Fingerprints in a Computer Science Course Profile: Early results from the IFNMG Data Warehouse

Igor Eleuterio[1], Marcos Bedo[2], Daniel de Oliveira[3], Luiz Olmes[4], Lucio Santos[1]
[1]IFNMG: Federal Institute of North of Minas Gerais, Montes Claros/MG, Brazil
lucio.santos@ifnmg.edu.br,igoralberte@gmail.com
[2]INFES/UFF: Fluminense Northwest Institute, Fluminense Federal University, S. A. Pádua/RJ, Brazil
marcosbedo@id.uff.br
[3]INFES/UFF: Institute of Computing, Fluminense Federal University, Niterói/RJ, Brazil
danielcmo@ic.uff.br
[4]UNIFEI: Inst. of Math and Computing, Federal University of Itajuba, Itajubá/MG, Brazil
olmes@unifei.edu.br

ABSTRACT

This manuscript reports the implementation of an Educational Data Warehouse (EDW) at the Federal Institute of North of Minas Gerais by using data from the academic system called *Cajuí*. The logical model of the system is the Fact Constellation with data persisted into the relational DBMS PostgreSQL. After the loading and setting of the EDW, we ran a set of analytic queries regarding courses from the Computer Science bachelor course at the IFNMG campus of Montes Claros for the 2013/2020 timespan. The data analysis indicates: *(i)* there were no significant differences in the academic performances of students enrolled by either standard entrance or Brazilian SISU exams, *(ii)* the number of unofficial dropouts reached up to 19% of students, *(iii)* the 19.51% of students that took a leave of absence and 15.38% of dropout bachelor candidates had completed at least ⅓ of courses from the entire graduation process, *(iv)* the first-year courses had more failing students than final-year courses, and the average grade of final-year courses was higher than those of other years, and *(v)* nearly 60% of students had at least one failure in either Algorithms and Data Structures or Calculus courses.

CCS CONCEPTS

• **Social and professional topics** → Computing education.

KEYWORDS

Information Systems for Education, Educational Data Warehouse, Data Analytics, Dropout, Computer Science

1 INTRODUCTION

Although information systems based on transactional databases increase productivity and consistency of data-driven environments, they may also generate an over-accumulation of stored data. In those scenarios, it is often impractical to analyze massive stored data without proper tools for the extraction of useful information. Data Warehouse (DW) systems are prime solutions for converting such a large mass of heterogeneous data from transactional sources into useful information that can be easily accessed by online analytical processing (OLAP) systems. The combination of DW and OLAP enables answering diversified and high-level data-oriented queries posed by decision-makers [7].

The rationale behind the implementation of DWs for enterprise systems is the broader understanding of the business data so that the decision-makers can react to adversities [5, 10] and prevent personal and financial losses [9]. Accordingly, decisions have increasingly been taken based on data (or information extracted from data) stored in DW environments [8, 14], which avoids the subjectivity of decisions made based on administrators' prior knowledge or intuition. Additionally, companies have also turned to DW and OLAP tools for addressing risk management [1, 23]. Recently, other business sectors than e-commerce and health [17] have started using such analytical tools for making data-oriented decisions, as the Education sector [12, 13]. In this case, the goal is to characterize student profiles as well as outline guidelines for optimizing both the educational institution's quality and costs. Several studies have also proposed DW designs for educational companies to provide insightful information, such as dropout trends [15, 19] and academic performance forecasting under certain conditions [2, 11].

The reviewed designs differ from each other, as they implement solutions for integrating information across multiple information sources (*e.g.*, relational databases and spreadsheets) into a central and queryable repository. Characteristics of institutions and regional regulations that vary from country to country also must be considered by the EDW Project [16], which renders the EDW reproducibility a burdensome challenge, sometimes unfeasible.

In this manuscript, we propose (with reproducible details) a new DW design for the Federal Institute of North of Minas Gerais, a Brazilian educational institution located in a vulnerable South American region. While the Federal Institute of North of Minas Gerais extensively uses an academic system, coined *Cajuí*[1], for storing data from eleven *campi* with graduation courses since 2007, such information was neither stored nor queried in a proper DW-oriented approach until 2019. This study discusses the implementation of the IFNMG Educational Data Warehouse (EDW), which we designed for the analysis of student profiles regarding their grades, dropouts, leaves of absence, and failures. Additionally, this study

[1]Academic system available at cajui.ifnmg.edu.br

EduComp'22, Abril 24-29, 2022, Feira de Santana, Bahia, Brasil (On-line)

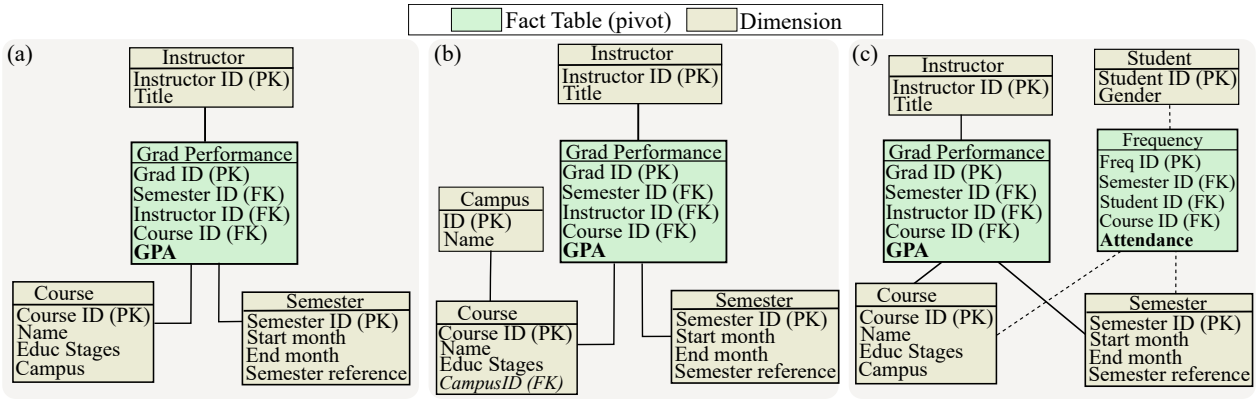

Figure 1: Dimensional design examples for a Data Warehouse. (a) Star Schema. (b) Snowflake. (c) Fact Constellation.

reports a preliminary data analysis we carried out with the `IFNMG` EDW for Computer Science students in the 2013/2020 timespan.

The remainder of this manuscript is organized as follows. Section 2 presents the preliminary concepts on data warehousing and its deployment in educational institutions, Section 3 describes the project and implementation of the `IFNMG` EDW, and Section 4 reports our preliminary analysis. Finally, Section 5 concludes the study and discusses future work.

