Intelligent Analysis of Students Profile about Dropout Factors: A Study in Information System Course Context

Wallyce Azy[1], Regina Braga[1], Victor Stroële[1], José Maria N. David[1], Fernanda Campos[1] Luciano J. Chaves[1], Luciana Campos[1]

[1]Computer Science Department– Federal University of Juiz de Fora

CEP 36036-900 – Juiz de Fora – MG – Brazil

wallyce.azy@ufjf.br, regina.braga@ufjf.br, victor.stroele@ufjf.br, jose.david@@ufjf.br, fernanda.campos@ufjf.br, luciano.chaves@ufjf.br, luciana.campos@ice.ufjf.br

Abstract. *Student dropout from higher education is still a challenge, imposing a financial and human burden and refusing students to learn. Brazil witnessed a university dropout rate of almost 55%. This work aims to analyze the factors that lead to student dropout from Information System courses, exploring the profile of students, using intelligent techniques. The information obtained can help reduce the evasion rate and identify key actions to control the problem. We used the Design Science Research methodology to conduct our study. An analysis with data from a university, considering the LGPD was conducted to verify the proposal's feasibility. Our results show that the solution can help identify key factors that lead to dropping out.*

1. Introduction

The dropout rate from higher education in Brazil, as reported by [Mapa do Ensino Superior no Brasil 2024], is a complex issue with interconnected factors. In private institutions, dropout rates reach almost 61%, while in public institutions, it is 40%. In-person courses have 52.6% dropouts, while distance courses have 64%. The implications arising from dropouts range from the reduction of institutions' financial and human resources to the loss of professional opportunities for students. The challenge of maintaining good academic performance and the student's socioeconomic context are interconnected elements that play central roles in dropout rates. In this work, we are particularly interested in dropouts in Information System (IS) courses, exploring the profile of students who have dropped out in recent years. The study focuses on aspects not frequently analyzed, such as student assistance and university quotas, among others. Student's performance in class throughout graduation is also examined.

Specifically, the study focuses on creating an ontology [Gruber, 1993] to provide specific means for student data and map new relationships, using SWRL rules [World Wide Web Consortium 2012] to discover patterns in these data. These patterns offer insights into the reasons for course dropout, mapping previously unknown relationships and directing the study to consider specific dropout patterns.

By processing student data using ontology and SWRL rules, we seek to identify the factors that lead to dropout. This process allows the analysis of general patterns and

DOI: 10.5753/sbie.2024.244921

trends that influence students' decisions to drop out. This article aims to establish an understanding of dropout drivers through ontological analysis of student data.

The work investigates the following research question: "*How to identify factors that lead to student dropout in IS courses?*" To verify the feasibility of the solution, we conducted a study with data from students of a specific institution's IS course, anonymized considering the General Data Protection Regulation (LGPD - Lei Geral de Proteção de Dados). The analysis encompasses data from more than 10 years. These results were confronted with interviews with IS course coordinators, who analyzed the results and provided some insights.

In addition to this introduction, this article presents the related work in Section 2, the methodology in Section 3, and the feasibility study and conclusions in Sections 4 and 5.

2. Related Works

Considering student profile analysis and personalization of educational services, [Ameen et al. 2012] explore the importance of ontology-based profiles in application personalization. [da Silva 2021] discusses an ontology to represent profiles of undergraduate students related to Computer Science to understand individual characteristics in academic and personal terms. Our proposal uses ontology inference to enrich the holistic profiles of students with a focus on social factors. [El-Rady 2020] details an ontological model to predict student dropout through the combination of academic, personal, and behavioral information and proactive interventions to improve retention and academic success. [Gutiérrez et al. 2022] analyzed the profile of university students using questionnaires. Our approach uses SWRL rules to analyze data considering also other factors, outlining an academic and social profile of students.

[Priyambada et al. 2021] analyze data from students in IS courses, creating temporal clusters according to academic performance. While Priyambada et al. focus on analyzing students' behavior patterns over time, our work is primarily based on academic and social data from students who dropped out, also considering temporal issues to predict recurring behaviors to mitigate dropout. [de Oliveira et al. 2021] explore student interaction in online learning environments using Social Network Analysis (SNA) and clustering techniques. Our approach proposes to create an intelligent solution to analyze student profiles and infer insights about the reasons for dropout. [da Silva et al. 2021] reviewed the literature on intelligent services applied to distance learning. Our focus is on intelligent services for dropout control. Insights discussed in [da Silva et al. 2021] helped specify our work proposal.

