39 Mining ENADE Data from the Ulbra Network Institution

Heloise Acco Tives Leão, Edna Dias Canedo, Marcelo Ladeira, and Fabiano Fagundes

Abstract

The National Institute of Educational Research and Studies (INEP) provides ENADE data for Higher Education Institutions (IES) from Brazil. This data is a rich source of support in improving the quality of education offered by these IES, but requires the application of data mining techniques to achieve the standards of the learning process and thus achieve improved academic performance of students in different courses. This paper aims to present the steps of mining the data provided by INEP, which will enable the identification of standards for the IES analyzed, as well as serve as a guide for other IES that wish to follow a similar process.

Keywords

Data mining · CRISP-DM · Association algorithm · Apriori

39.1 Introduction

The National Student Performance Exam (ENADE) is part of the National Higher Education Evaluation System (Sinaes) and the National Institute of Educational Studies Teixeira (INEP—http://portal.inep.gov.br/enade/) has conducted it annually since 2004.

The exam (ENADE) is carried out with a selected sample of first and last year undergraduate students from the Higher Education Institutions of Brazil in order to evaluate the quality of the higher education courses and to make a unique classification for undergraduate courses in Brazil.

Among the Enade steps are the Student Questionnaire, the Test and Courses Coordination Survey. This study will present the mining of collected data regarding the Student Questionnaire, organized and made available by INEP. The years from 2014 to 2016 were chosen for analysis and the Institutions linked to the Lutheran University of Brazil (ULBRA—http://www.ulbra.br/) chosen as the scope of the mining project, whose general objective is to identify the most relevant complaints of the students and, from these, propose alternatives of teaching process improvement.

The remainder of this paper is organized as follows. Section 39.2 presents the Literature Review. Section 39.3 presents the understanding of the business and an understanding of data. Section 39.4 presents data preparation and Modeling. Section 39.5 presents the evaluation. Section 39.6 presents an implementation. Section 39.7 presents the conclusions and future studies.

39.2 Literature Review

Data mining is a multidisciplinary research area, involving Database, Statistics and Machine Learning [1] and has been used for knowledge discovery in databases.

According to Jang et al. [2], identifying patterns in the mining patterns is one of the most important tasks in order to extract significant and useful information from raw data. This task aims to extract sets of items that represent some sort of homogeneity and/or regularity in data.

Among the techniques considered efficient for data mining, there are rules of association that seek to find links between attributes, that is, they assume that the presence of an attribute in an event implies the presence of another attribute in the same event [3].

The use of mining with association rules is appropriate to analyze educational data, which is intended to identify

H. A. T. Leão (✉) · E. D. Canedo · M. Ladeira
Computer Science Department, University of Brasília (UnB), Brasília, Federal District, Brazil

F. Fagundes
Computer Science Department, Centro Universitário Luterano (ULBRA) de Palmas, Palmas, Tocantins, Brazil

S. Latifi (ed.), *Information Technology – New Generations*, Advances in Intelligent Systems and Computing 738,
https://doi.org/10.1007/978-3-319-77028-4_39

patterns of the learning process and improve the academic performance of students in different courses [4].

Kumar and Chadha [5] identifies the application of data association rules to improve the quality of management decisions to provide quality education. This is done to analyze the data and find out factors affecting academic achievement in order to increase chances of student success.

Among the association algorithms, Apriori was proposed by Agrawal et al. and Librelotto and Mozzaquatro [6, 7], and is the most widely used to discover association rules, since it thoroughly tests the attributes in search of expanded rules that represent the pattern of its society [2].

When using the Apriori algorithm, there are measures that influence the discovery of the rules: support, which is percentage of cases in which contains both A and B; confidence, which is the percentage of cases having A, and contains B; and lift, which is the confidence rate with the percentage of cases containing B [8].

39.3 Methods

Based on comparative research of data mining techniques [2, 9, 10] and data mining from the educational field: [4, 5, 11], the Apriori algorithm was chosen for the discovery of rules of association in this study.

The methodology used comes from the literature review and the choice of technique and algorithm. It consists in knowing the environment to which the data refer and using the CRISP-DM reference model to perform the mining steps, perform tests to evaluate the results found, present results and conclusions of the project as well as the suggestions of ways to implement improvements to increase academic satisfaction.