2 PRELIMINARIES

2.1 Data Warehousing

Data warehouses are repositories for large and stationary data gathered from different information sources through proper Extract, Transform, and Load (ETL) processes [10, 24]. By definition [8], data warehouses must handle data in a time-variant, non-volatile, and subject-oriented manner, which creates a core of data mass to be queried and grouped by distinct levels of granularity. Such data handling enables the analysis of clean and consolidated information, which has provided useful insights and data-oriented knowledge for several business domains [10, 13, 23].

DW systems distinguish themselves from transactional databases [3] and Data Lakes [18] by their non-volatile and subject-oriented properties. While transactional databases are constructed to ensure ACID properties by using transaction-oriented removal and update operators [3], DWs are designed to optimize only insertion and retrieval, which may disregard the ACID aspect of updates [10]. DWs are also built for subject-oriented query loads following a logical model-compliant schema, while data lakes are designed for the storage of generic raw data and information in schema-less structures that are hard to query regarding specific-purpose analytics [18]. Finally, unlike competing approaches, DWs heavily rely on *dimensional modeling* [10] for keeping their logical structure portable and intelligible, as well as to optimize query performance [8].

Essentially, the dimensional model uses *fact tables* or *dimensions* for storing data according to their timestamps and granularity. The facts include values (attributes) related to the DW subject of interest, and the dimensions can be seen as the context where the fact values are examined. In practice, a dimensional model is constructed following three design patterns, namely *(i)* Star Schema, *(ii)* Snowflake, and *(iii)* Fact Constellation [7]. Star Schema (the most common DW design [7]) relies on a pivotal, central table that drawn (merging and grouping) values and measures from the remaining dimension tables (Figure 1(a)). The drawback of Star Schemes is that the pivot table concentrates a great number of joining attributes, which may result in growing data redundancies. The Snowflake scheme addresses this issue by splitting the dimension tables into several normalized tables with "one-to-many" relationships between normalized attributes. For instance, table `Course` was normalized regarding the *campus* attribute in the example of Figure 1(b). A side-effect is that extra join operations are required to process the queries, which makes the retrieval operators costlier than those in Star Schema.

Therefore, Star Schema and Snowflake are competing approaches to be chosen according to the availability of secondary memory and computational processing power [10]. Finally, the Fact Constellation design (see Figure 1(c)) enables sharing a set of dimension tables and connects distinct models of Star Schemas (`Frequency` and `Grad Performance`). Such an approach assumes the memory hierarchy is set to manage attribute redundancy and enables the DW to handle more than one subject-oriented analytics context. The scope of those analytics is defined by the set of logically connected pivotal fact tables, which ensures query performance.

Dimensional modeling is tightly coupled with the DW project methodology, which can be divided into *Top-Down* and *Bottom-Up* approaches [6, 8, 10]. The *Top-Down* strategy provides a whole consistent dimensional view regarding each real-world organizational piece and the stored data, *i.e.*, a subset of the DW data that is directly linked to a local portion of the modeled organization, *e.g.*, a department, or a faculty, in the case of a real-world Education institution. On the other hand, the *Bottom-Up* approach prioritizes individual analytics by assuming independent data portions rather than the whole dimensional view. Accordingly, for each piece of the real-world organization, one independent repository is set by creating a copy of the transactional data, structured logically according to the local part of the dimensional model. Those local-independent data portions can eventually be integrated into a single repository, representing the whole modeled organization. In practice, *Top-Down* and *Bottom-Up* are competing approaches regarding budget constraints.

The deployment of a DW also depends on mapping the dimensional model into a physical implementation. Usually, three possible approaches are employed for such a mapping [7], namely *(i) Relational Online Analytical Processing* (ROLAP), *(ii) Multidimensional Online Analytical Processing* (MOLAP), and *(ii) Hybrid Online Analytical Processing* (HOLAP). Although ROLAP approaches rely on relational DBMSs for data storage and maintenance, they add an extra Online Analytical Processing (OLAP) middleware on top of the DBMS engine to support data analysis. Accordingly, ROLAP overall performance is usually bounded by the coupled relational DBMS. In contrast, MOLAP enhances data handling by adopting multidimensional views implemented by vector structures and improves data retrieval with fast indexing of precomputed and summarized data. HOLAP implementations combine the support for large data storage derived from ROLAP with the performance of MOLAP approaches.

Following the DW project pipeline, the data ingestion on the DW physical implementation includes the ETL design for *(i)* the collection, cleaning, and integration of data from multiple sources (*e.g.*, transactions, tables, or even raw data from sheet files), and *(ii)* the schema-oriented persistence of those conformed data into a proper DBMS (*e.g.*, relational or columnar). The schema must comply with the time-variant property of DWs by including timestamp attributes for the retrieval of consolidated data in distinct timespans. For instance, the grades of a regular student can be consolidated by intervals of months, semesters, or years. DWs enable the querying of multidimensional data by distinct levels of granularity after data ingestion (through dynamic and interactive filters and reports) [7, 13, 25].

The OLAP operators that support the reports include *(i) drill down/drill up*, which retrieves data by the granularity specified in the dimensional attribute, *(ii) slice*, which provides details for a specific dimension (fact table), and *(iii) pivot*, which fetches pairs of related two-dimensional data [10].

2.2 Educational Data Warehouses

Educational Data Warehouses (EDWs) are DWs designed for the analysis of data from teaching and learning within educational and academic information systems [13, 15, 20]. The EDW goals range from data-oriented forecasting of academic performances [2, 19] to dropout trends [15, 20, 25] regarding diversified and national-wide groups of schools and universities [15, 20, 21]. For instance, the studies in [13] and [25] review the methodologies for constructing an EDW whose focus is on either *(i)* describing possible users and data to be managed by a DW system [13] or *(ii)* discussing the business decision process to be carried after the DW querying [25]. While those proposals set the global stage for DW usage over educational data, they lack the implementation details that must be adjusted according to regional laws and rules. In this manuscript, we tackle this issue by investigating the construction scenario of an EDW for a Brazilian educational institution.

Recently, the study in [4] has proposed a DW project and implementation for the analysis of remote educational data from Brazilian government repositories. The authors' proposal incorporated data from e-MEC and *Universidade Aberta do Brasil* (UAB) into a ROLAP DW (constructed with a *Bottom-Up* approach and the Fact Constellation model) for drilling several aspects of e-learning graduation courses. Their results indicate more Brazilian students tend to adopt e-learning, particularly those in peripheral areas and country parts. However, their proposal is designed for national-wide analysis and carried out without data from in-person courses, hindering the decision-making process for particular cases.

The proposal in [15] follows a different DW structure for the evaluation of dropouts in graduate courses at the *Paraná Federal University of Technology* – Brazil. The DW was deployed by using materialized views on DBMS Oracle, *i.e.,* a *virtual DW*, and the Fact Constellation was used as the logical model. Copies from transactional data were not necessary for their implementation due to the use of materialized views. The authors also embedded an index – coined "course hardness" – into the reports for measuring in scale the quality of a particular course based on the students' approval. Unfortunately, their approach is tightly coupled to a particular commercial DBMS so that portability becomes a challenging issue for project replication and migration.