[Kumaran and Malar 2023] propose a dropout prediction model in online learning environments, employing an iterative classification algorithm. Our approach provides an ontology for analyzing aspects that can be decisive for evasion. [Ajoodha et al. 2020] highlight the importance of mathematical and computing skills, which have implications for guiding and supporting students on their academic trajectory. Our work explores student skills, student academic performance, and social aspects to prevent dropout. [Gonzalez-Nucamendi et al. 2022] used questionnaires to assess Multiple Intelligences. Our study concentrates on academic performance and social characteristics and uses intelligent techniques for data analysis. [Vinker and Rubinstein 2022] analyzed how students interacted with programming tasks in an online computer science course and developed Machine Learning (ML) models to predict when students might drop out.

Our work and Vinker's are similar. Our focus uses more general metrics related to academic and social aspects. [Saqr et al. 2023] investigate the relationship between student engagement in online courses and their academic performance over a four-year program. Our work aims to help low-performing students by analyzing academic metrics and social issues. Saqr's work presents some insights that we use to improve predictions. In a similar work, [Da Cruz et al. 2023] propose a new method using ML techniques to analyze higher education dropouts in IS and other courses. Our work uses ontologies to pre-process data and then uses inference rules and semantic information related to academic and social profiles to identify factors that lead to dropout. [Aguirre and Pérez 2020] investigate university dropout using ML and polynomial regression techniques. We use ontologies to pre-process data and inference processing to identify factors that lead to dropouts. Our work is based on historical data and adds a semantic approach to the analysis conducted.

3. Methodology

We developed the work following the Design Science Research (DSR) methodology [Wieringa 2009]. The DSR proposes the execution in cycles. The artifact is evaluated in each cycle, and if new improvements are necessary, a new development cycle can be proposed.

3.1. EducAAr Architecture

In the first cycle, we analyzed related literature and considered our practical experience with similar projects at a specific Brazilian institution. As a solution, we specified an intelligent architecture, EducAAr (Educational Analysis Architecture), to identify key factors that lead to university dropout. We combined data analysis techniques, using ontologies, to analyze the profile of students and predict factors that lead to dropout.

Figure 1 presents the main components of the architecture. To help understand the solution, the datasets, models, and methods are detailed in Section 4, which explains the EducAAr use in each phase. A critical characteristic of EducAAr architecture[1] is that it uses an ontological model to represent the dynamicity of knowledge generated. The semantic profiles of students are dynamic. They evolve based on their activities, behavior, grades, and disciplines. The ontology receives constantly new data to reflect this dynamicity. As it is a work in process, it is important to note that the Machine Learning component is not part of the first DSR cycle.

[1] The EducAAr architecture was implemented in Python, using the owlready2 library, the Hermit reasoner for information inference, and PyCaret library for ML processing.

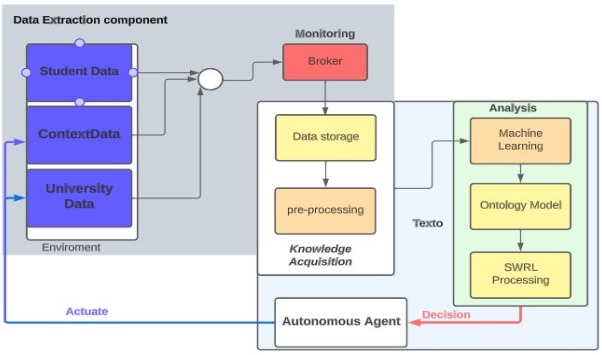

Figure 1 - EducAAr Architecture Main Components

Data Analysis component: the contents stored and processed in the ontology are retrieved, to be processed by ML techniques to identify factors for dropout. This component has two modules: (i) the module for discovering new connections between data, which uses the Ontology model and Semantic Web Rule Language (SWRL) processing (Figure 1) to analyze the relationships between them, and (ii) Machine Learning, that is not detailed in this article.

Data Extraction component: the data is obtained and integrated through a broker, through an API for combining data from several sources. The acquired data is combined considering contextual information (important in our personalized approach) and other data, such as courses, disciplines, and average grades from a specific discipline. Other data can also be combined, such as extracting educational and holistic information from comments, search results, posts, tweets, blogs, logs, etc. This data is instantiated into the ontology to be processed, using inference algorithms.