The CRISP-DM [12] reference model defines a set of sequential steps to guide the data mining and allows for the mining process to be fast, reliable and with more management control. It also comprehends the stages of business understanding, data comprehension, data preparation, modeling, evaluation and implementation [8]. The adoption of these phases help define the flows used to run the mining project.

39.3.1 Business Understanding

The IES Ulbra network consists of seven institutions, and their students take the ENADE exam since 2005, but to date there hasn't been a study about the information given by their students in mandatory ENADE questionnaires.

The main purpose of this study is to identify the most relevant student complaints through the evaluation of the student questionnaire responses, and from that to propose new ways to improve the quality of education provided by Ulbra network. It is also expect from the implementation of this project:

- To improve understanding of the socioeconomic profile of students in order to find ways to reduce absence [13];
- identify problems in physical infrastructure and services;
- To assess the didactic and pedagogical organization of the institutions from the perception of students for further study on process improvements;
- To expand academic and professional training opportunities in order to make them gradually more appropriate to the labor market.

The data used to carry out this mining project comes from the INEP bases, that is, public databases of free access. This way, other IES can replicate the models and standards identified in this project. Criteria for success of this project will be the identification of the factors with the highest index of dissatisfaction by academics and, based on this, to propose improvements or processes that contribute to the quality of the education offered as well as the increase of the students' satisfaction with the institution and with the undergraduate course they take.

The main tool used in this mining project is R-Studio, which has packages and tools to perform most of the mining steps.

39.3.2 Data Understanding

Initial data for the mining project was extracted from the INEP website (microdata http://portal.inep.gov.br/) and refers to a micro data file of each year, the variables dictionary and the student questionnaire. The period defined for the application of this mining was from 2014 to 2016.

From detailed analysis of data collected, understanding of attributes, and integrated database generation, a total of 9858 observations and 70 attributes was reached, as shown in Table 39.1.

Continuing the task of data comprehension, we identified the use of Likert scale with six points to the attributes of the Student Questionnaire. The process of attributes cleaning and the integration of some of them must be performed in the data preparation stage, with the objective of minimizing problems in the process of knowledge discovery, such as eliminating inconsistencies and adding value to the data.

Table 39.1 Attributes identified in the raw data

Attributes	Amount
Identifiers (year and institution)	2
Socioeconomic	26
Infrastructure and physical facilities	12
Didactic-pedagogical organization	23
Academic and professional training opportunities	7
Total	70

39.4 Preparation of Data

The first activity performed in the data preparation was the cleaning of such data. As the evaluated data was extracted from questionnaires with predefined answer parameters, and were electronically answered, no outstanding values were identified.

Observations with more than 50% of blank attributes were considered missing values, and were often identified.

Table 39.2 shows the number of observations of each year which have been eliminated to avoid problems in the quality of analysis to be performed.

Table 39.2 Quantitative observations with at least 50% of missing data

	2014	2015	2016
Total gross	4165	3860	1833
Missing data	886	807	177
Frequency of absence	21.3%	20.9%	9.7%
Total refined	3279	3053	1656

The process of evaluation of attributes was manual, as it was required to have an understanding of variables by consulting the data dictionary and checking the scale/pattern used in each one.

As all data was obtained from a single data source, just one transformation process was necessary to change the data value of 7 and 8 for “null” values, so as to not influence the process of merging of attributes. 6689 entries with value “7—I cannot answer” and 5818 observations with value “7—Not applicable” were identified, corresponding respectively to 2.07%, and 1.80% of 322.689 data entries on the Student Questionnaire.

With the support of a specialist, who is accompanying all phases of this project, it was possible to identify and exclude 16 attributes that represent socioeconomic variables from the project because they are not significant for the intended results. Example of excluded variables: q3, which deals with the nationality of the students, and q16, which stores State data on secondary education completion.

To identify correlation between the attributes of the Student Questionnaire, the Pearson correlation algorithm was applied in software R, which identified 5 attributes with correlation above 0.7 and could be deleted.

For analysis of other attributes of the Student Questionnaire, the histogram and the average of the other attributes were generated to help the individual analysis of each attribute. Thus, another 27 attributes were deleted, because all of them possess an average satisfaction of 5.18, and the maximum satisfaction score is 6.