Finally, the approach in [20, 21] follows an analogous ROLAP strategy for designing a DW focused on predicting dropout trending patterns, which considers correlations among a fixed number of attributes. The results pinpoint the initial scores on the admission exam correlate with both students' performance and dropouts. However, other attributes went unexamined as possible causes.

Inspired by those previous designs and data-driven results, in this paper, we propose and implement an Educational Data Warehouse for `IFNMG`. Our information system provides analytics (with reports and dashboards) for distinct subject-oriented queries and levels of granularity, which we show as a case study for the Computer Science course.

3 THE `IFNMG` EDUCATIONAL DATA WAREHOUSE

3.1 Dimensional modeling

Although the design of an EDW has been explored in the literature, it plays a crucial and support role in the analytical process of querying educational information. Academic systems can vary according to regulations and institutions, so whenever data is poorly imported into the DW, the analytical query is harmed as well, and the decision support application will be plight down. Figure 2 illustrates the conceptual methodology we used for building the `IFNMG` Educational Data Warehouse.

Step (1) is a twofold process whose initial stage depends on the real-world meaning of "success" of students' activities. Meetings were held with decision-makers, *i.e.*, *pro-rectors*, to define the information requirements and which indicators should be measured or taken for that purpose. The final stage of the twofold step is a data assessment from the *Cajuí* academic information system regarding *(i)* the availability of information throughout 12 years of storage, *(ii)* the levels of data granularity and their available format. The *Cajuí* academic database is designed in three different schemas (basic, undergraduate, and technician), having a total of 93 tables stored in the relational DBMS PostgreSQL.

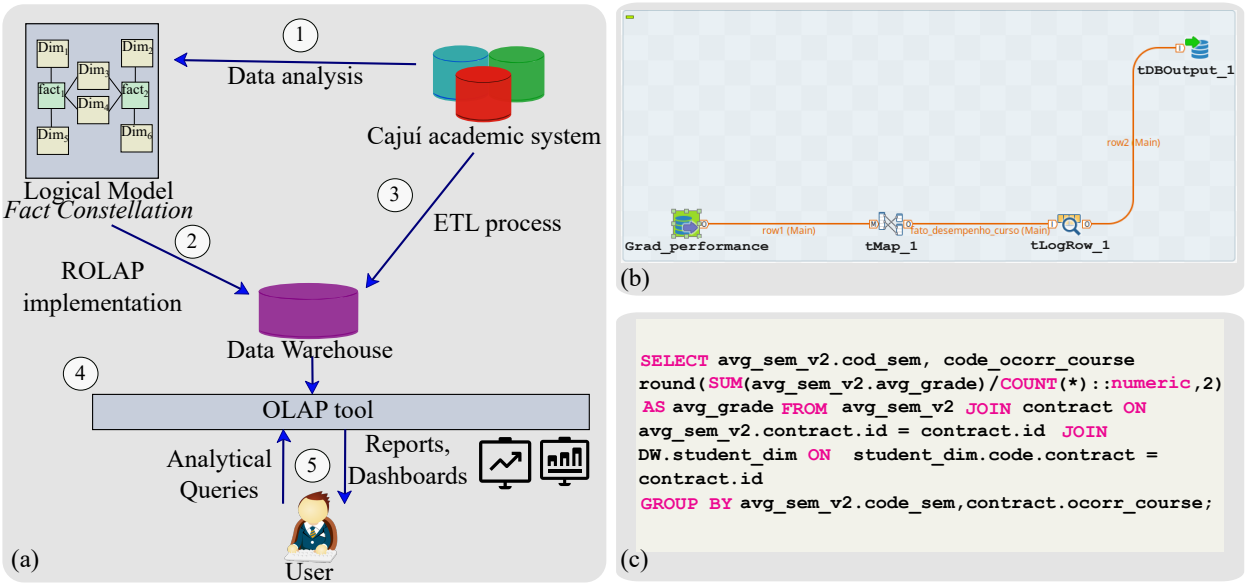

Figure 2: Conceptual methodology of IFNMG EDW. (a) Overall data searching pipeline. (b) The model for ETL process that loads student grades into a fact table. (c) The implementation of the modeled ETL process for extracting student grades from the relational tables of *Cajuí* system.

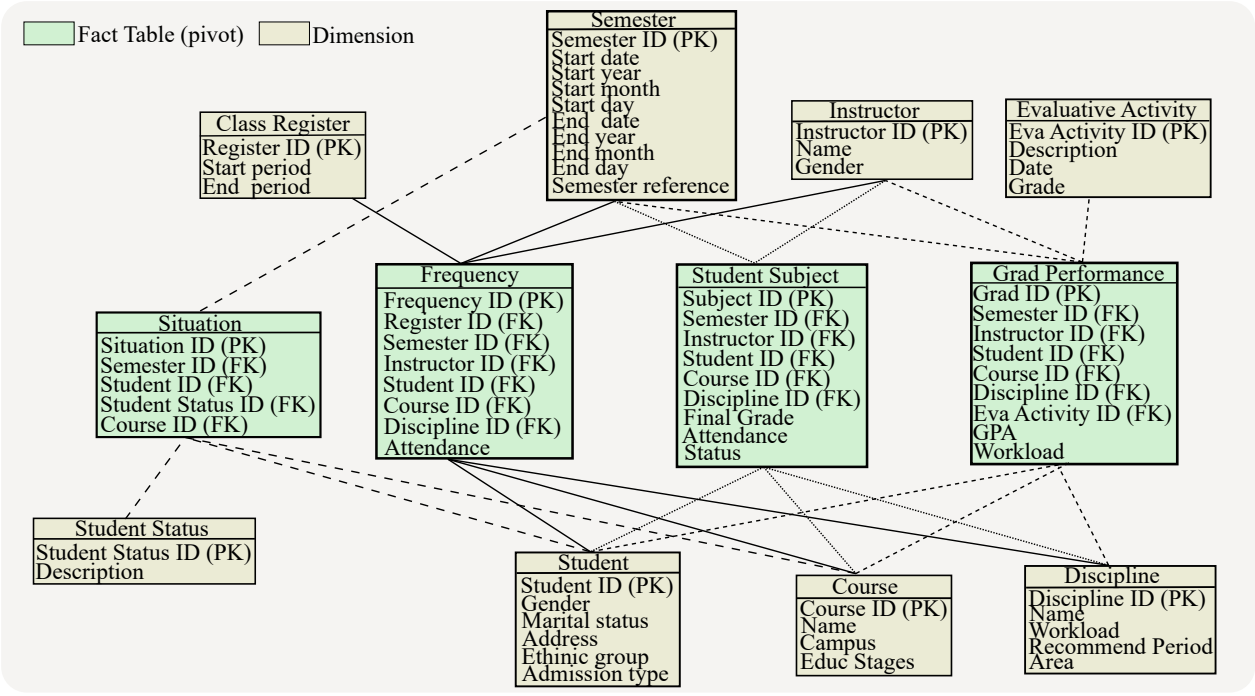

Figure 3: The *Fact Constellation* schema of IFNMG EDW with dimensional and pivot tables.