4. Feasibility Study

This feasibility study considers the use of EducAAr to identify factors that lead to students dropping out. The aim is to help course coordinators and educational leaders decide the best strategies for mitigating the dropout. In this initial phase, we used the Data Extraction Component, the Ontology Model, and SWRL processing from the Data Analysis component. The RQ (Research Question) analyzed was "*How to identify factors that lead to student dropout in IS courses?*" The data extracted is available at https://github.com/WALLYCE/OntologyDropOut We used EducAAr's Data Extraction Component to access the data. This data processing involved 439 students and their academic histories, comprising 13,518 records with student assistance type, quotas, and grades.

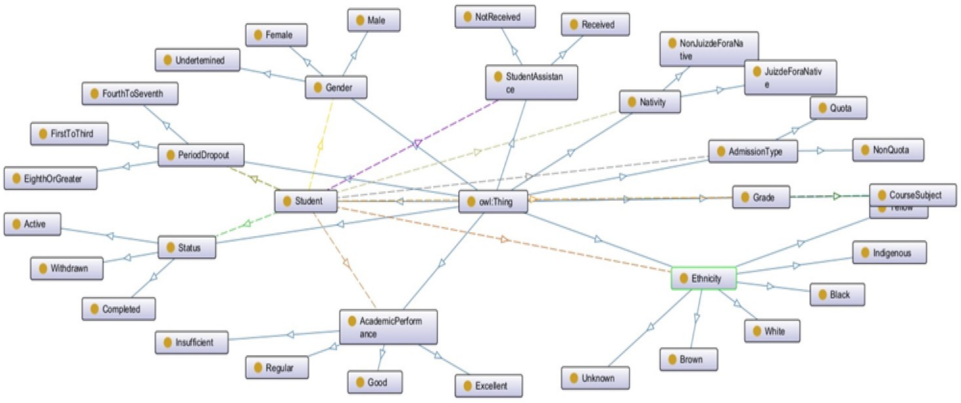

Figure 2 - Ontology classes and relationships

To develop the ontology, the six phases of the "Methontology" [Fernández-López et al., 1997] were executed. Classes and their relationships were defined to represent the variables identified as relevant to the study. Table 1 presents details of the main classes of the ontology, and Figure 2 presents a diagram with the main classes.

Table 1- Main ontology classes

Student: Who is or was a student in the course?	**Status**: Represents student's status in the course, with subclasses: Active, Completed, Withdrawn
Nativity: The student's place of birth, with subclasses: Born in the city of the institution or not.	**PeriodDropout**: The period in which the student dropped out, with subclasses: First to Third, Fourth to Seventh, Eighth or Higher
Ethnicity: student self-declared. Subclasses: White, Black, Yellow, Brown, Indigenous, Unknown.	**Grade**: The grade obtained by the student during the course.
Gender: The biological gender of the student, with subclasses: Male, Female, Indeterminate.	**CourseSubject**: Represents a specific course subject.
AdmissionType: Describes how the student entered the institution, with subclasses: Quota: Students admitted through the quota system, NonQuota: Students admitted through open competition.	**StudentAssistance**: Represents if student received financial assistance from the institution, either through projects or scholarships, with subclasses: Received, NotReceived. [2]
AcademicPerformance: The student's performance throughout the course: $\text{Academic Performance} = \frac{\sum_{i=1}^{n} \text{Grade}_i \times \text{Credits}_i}{n}$	We define 4 performance levels: insufficient from 0 to 60, Regular from 60 to 70, Good from 70 to 80, and Excellent from 80 to 100.

An OWL file with the specification of all classes, object properties, rules, and related queries was generated. The data was loaded into the ontology, considering all students from a specific institution IS course and their social and academic characteristics. The students' history throughout their degree was also instantiated. The Data Extraction Component performed this data integration. To illustrate the use of the inference mechanism, two rules (one property chain and a SWRL rule) specified in the ontology are presented: (a) StudentObtainedGrade o GradeBelongsToCourseSubject->StudentCompletedSubject; (b) Student(?s) ^ StudentHasStatus(?s, Completed) -> sqwrl:select(?s, "Completed"). Figure 3 presents the results (we are using the protégé

[2] Student assistance is continuous, i.e., students who receive assistance only lose it in cases of academic failure. At this stage, we do not consider the period of assistance, but whether the student received it throughout their academic life.

tool screenshot to enable the visualization of the inference mechanism – in yellow) of the inference mechanism using the property chain (a).