The data set resulting from this preparation phase consists of 22 attributes, 7988 observations and 177,795 data entries. They are:

- Two attributes (year, IES) that will be used to make the separation of data and subsequent comparison of the results of each institution.
- Ten qualitative attributes that represent the socioeconomic data chosen by the expert to set the context for the inclusion of students.
- Ten remaining attributes of the Student Questionnaire which were filtered through the correlation and analysis of higher incidence of complaints by students.

From the study of data, cleaning, classification and analysis of attributes performed at this stage, it is possible to define models, select techniques and parameters to perform the mining process. The Modeling step of this project performs and presents these steps.

39.4.1 Modeling

In order to perform the attributes modeling of this project, variables association mechanisms will be used, together with the application of the Apriori algorithm.

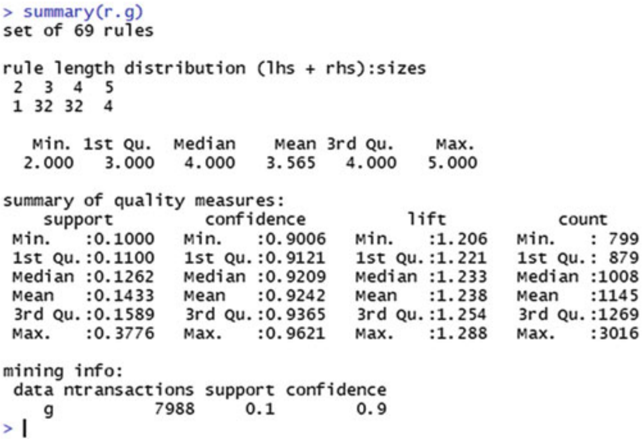
As such algorithm is based on nominal or binary attributes, the values of the Student Questionnaire attributes had to be adapted. In accordance to the expert, it was defined that for 5 s and 6 s on the Likert scale, attributes would receive the “no” value, indicating no improvement pointed by students. For Likert scale values from 1 to 4, attributes would receive the “yes” value, indicating a need for analysis and evaluation of opportunities for improvement.

The data processing phase resulted in ten attributes of the Student Questionnaire. These were separated in five attributes for infrastructure and physical facilities, three attributes for the didactic and pedagogical organization, and two attributes for the academic and professional training opportunities. All these attributes have been adapted to the defined pattern.

The Apriori algorithm was applied and generated 69 association rules. Figure 39.1 presents the summary of statistics (Minimum, 1st Quartile, Median, Mean, value as average, Q1, Q3° Quartile and Maximum Value) to support parameters, confidence, lift and counter resulting from the application of Apriori algorithm.

The results generated by Apriori were filtered to identify just the rules that brought associations with the attributes derived from the Student Questionnaire. This filter generated a subset of 30 rules presented in Fig. 39.2, which will compose the result of the mining process.

With the definition of rulesets for mining, the modeling phase of the project is complete and it must move on to the evaluation of the generated rules, which will be held in Sect. 39.5.



```
> summary(r.g)
set of 69 rules

rule length distribution (lhs + rhs):sizes
 2  3  4  5
 1 32 32  4

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   4.000   3.565   4.000   5.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.1000   Min.   :0.9006   Min.   :1.206   Min.   : 799
 1st Qu.:0.1100   1st Qu.:0.9121   1st Qu.:1.221   1st Qu.: 879
 Median :0.1262   Median :0.9209   Median :1.233   Median :1008
 Mean   :0.1433   Mean   :0.9242   Mean   :1.238   Mean   :1145
 3rd Qu.:0.1589   3rd Qu.:0.9365   3rd Qu.:1.254   3rd Qu.:1269
 Max.   :0.3776   Max.   :0.9621   Max.   :1.288   Max.   :3016

mining info:
 data ntransactions support confidence
    g          7988     0.1        0.9
> |
```

Fig. 39.1 Statistics generated by applying the rules Apriori algorithm

39.5 Evaluation

The evaluation step is responsible for analyzing the model (or models) obtained in more detail, by reviewing the work done until this point to enable the validation of the final model.

In Fig. 39.3 it's possible to see the array of previous rules (LHS) with the successor attributes (RHS). The color intensity of the circles represents the scale of the **lift** parameter and the size of the circles indicate information on the support parameter.