The output of Step (1) (Figure 2) is the logical model with a *Fact Constellation* schema sharing dimensions among fact tables – Figure 3. We designed eight dimension tables, namely `Semester`, with class periods; `Instructor`, with identifiers and genders of professors; `Student`, with birth dates, genders and ethnic groups; `Course`, with names, campus, and educational stages; `Discipline`, with

names, workloads, and recommended semesters; `Class Register`, which keeps information of class contents; `Student Status`, which stores possible student status throughout the course, *e.g.*, dropout; and `Evaluative activity`, which describes the date, description and score of student activities, such as exams and labs.

Additionally, we designed four fact tables joining with the `Frequency` table for measuring the students' attendance in a particular lecture for a given semester. The `Situation` table measures the number of students that either took a leave of absence or dropped out from a given course and semester. The `Grade Performance` dimension measures the achieved grades and workload, while the `Student Subject` table stores the final grades, attendance, and status in a given course for a particular student and semester.

Steps (2) and (3) (Figure 2) show the implementation and data loading tasks, respectively. As for physical modeling (Step (2)), the Fact Constellation dimensional model was mapped into a relational scheme, following a PostgreSQL-based ROLAP approach. The *Extract-Transform-Load* process (Step (3)) was then carried out by the open-source tool *Talend Open Studio*[2]. It provides several components to connect, aggregate, filter, and conform data from different formats and sources.

The building processes for the DW were separated into *Talend jobs* (units of work) defined for tables of interest (fact or dimension), while several other *jobs* (Figure 2(b)) queried the *Cajuí* database (Figure 2(c)) in parallel and performed data transformations. Steps (4) and (5) are related to the report visualization. The *Knowage Analytics Suite*[3] was used as the OLAP server extension, as it enables displaying data in different formats, *e.g.*, as charts and tables, as well as filtering subjects from the dashboard.

3.2 Modeling and implementation of *OLAP* dashboards

While the EDW is responsible for effectively storing and querying consolidated educational data, the result sets are interactively presented by *OLAP* dashboards with friendly user interfaces to decision-makers. Therefore, the integration of the EDW processes with those external tools is the first requirement for the design of an *OLAP* dashboard. We model that integration by using various Unified Modeling Language (UML) diagrams, which are summarized in Figure 4. It presents the *activity diagram* whose first stage is collecting data from both the *Cajuí* academic system (composed of relational tables) and secretary spreadsheets (separated and isolated files). As detailed in Section 3.1, relevant data are loaded into the DW through an ETL process so that cleaned and consistent data are finally stored into multidimensional tables.

The ETL processes verify atomic data consistency through a set of validations and rules. For instance, if the entry is repeated then the ETL process unifies the repeated values, merging possible diverging attribute values. Likewise, if attributes are missing then the ETL process browse other data sources (*e.g.*, tables and spreadsheets) to complete the tuple. Finally, multidimensional tables are queried, and their results are presented through dynamic dashboards. Such integration is performed with DW connectors that

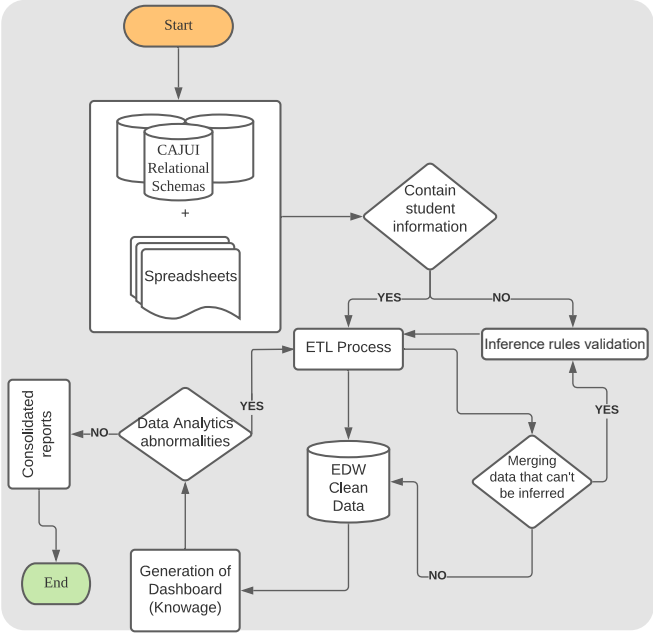

Figure 4: Diagram flow for the presentation of *OLAP* dashboards and consolidated reports.

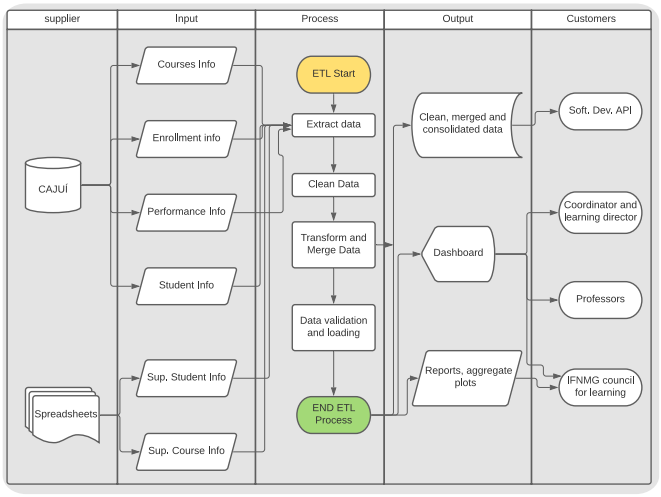

Figure 5: SIPOC diagram for accessing the `IFNMG` EDW data and reports.

lead to predefined dashboards and reports, being direct querying to DW also available through SQL statements.

Figure 5 presents the SIPOC (supplier > inputs > process > outputs > customer) diagram for accessing the results from the `IFNMG` EDW. The *suppliers* are the original data sources based on the *Cajuí* information system and spreadsheets, while the *inputs* consist of the information collected from those sources. The core activities (*process*) include the consolidation of cleaned information extracted

[2] Avaliable at https://www.talend.com/products/talend-open-studio/
[3] Available at https://www.knowage-suite.com/site/home/

by the ETL processes so that the *outputs* are consistent relations retrieved after consistent educational data. Finally, data are displayed to distinct decision-makers through dynamic dashboards, which include reports for students, professors, and the administration board (See Figures 12, and 14)[4].

