices. *Computers & Education*, 51(2):594 – 608.

SBC (2016). Educação superior em computação, estatísticas 2016. *Sociedade Brasileira de Computação-SBC. Disponível em:< http://www. sbc. org. br/documentos-da-sbc/summary/133-estatisticas/1074-educacaosuperior-em-Computação-estatisticas-2016>.*, 7.

Stout, J. G., Grunberg, V. A., and Ito, T. A. (2016). Gender roles and stereotypes about science careers help explain women and men's science pursuits. *Sex Roles*, 75(9):490–499.