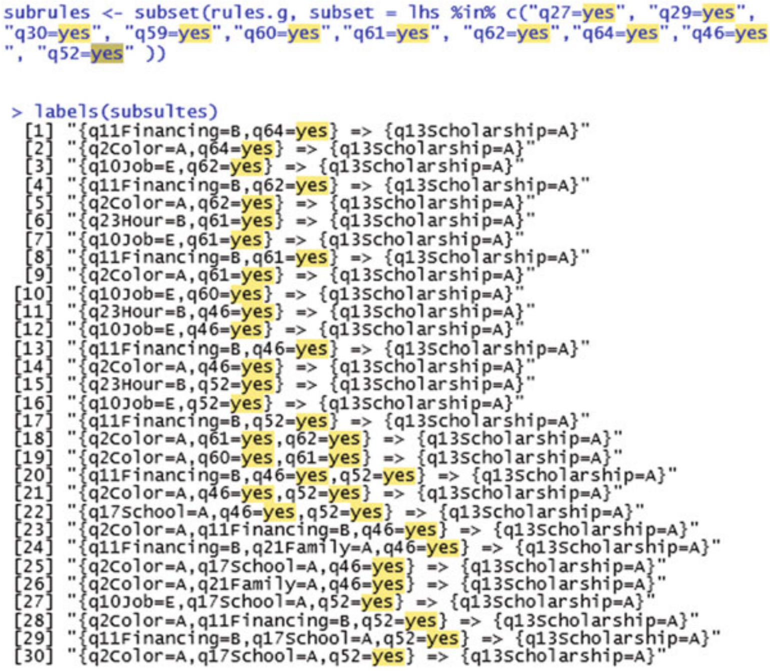
Although most of the rules generated by Apriori point to the same successor attribute (q13 = A, indicating that students do not have a scholarship to finance their studies), the generated rules do not become invalid as this group of students represents the highest percentage (74.8%) of the students who answered the Student Questionnaire.

Figure 39.4 presents the list of 30 rules that are in analysis through color that indicate the **lift**, which has a minimum value of 1.208 and a maximum value of 1.278, and the size of the circles represent the support parameter, which has a minimum value of 0.1 and a maximum value of 0.174.

In the center of the graph lies the successor of the rules generated by Apriori. Around this attribute are the attributes that represent additional portions of the population (q17School = A, q21Family = A, q11FinanCourse = B, q2Color = A, q10Work = E, q23Hours = B).

In regards to other attributes, they are inferred to represent the greatest complaint indexes (Student Questionnaire responses with values from 1 to 4 on the Likert scale) where:

Fig. 39.2 Statement subset of mining rules



```
subrules <- subset(rules.g, subset = lhs %in% c("q27=yes", "q29=yes",
"q30=yes", "q59=yes","q60=yes","q61=yes", "q62=yes","q64=yes","q46=yes
", "q52=yes" ))

> labels(subsultes)
 [1] "{q11Financing=B,q64=yes} => {q13Scholarship=A}"
 [2] "{q2Color=A,q64=yes} => {q13Scholarship=A}"
 [3] "{q10Job=E,q62=yes} => {q13Scholarship=A}"
 [4] "{q11Financing=B,q62=yes} => {q13Scholarship=A}"
 [5] "{q2Color=A,q62=yes} => {q13Scholarship=A}"
 [6] "{q23Hour=B,q61=yes} => {q13Scholarship=A}"
 [7] "{q10Job=E,q61=yes} => {q13Scholarship=A}"
 [8] "{q11Financing=B,q61=yes} => {q13Scholarship=A}"
 [9] "{q2Color=A,q61=yes} => {q13Scholarship=A}"
[10] "{q10Job=E,q60=yes} => {q13Scholarship=A}"
[11] "{q23Hour=B,q46=yes} => {q13Scholarship=A}"
[12] "{q10Job=E,q46=yes} => {q13Scholarship=A}"
[13] "{q11Financing=B,q46=yes} => {q13Scholarship=A}"
[14] "{q2Color=A,q46=yes} => {q13Scholarship=A}"
[15] "{q23Hour=B,q52=yes} => {q13Scholarship=A}"
[16] "{q10Job=E,q52=yes} => {q13Scholarship=A}"
[17] "{q11Financing=B,q52=yes} => {q13Scholarship=A}"
[18] "{q2Color=A,q61=yes,q62=yes} => {q13Scholarship=A}"
[19] "{q2Color=A,q60=yes,q61=yes} => {q13Scholarship=A}"
[20] "{q11Financing=B,q46=yes,q52=yes} => {q13Scholarship=A}"
[21] "{q2Color=A,q46=yes,q52=yes} => {q13Scholarship=A}"
[22] "{q17School=A,q46=yes,q52=yes} => {q13Scholarship=A}"
[23] "{q2Color=A,q11Financing=B,q46=yes} => {q13Scholarship=A}"
[24] "{q11Financing=B,q21Family=A,q46=yes} => {q13Scholarship=A}"
[25] "{q2Color=A,q17School=A,q46=yes} => {q13Scholarship=A}"
[26] "{q2Color=A,q21Family=A,q46=yes} => {q13Scholarship=A}"
[27] "{q10Job=E,q17School=A,q52=yes} => {q13Scholarship=A}"
[28] "{q2Color=A,q11Financing=B,q52=yes} => {q13Scholarship=A}"
[29] "{q11Financing=B,q17School=A,q52=yes} => {q13Scholarship=A}"
[30] "{q2Color=A,q17School=A,q52=yes} => {q13Scholarship=A}"
```