4 CASE STUDY – THE COMPUTER SCIENCE COURSE

In this section, we discuss the data analytics provided by IFNMG EDW. The records include 11,106 students from 11 *campi* of 23 graduation courses in different locations, and the data ranges from 2007 to 2019 – A time window of 12 years. In particular, a preliminary analysis is presented for the Computer Science (CS) graduation at the IFNMG campus of Montes Claros after a double data checking from 2013 to 2019. The bachelor course has registered 281 enrollments during the evaluated timespan, whereas 20.28% of the candidates were registered as female students in contrast with the 49.41% overall IFNMG enrollment ratio for the same gender of students. Analytics also show female CS students usually have a substantially better overall grade in the entire bachelor course than male students, *i.e.*, an average of 53.54/100 in comparison to 44.43/100-grade points. Such enrollment cipher(combined with those consolidated performances) suggests the reinforcement of affirmative policies for female students in computer science has the potential to *(i)* enhance equality and diversity, and *(ii)* improve overall academic grades. Those data-driven findings are being considered by the administration board for next enrollments.

Since 2014, undergraduate students can apply to CS by approving one of two different exams: the standard entrance exam ("*Vestibular*") or the Brazilian ENEM/SISU exam. The number of available positions is divided in half for each of the admission criteria, *i.e.*, 50% to *Vestibular* candidates, and 50% to SISU candidates. The passing performance of students admitted by those different criteria is presented in Figure 6. Little-to-nothing differences were observed in the number of approvals for students of the two admission types in the 2014 – 2018 interval. The largest differences were observed in 2016 when SISU students approved 65% of times against 59% of *Vestibular*, and 2017, in which *Vestibular* students achieved a 57% approval ratio against the 49% of SISU. Due to that variance and the sample size, we conclude there are no significant differences between the performance of students regarding admission type.

In the examined time-spam, data also showed only 20%, on average, of enrolled students in Computer Science are females – See the dashboard in Figure 13 for a complete social profile of CS students. Therefore, they contribute to the minor portion of the unitary approval ratio in the stacked bars of Figure 6 (average upper bound of 20%). It is important to highlight, however, female students have a proportional pass ratio superior to male students per year and average. For instance, female students' pass rate was 66% against the 40% of the male students (global average of 44%) in the year with the lowest passing ratio (2014).

We also performed an EDW analysis regarding the informal dropouts of CS students. According to the IFNMG rule book of 2014–present, student admission occurs only in the first semester of each year (40 candidates each time), but bachelor candidates are allowed

[4]Demo available at http://ifnmg.edw.cloudns.nz/

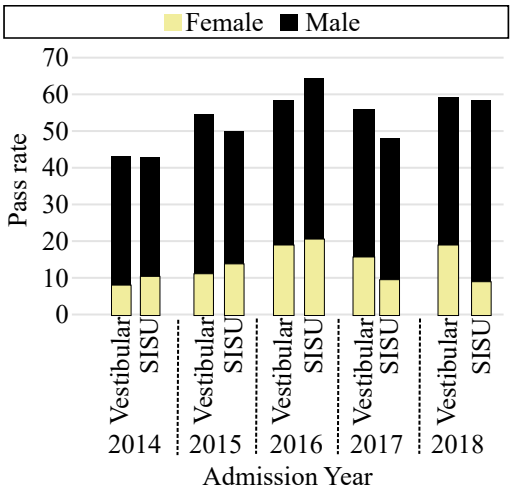

Figure 6: Passing ratio by admission type and gender.

to leave for two straight semesters. Figure 7 shows that the course entrances were nearly constant in time, but the sum of leaves of absence and dropouts has increased with time. Accumulated values indicate a total of 26 students dropped out, *i.e.*, officially left, from the course while 41 took a leave of absence. If we remove those 67 students from a total of 281 enrollments, then 214 regular students are expected to be attending the lectures regularly. However, at the beginning of 2019, only 153 students were enrolled in at least one course, which suggests 61 bachelor candidates might have unofficially dropped out, representing more than 19% of students. Figure 7 also shows the number of female enrollments doubled in numbers from 2014 to 2018 (15% to 31%), with females presenting a much smaller ratio of dropouts and absences than other students.

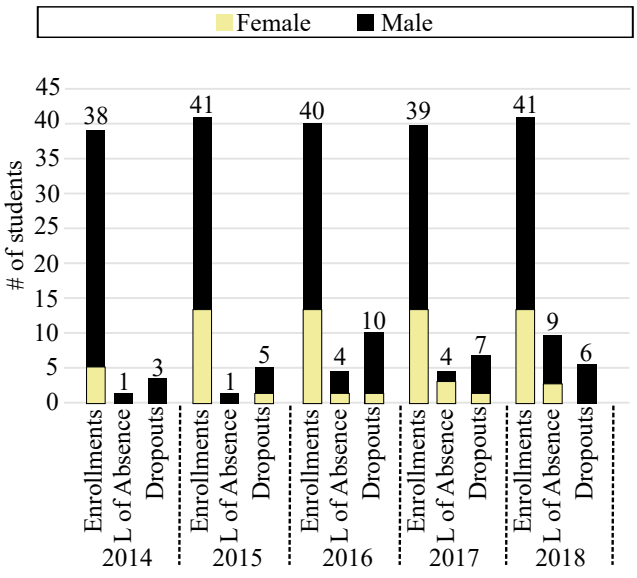

Figure 7: Enrollments, dropouts, and leaves.

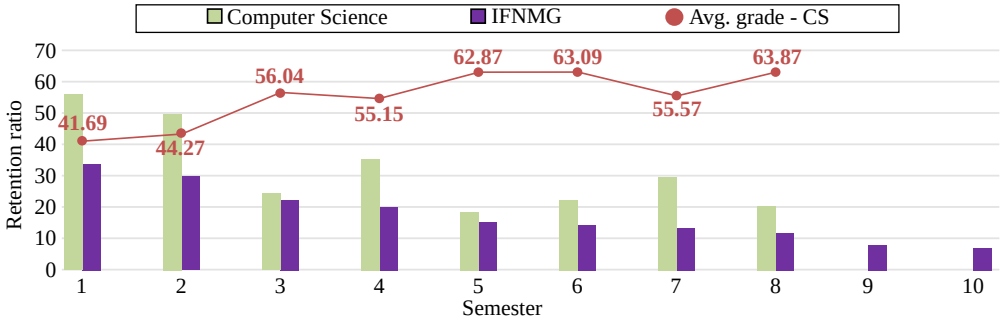

Figure 8: Retention ratio and CS students' average grades by semester.

To shed light on the real status of students that either took a leave of absence or dropped out, we *drilled* the joined EDW fact tables `Student Subject` and `Situation`. Figure 9 summarizes the results for the queries over the merged tables, which indicate the majority of bachelor candidates that dropped out had less than six failures, and students in leaves of absence are likely to have more than six failures. On the other hand, Figure 10 shows dropout students had passed at most five courses (69.23%), while students in leaves of absence had usually passed in more than one course (58.54%). Analytics also show 15.38% of dropout bachelor candidates and 19.51% of students in leaves of absence had at least 16 approvals before leaving the graduation, *i.e.*, they had already completed ½ of a Computer Science degree with 48 courses.