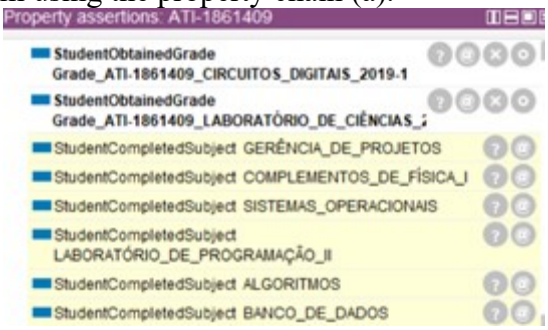

Figure 3 - Results of reasoner processing (in Protégé tool)

4.2 Results

With reasoner processing on the data instantiated in the ontology, using specified SWRL rules or SPARQL queries, we obtained relevant information to help coordinators, and educational leaders make decisions about mitigating dropout rates in Information Systems courses based on the factors highlighted by our solution. The SPARQL and graphs of some of the analyses are not shown for space restrictions. However, they can be reached at https://github.com/WALLYCE/OntologyDropOut .The IS course, as expected, is a predominantly male course. This male predominance is not unique to this course but reflects a broader trend in the industry. Despite ongoing efforts to foster gender diversity in this field, there remains a need to intensify inclusion, encourage female participation, and build a more equitable environment. Based on data analysis and processing of specific inferences, the IS course has a high dropout rate. Our analyses highlight a significant dropout rate in the IS course, becoming even more evident when we examine the completion rates, which are around 10.7%. Although the course is familiar, the persistence of many active students also suggests a considerable challenge for students in completing the course. Considering student performance, the performance profile of students who completed and those who dropped out offers a comprehensive view of academic quality and the challenges faced. In the group of graduates, success is notable, with 34% achieving excellent performance and 40.4% achieving good performance. Although a considerable portion, 14.9%, performed regularly, and a smaller portion, 10.6%, performed poorly.

Among the students who dropped out (Figure 4), the predominance of 88.5% with insufficient performance indicates the correlation between academic difficulties and the decision to abandon the course. In contrast, 7.2% showed average performance, while only 1.4% achieved excellent performance, and 2.9% achieved good performance, highlighting an important and clear connection between academic performance and students' course dropout. The SPARQL1 query generated the graph in Figure 4.

SPARQL1: Academic Performance of dropout Students

```
SELECT ?subclasse (COUNT(?student) AS ?count) WHERE { ?student rdf:type ont:Student. ?student ont:StudentHasStatus ?status. ?status rdf:type ont:Withdrawn. ?student ont:StudentHasAcademicPerformance ?performance.

?performance rdf:type ?subclasse. ?subclasse rdfs:subClassOf ont:AcademicPerformance.}GROUP BY ?subclasse
```

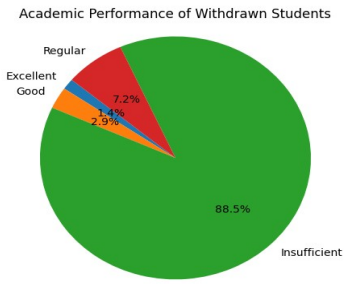

Figure 4 - Academic performance of dropout students

Analyzing the data regarding student assistance (Figure 5), it is highlighted that students who receive financial support from the university are significantly more likely to complete the course. Of the students who benefited from financial support, 40.2% completed their degree, while 59.8% chose to abandon the course. In contrast, dropout rates are notably higher among students who did not receive financial support, reaching 93.9%, with only 6.1% achieving course completion. These data highlight the substantial importance of financial assistance to students, highlighting that the probability of a student completing the course when receiving this support is approximately seven times greater than that of a student without assistance.

Furthermore, financial assistance significantly reduces the probability of dropping out, providing a reduction of almost 34%. These findings reinforce the relevance of effective student assistance policies as facilitators of academic completion and as tools for preventing student dropout. Data on the dropout period are distributed over time, with a significant concentration in the initial periods and an increasing proportion in the advanced periods. Dropout in the initial periods may reflect inadequate course choices or even an opportunity for students to migrate to other courses that are more aligned with their expectations. This initial phase often serves as a period of experimentation for students where they can reevaluate their academic preferences and goals.