Fig. 39.3 Relationship matrix leading and succeeding rules

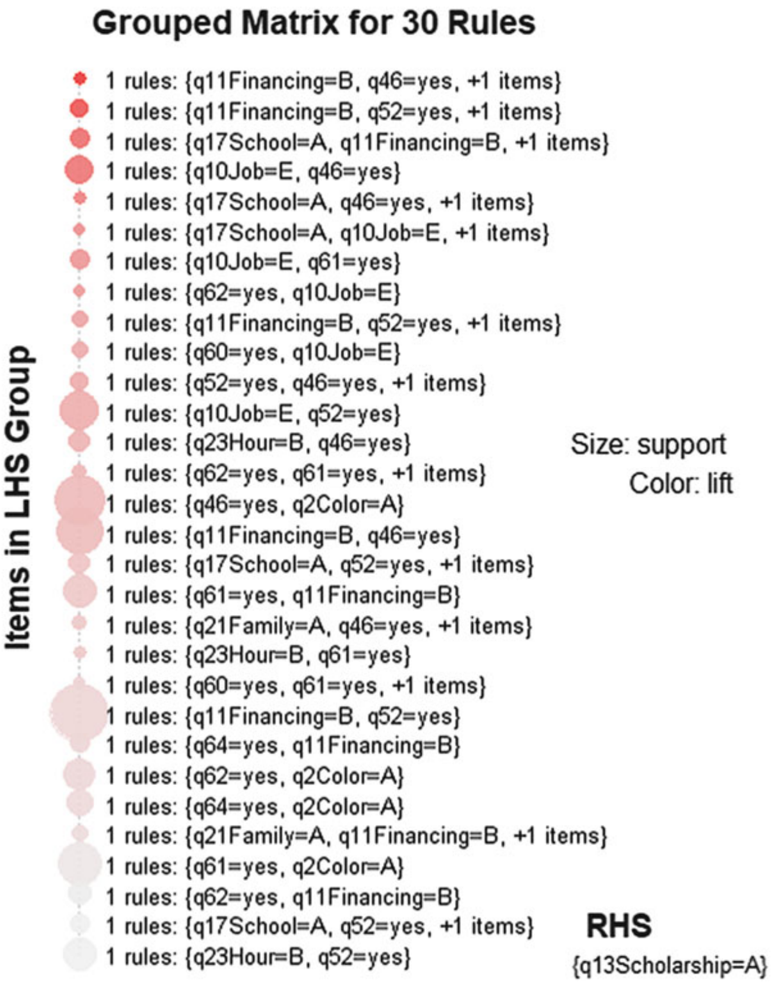
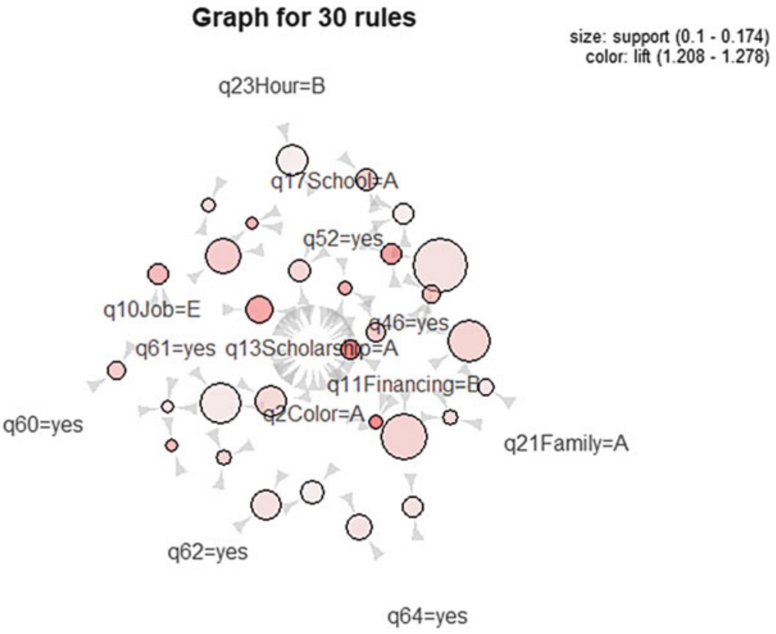