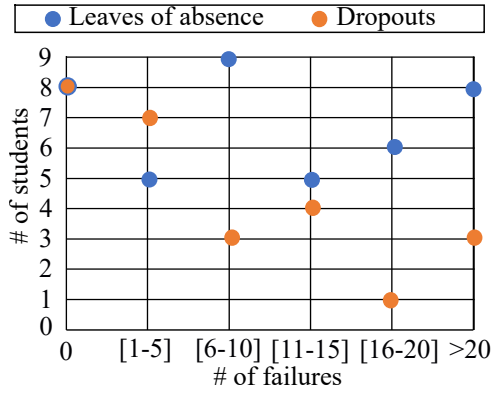

Figure 9: Number of failures until the student takes a leave of absence or drops out.

Another data aspect we obtained from the EDW system is the failure ratio by *ideal semesters*, which covers the courses that the bachelor candidates must be attending in a specific semester according to the indications of both CS and `IFNMG` technical academic guidelines. The results are visually represented in the EDW dashboard, as the hybrid graph of Figure 8. Surprisingly, data findings show, on average, the number of CS failures is higher than other `IFNMG` courses in every one of the eight semesters of the CS course (other `IFNMG` courses may last up to ten semesters).

Accordingly, we continue querying the EDW fact tables to verify how much that high failure ratio impacts the students' *academic curricula*. The results are included in Figure 8 as the line that describes the students' average grade as a $[0, 100]$ score. Data-driven results show not only the CS failure ratio tends to be higher in the first-years, but the average pass grades in the first-years are lower. Such a trend is reverted throughout the student formation, in which courses in the last years present a lower failure ratio and better average grades.

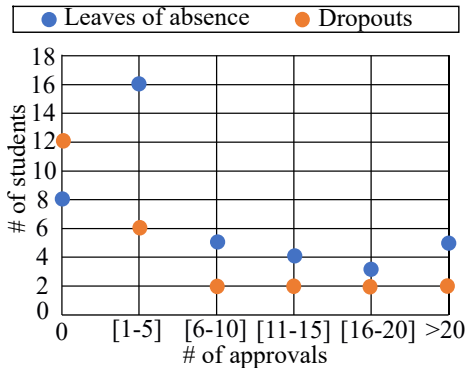

Figure 10: Number of approvals until the student takes a leave of absence or drops out.

Table 1: Ranking of courses with highest failure rate.

Ranking	Course	Ideal semester	Failure Rate
#1	Discrete Mathematics	2	69.81%
#2	Algorithms and Data Structures I	1	68.87%
#3	Analytical Geometry	1	64.73%
#4	Calculus I	1	63.18%
#5	Undergraduate Thesis I	7	58.14%
#6	Fundamental of Electronics	1	56.77%
#7	Computer Networks	4	55.67%
#8	Data Structures and Algorithms II	2	55.19%
#9	Scientific Methodology	2	54.50%
#10	Linear Algebra	2	49.17%

We pursue that investigation of the high failure ratio in CS first-year courses by drilling the EDW and unveiling the number of failures and passes by year. Table 1 shows the top-ten courses with the highest retention ratio, in which eight entries are first-year courses. In particular, the top−4 courses with the most retention are

from first-year courses, and they present a pass ratio lower than 40%. The remaining first-year entries are also listed in Table 1, whose retention ratio covers more than half of enrolled students. The EDW drill query also presents the number of students that have multiple failures in the same courses, as the bar chart in Figure 11. Such findings show 170 (total) students failed at least once in Calculus I (60.50% of attempts), whereas 29 students failed twice and 33 failed three or more times until the approval in the course. Likewise, 165 students have also failed at least once in Algorithms and Data Structures I (58.72% of attempts).

Unlike other subjects from the first period, in which the number of students who fail twice or at least three times is substantially lower than the number of students who fail once, in the Discrete Mathematics course (second period), the number of students who fail twice is practically 60% of the number of those who failed only once. When we consider the gender of the students, the proportion of retention ratio per group was constant for all examine failures within the top-4 courses on Table 1, *i.e.*, there was no correlation between gender and failing a course. Finally, the dashboard in Figure 15 shows the accumulated dropouts (84) by students that failed at least one of the top-4 courses on Table 1, which accounts for 47.7% of all accumulated CS dropouts (176). This finding devises a direction to future investigations regarding dropouts *vs.* failures regarding groups of CS courses.

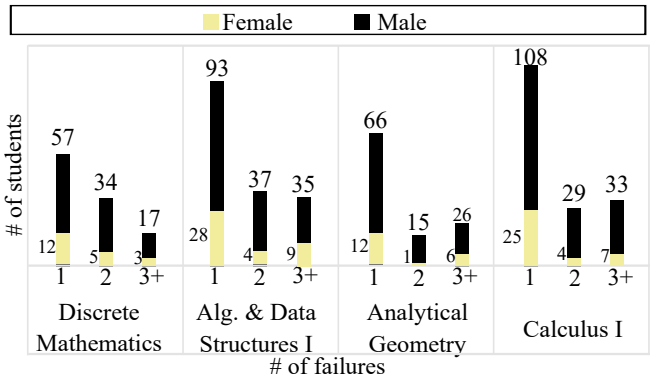

Figure 11: The consecutive number of attempts until the approval regarding the top-4 courses with the highest failure rate grouped by year and gender.

5 CONCLUSIONS AND FUTURE WORK

This manuscript has discussed the project and implementation of an EDW at IFNMG that provides data analytics regarding student profiles. The Data Warehouse was designed by using the *Fact Constellation* model persisted in the RDBMS *PostgreSQL* and accessed by the *Knowage Analytics Suite* tool. A preliminary study considering the computer science students of Montes Claros campus shows the type of information is available for the decision-makers at the institution, covering academic performances in several dimensions, *e.g.*, admission entries and course semesters. Furthermore, student dropouts can be analyzed through failures and pass ratio by courses, and compared against other institute courses. We plan to integrate machine learning and EDW to address pattern detection as future work [22].

6 ACKNOWLEDGMENTS

The authors thank the IFNMG Administration Board and its Information Technology Management for providing full access to the *Cajuí* system. The authors also thank CAPES/CNPq (G. 434421/2018-9 Universal CNPq, 313238/2018-9 PQ), FAPERJ (G. E-26/202.806/2019), FAPESP (G. 2021/06564-0), and IFNMG/PROAPE/PIBIC (G. E-No 60/2019, E-No 23/2021) for their financial support. Author M. Bedo was on leave from the Fluminense Federal University (UFF/Brazil).