Fig. 39.4 Graph interface 30 of the rules in question



- Academic and professional training opportunities, in questions: Q46—The institution offered opportunities for students to act as representatives in collegiate organs and Q52—The students were given opportunities to undertake exchanges and/or internships in the country.
- Infrastructure and physical facilities, in questions: q60—The course provided monitors or tutors to help students; q61—The infrastructure conditions of the classrooms were adequate; q62—The equipment and materials available for class practices were adequate for the number of students and q64—The environments and equipment intended for the practical classes were adequate to the course.

The next step in the mining project is the implementation of the model, where it is expected to consolidate the knowledge discovered with the created model and to verify the fulfillment of the project objectives.

39.6 Implementation

With the confidence factor of 0.9 executed in the modeling, whose results were interpreted in the evaluation phase, it was verified that an important set of attributes was not covered.

To get the results expected for this project, we seek to evaluate the full set of attributes of the Student Questionnaire; tests were carried out until the complete identified set of attributes was represented in the rules generated by Apriori. The configuration made to meet this premise was with the use of the confidence parameter set to 0.6.

The result of new implementation of Apriori algorithm revealed the existence of 753 rules, and the amount of rules related to the attributes of the Student Questionnaire are:

- Academic and professional opportunities
  - Q46: 25 rules;
  - Q52: 50 rules;
- Physical infrastructure and facilities
  - Q60: 33 rules;
  - Q61: 31 rules;
  - Q62: 26 rules;
  - Q64: 6 rules;
- Didactic and pedagogical Organization
  - Q27: 8 rules;
  - Q29: 7 rules;
  - Q30: 5 rules;
  - Q59: 7 rules.

This quantify of generated rules can be seen Fig. 39.5, where it can be seen (by marking made with blue lines and the very thickness of the red lines) that the attributes with a higher incidence in rules are Q46 = yes, q52 = yes, q60 = yes, q61 = yes and q62 = yes. Based on this predominance, this set of attributes was determined as factors that can be better developed in the IES, and thus increase the quality of teaching offered and, consequently, the students' satisfaction.

The Analysis of two factors related to opportunities for academic and vocational training enabled the identification of a profile of students who were previously unknown to analyzed IES, regarding dissatisfaction with items of this nature. The identification of this profile is a starting point for defining improvements in the IES.

These are mostly white students (q2Cor = A), who do not use student financing mechanisms to pay the tuition fees (q11FinanCurso = B), who do not have scholarships to support their study (q13Bolsa = A), who went through high school in a public school (q17Escola = A) and have at least one family member who has completed higher education (q21Familia = yes).