REFERÊNCIAS

[1] R. Bouman and J. Van Dongen. 2009. Pentaho Solutions. *Business Intelligence and Data Warehousing with Pentaho and MYSQL.*
[2] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Erven. 2019. Educational Data Mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research* 94, 1, 335 – 343.
[3] H. Garcia-Molina. 2008. *Database systems: The complete book.* Pearson.
[4] R. Gouveia and C. Freitas. 2018. Implementação de um Data Warehouse para Análise de Dados Abertos Governamentais da Educação a Distância. *Tear: Revista de Educação Ciência e Tecnologia* 7, 2, 1 – 15.
[5] B. Griesemer. 2009. *Oracle Warehouse Builder 11g.* Packt Publishing Ltd.
[6] J. Guan, W. Nunez, and W. 2002. Institutional strategy and information support: The role of data warehousing in higher education. *Campus-wide Information Systems.*
[7] J. Han, M. Kamber, and J. Pei. 2012. *Data mining concepts and techniques.* Elsevier.
[8] W. Inmon, J. Welch, and K. Glassey. 2005. *Building the Data Warehouse.* Sons Inc, New York.
[9] M. Júnior, M. Mendonça, and F. Rodrigues. 2009. Data warehousing in an industrial software development environment. In *Software Engineering Workshop.* IEEE, 131–135.
[10] R. Kimball and M. Ross. 2002. *The Data Warehouse Toolkit: The complete guide to dimensional modeling.* Wiley, New York.
[11] L. Manhães, S. Cruz, and G. Zimbrão. 2014. WAVE: an architecture for predicting dropout in undergraduate courses using EDM. In *Symposium On Applied Computing.* ACM, 243–247.
[12] O. Moscoso-Zea and S. Luján-Mora. 2016. Datawarehouse design for educational data mining. In *International Conference on Information Technology Based Higher Education and Training.* IEEE, 1–6.
[13] O. Moscoso-Zea, J. Paredes-Gualtor, and S. Luján-Mora. 2018. A holistic view of data warehousing in education. *IEEE Access* 6, 64659–64673.
[14] F. Nunes, M. Júnior, J. Junior, L. Costa, and E. Recchi. 2019. Galactus – Um ambiente inteligente para apoio à tomada de decisão no âmbito do Ministério Público de S.E.. In *Simpósio Brasileiro de Sistemas de Informação.* SBC, 153–156.
[15] J. Oliveira Jr., L. Bastos, and C. Kaestner. 2015. Uma Abordagem de Data Warehouse Educacional para Apoio à Tomada de Decisão. In *Anais do Congresso Brasileiro de Informática na Educação.* SBC, 1064–1073.
[16] F. Pfeffer. 2008. Persistent inequality in educational attainment and its institutional context. *European Sociological Review* 24, 5, 543–565.
[17] B. Rance, V. Canuel, H. Countouris, P. Laurent-Puig, and A. Burgun. 2016. Integrating heterogeneous biomedical data for cancer research: the CARPEM infrastructure. *Applied Clinical Information* 7, 2, 260.
[18] F. Ravat and Y. Zhao. 2019. Data lakes: Trends and perspectives. In *International Conference on Database and Expert Systems Applications Conference.* Springer, 304–313.
[19] C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker. 2010. *Handbook of educational data mining.* CRC press.
[20] G. Santos, A. Bordignon, D. Haddad, D. Brandão, L. Tarrataca, and K. Belloze. 2019. Data Warehouse Educacional: Uma visão sobre a Evasão no Ensino Superior. In *Simpósio Brasileiro de Banco de Dados.* SBC, 235–240.
[21] G. A. S. Santos, K. T. Belloze, L Tarrataca, DB Haddad, AL Bordignon, and DN Brandao. 2020. EvolveDTree: Analyzing Student Dropout in Universities. In *International Conference on Systems, Signals and Image Processing.* IEEE, 173–178.
[22] D. Saraiva, S. Pereira, R. Braga, and C. Oliveira. 2021. Análise de Agrupamentos para Caracterização de Indicadores de Evasão. In *WEI.* SBC, 238–247.
[23] B. Shin. 2002. A case of data warehousing project management. *Information & Management* 39, 7, 581–592.
[24] V. Theodorou, A. Abelló, M. Thiele, and W. Lehner. 2017. Frequent patterns in ETL workflows: An empirical approach. *Data & Knowledge Eng.* 112, 1 – 16.
[25] William Villegas-Ch, Xavier Palacios-Pacheco, and Sergio Luján-Mora. 2020. A business intelligence framework for analyzing educational data. *Sustainability*

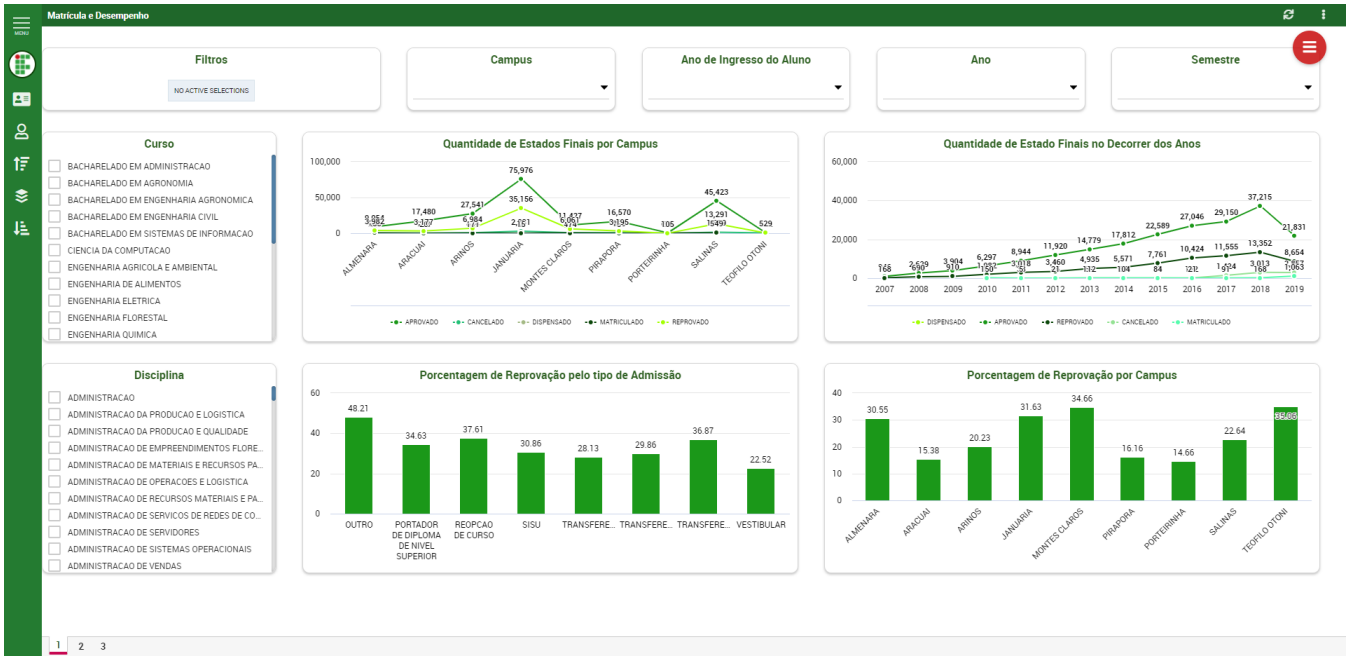

Figure 12: IFNMG EDW overall dashboard regarding every degree and *campus*. Decision-makers can examine individual courses, enrolled students, and their grades filtering and grouping by semesters and admission. Student grades grouped by enrollment are presented on the upper right, while the number of approvals and average grades is presented on the lower right.