The analysis of the items that refer to the Physical Infrastructure of IES resulted in the same pattern of students who present most of the complaints regarding the Institution. In addition, it was possible to highlight relevant factors for the identification of improvements in the institutions of the Ulbra network, and which are simple to implement. One example

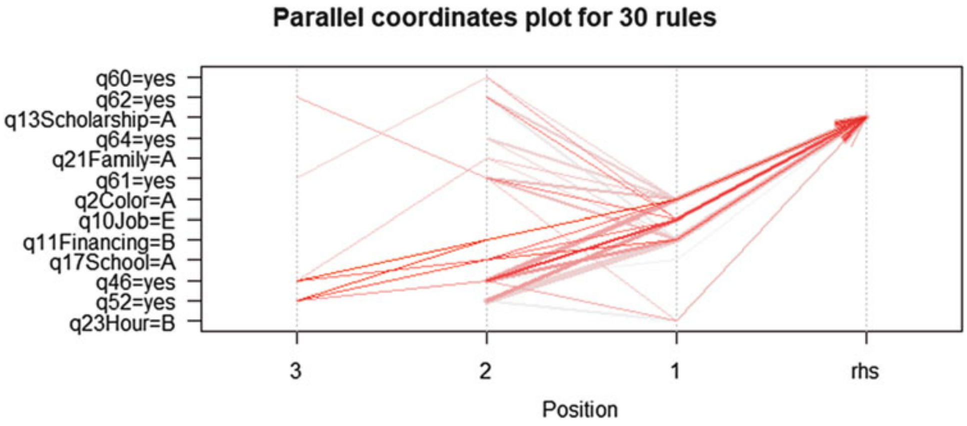


Fig. 39.5 Attribute frequency in Apriori algorithm rules with reliable 0.6

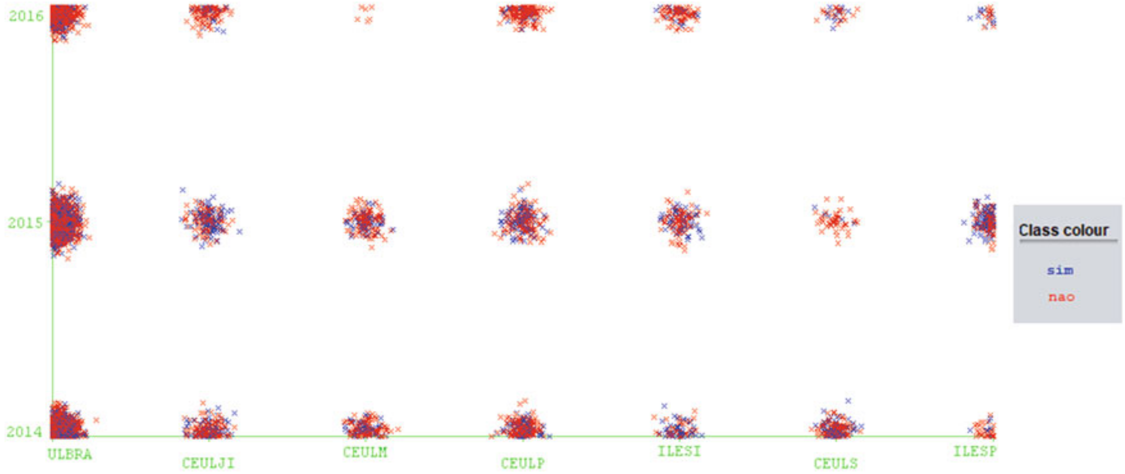


Fig. 39.6 Satisfaction of the representation of the students

of these factors is the increasing the availability of monitors or tutors for the students.

The mining assessment for these students also warranted the generation of Fig. 39.6, where the blue color showcases the amount of students who showed some dissatisfaction in regard to the items listed in the questionnaire. The graph shows the institutions of the Ulbra network in the "X axis" and the year of the evaluation in the "Y axis".

It can be seen from the evaluation of the image that, although the satisfaction index with the evaluation questions of IES is mostly positive (red x), there is a frequent occurrence of students' dissatisfaction (blue x).

This is proof that the data coming from the Student Questionnaire is relevant for proposing improvements in IES and thus trying to increase the level of academic satisfaction.

39.7 Conclusions and Future Work

With the execution of the steps of the CRISP-DM methodology with data provided by INEP to carry out this mining project, it was possible to display a large set of rules regarding the Student Questionnaire answered by students to take the Enade.

There was a need for cleaning, transforming, and adjusting the data to fit the Apriori algorithm chosen to perform the association task attributes.