Figure 13: IFNMG EDW dashboard with students' social profiles. Charts indicate marital status, gender, enrollments per year, ethnicity, and admission type.

12, 14, 5745.

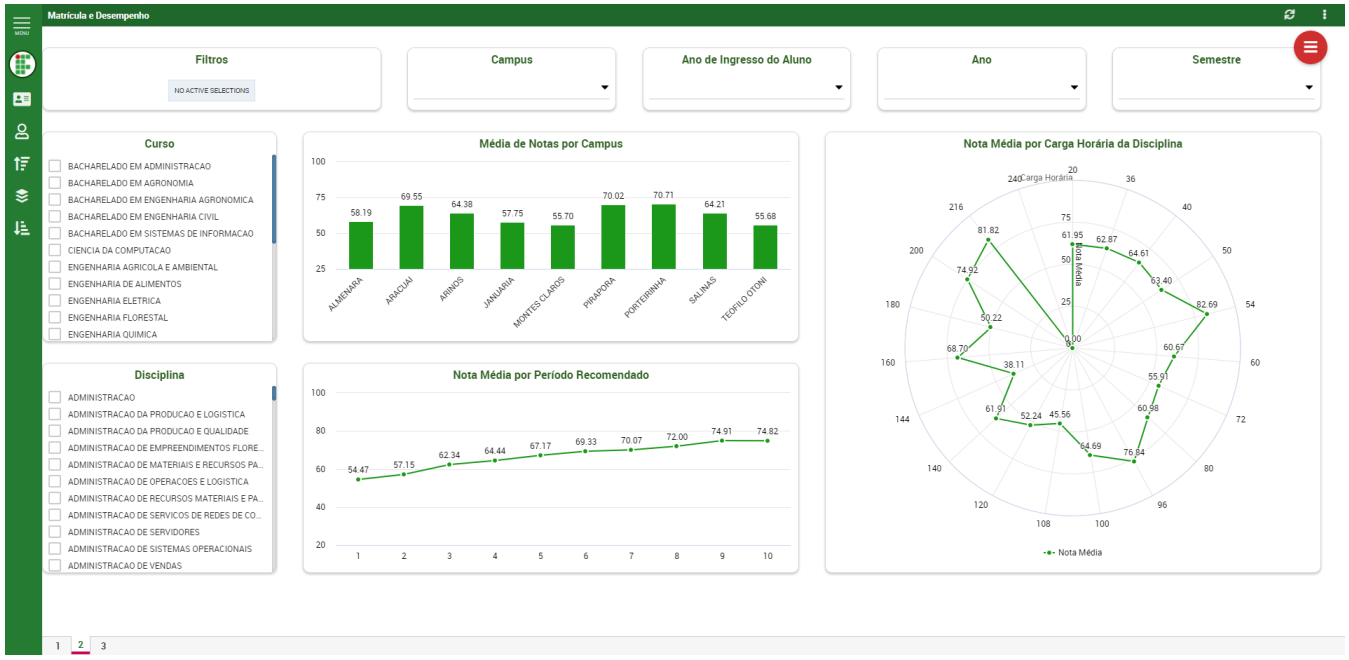

Figure 14: IFNMG Course Dashboard. Average grades by *campus* are presented by the vertical bar plot, whereas the radar chart shows the average grade per CS course. Average grades by recommended semesters are displayed as the line plot.

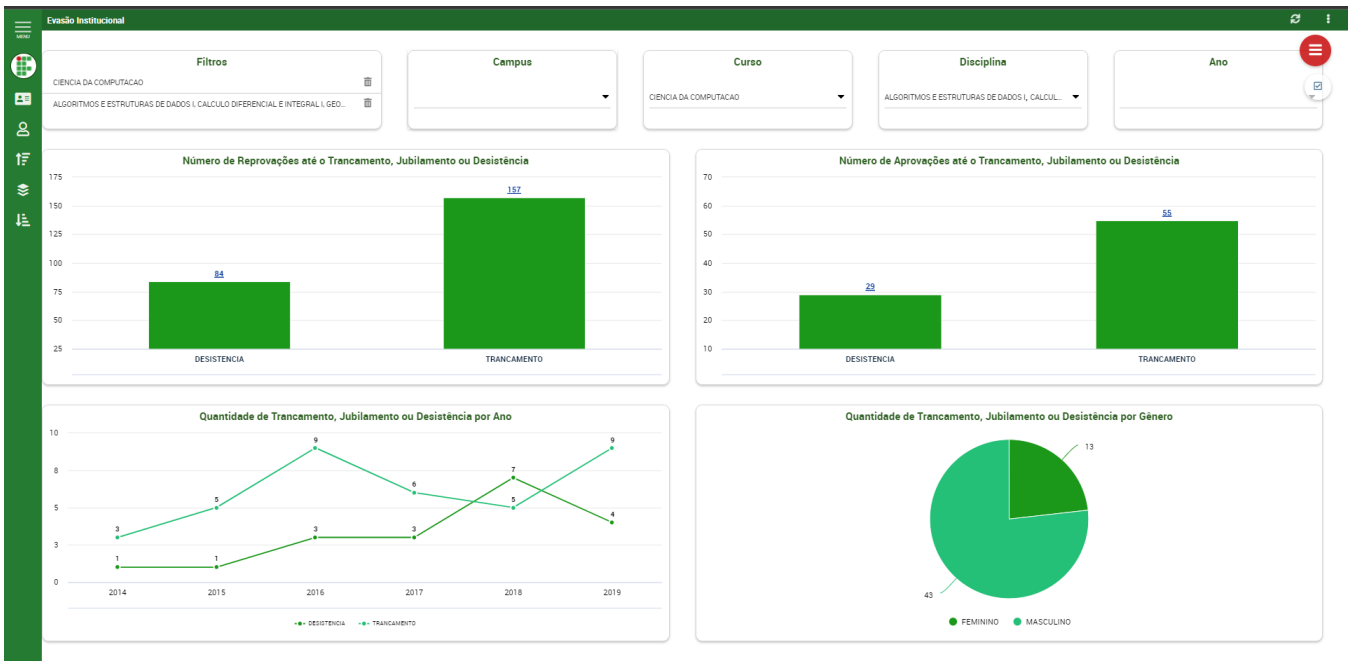

Figure 15: Dropout dashboard for the top-4 courses with highest failure ratio. Approvals and failures are presented as bar plots with the bottom left line plot detailing the information by year and gender. Plots can be drilled with a click.