The rules generated, together with the graphs executed in the software R, were easy to understand and of great value for the interpretation of the results, which led to the best understanding and consequent attention to the main factors of students' dissatisfaction.

As a highlight of mining there is the identification that the factors that evaluate questions related to didactic-pedagogical, practices, despite being related by some students, do not represent the greatest factor of dissatisfaction.

The attributes that evaluate questions concerning the academic and professional training opportunities are the ones with the most dissatisfaction among students and require a specific study for bringing improvements.

For a better understanding of the questions related to Physical Infrastructure, it's recommended to apply specific questionnaires to survey the needs in each IES and thus set up an activity plan in order to meet the needs of the students. It is worth mentioning that some of these factors can have efficient solutions and simple implementation such as the increase in the number of tutors and monitors for the disciplines.

The conclusion is that this mining project has been of great value for analyzing the data mass that existed and saw no purpose. The rules generated by the application of the Apriori algorithm were satisfactory to the achievement of the project objectives.

As a continuation of this study, we expected the IES to use these results to implement factors that contribute to students' satisfaction and study environment. As a future research, the mining of the data provided by INEP every year is expected, in accordance to the rules modeled here.

It is important to inform that the data used to carry out this study is from public free access, making it so that the standards defined in this work can be replicated by other universities in Brazil.

References

1. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD process for extracting useful knowledge from volumes of data. Commun. ACM **39**(11), 27–34 (1996). https://doi.org/10.1145/240455.240464
2. S. Jang, K. Park, Y. Kim, H. Cho, T. Yoon, Comparison of h5n1, h5n8, and h3n2 using decision tree and apriori algorithm. J. Biosci. Med. **3**(06), 49 (2015)
3. L.A. da Silva, S.M. Peres, C. Boscarioli, *Introdução à mineração de dados: com aplicações em R* (Elsevier Brasil, Rio de Janeiro, 2017)
4. G. Mobasher, A. Shawish, O. Ibrahim, Educational data mining rule based recommender systems, in *CSEDU 2017—Proceedings of the 9th International Conference on Computer Supported Education*, vol. 1 (Porto, Portugal, April 21–23, 2017), pp. 292–299. [Online]. https://doi.org/10.5220/0006290902920299
5. V. Kumar, A. Chadha, Mining association rules in students assessment data. Int. J. Comput. Sci. Issues **9**(5), 211–216 (2012)
6. R. Agrawal, T. Imieliński, and A. Swami, Mining association rules between sets of items in large databases, in *93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, vol. 22, no. 2 (ACM, 1993), pp. 207–216
7. S.R. Librelotto, P.M. Mozzaquatro, Análise dos algoritmos de mineração j48 e apriori aplicados na detecção de indicadores da qualidade de vida e saúde, Revista Interdisciplinar de Ensino, Pesquisa e Extensão, vol. 1, no. 1 (2014)
8. P. Kalgotra, R. Sharda, Progression analysis of signals: Extending CRISP-DM to stream analytics, in *2016 IEEE International Conference on Big Data, BigData 2016* (Washington, DC, USA, December 5–8, 2016), pp. 2880–2885 [Online]. https://doi.org/10.1109/BigData.2016.7840937
9. J.M. Luna, F. Padillo, M. Pechenizkiy, S. Ventura, Apriori versions based on mapreduce for mining frequent patterns on big data. IEEE Trans. Cybern. (2017)
10. C. Woo Kim, S.H. Ahn, T. Yoon, Comparison of flavivirus using datamining-apriori, k-means, and decision tree algorithm, in *2017 19th International Conference on Advanced Communication Technology (ICACT)* (IEEE, 2017), pp. 454–457
11. S. Ougiaroglou, G. Paschalis, Association rules mining from the educational data of ESOG web-based application, in *Artificial Intelligence Applications and Innovations—AIAI 2012 International Workshops: AIAB, AIeIA, CISE, COPA, IIVC, ISQL, MHDW, and WADTMB, Halkidiki, Greece, September 27–30, 2012, Proceedings, Part II* (2012), pp. 105–114
12. R. Wirth, J. Hipp, Crisp-dm: towards a standard process model for data mining, in *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (2000), pp. 29–39
13. R.M. Hoed, Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação, Dissertação, Universidade de Brasília, 2016. [Online], http://repositorio.unb.br/handle