SPARQL2: Students who did not receive student assistance

```
SELECT ?subclasse (COUNT(?student) AS ?count)WHERE { ?student rdf:type ont:Student. ?student ont:StudentReceivedStudentAssistance ?assistance. ?assistance rdf:type ont:NotReceived.

  ?student ont:StudentHasStatus ?status. ?status rdf:type ?subclasse. ?subclasse rdfs:subClassOf ont:Status.

FILTER (?subclasse = ont:Completed || ?subclasse = ont:Withdrawn)}GROUP BY ?subclasse
```

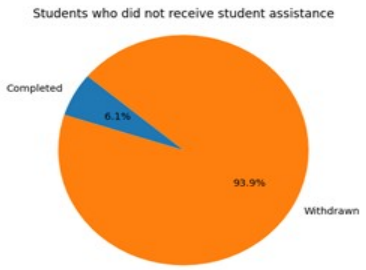

Figure 5 - Without financial support

In the case of dropping out in more advanced periods, the correlation with low academic performance is notable since 88% of students who dropped out showed insufficient performance. Academic difficulties may play a significant role in this process. However, it is important to note that there are exceptions, such as students with satisfactory performance who also decided to abandon the course. A possible reason for

this could be an early entry into the job market, leading these students to prioritize professional experience over continuing their studies.

A very controversial factor is the performance of students who enter through a quota system, whether social or ethnic. We noticed a certain similarity in the numbers when examining these data results. While quota students have a completion rate of 12.5%, those who were not admitted through quotas register a higher rate, reaching 22.5%. As already highlighted in previous graphs, students' financial issues stand out as a crucial factor in their retention at university. It is important to note that some of the admission methods through quotas cover students with an income of, at most 1.5 minimum wage. The difference in completion rates between the two groups suggests that financial situation significantly influences students' academic trajectories.

4.3 Observations and Discussion

We can answer the competency questions using the data analysis results. However, to verify the results of these analyses, we carried out a survey[3] with two IS course coordinators. We named them as Coordinator1 and Coordinator2. Next, we present the results confronted with the coordinator's answers.

Does academic performance influence student dropout rates? If so, what is the extent of this influence? Considering the results presented, academic performance directly influences the dropout decision. Coordinator1 highlighted that a significant cause of dropout is the difficulty in passing subjects, especially mathematics, contributing to low academic performance. Coordinator2 mentioned that more than 70% of students have an Academic Performance Index (ARI) below 60, which indicates that low performance is a critical factor in dropouts, particularly in the first periods of the course.

Is admission through quotas correlated with dropout rates or academic performance? The analysis presented in section 4.2 support the argument for the importance of financial assistance to students, highlighting that the probability of a student completing the course when receiving this support is approximately seven times greater than that of a student without assistance. Coordinator1 observed that quota students face more financial difficulties but have more opportunities for student assistance. Coordinator2 mentioned a study that analyzed the admission profile and concluded that quota students suffer more retention and drop out.

Can student assistance effectively reduce dropout rates? Coordinator1 commented that students in vulnerable situations have a greater chance of dropping out, but student assistance has been effective, despite still focusing mainly on financial aid. Coordinator2 suggested that the student assistance policy needs to improve, especially with a focus on leveling and pedagogical and psychological support.

Can the Information Systems students' earlier insertion in the job market influence the dropout rates? We can have evidence, based on the results discussed on section 4.2. However, we cannot affirm that. Coordinator1 mentioned that dropouts between the middle and final periods of the course occur mainly among students who are already in the job market and do not see the need to complete their degrees to progress in their careers. Coordinator2 highlighted that many students work during the day and study at night, which makes it difficult to reconcile both activities, contributing to low performance and dropout rates.

Therefore, considering our RQ: **"How to identify factors that lead to student dropout in Information System courses?"**, supported by the answers of CQ and two IS

[3] https://drive.google.com/file/d/1tKgMeSSUzr40aB2c3Y0SfaQMYJoEyTrG/view

coordinators' analysis of the results, we can have evidence of some factors that can affect student dropout, supported by data extracted from the Information System course at a specific Brazilian institution. The main factors highlighted in our work were: Academic performance; Quota availability; Financial assistance; Early entrance in the job market[4]. In this vein, specific actions can be taken by coordinators and educational leaders considering these factors.

4.4 Next Steps

For the next steps, we will use ML techniques to make individual predictions for students, using the insights obtained. These insights will be essential to inform the model about patterns and indicators associated with student dropout, thus enabling more accurate and effective analysis.

In the next steps of this work, this semantically structured information will be the basis for applying machine learning techniques for individualized predictions. Ontologies are also used to prepare data for ML processing. Preparing data for use by AI algorithms is essential for better processing and insights, and using ontologies is considered a suitable method for this preparation [Ozkaya 2023], especially when working with a "cold start" or the need to clean or add semantic information to data.

5. Conclusions

This work presented an architecture for students' dropouts in IS undergraduate courses by extracting data from students' profiles and context. EducAAr aims to assist coordinators and educational leaders in analyzing student dropout. EducAAr implemented data analysis techniques, favoring the understanding of the data. It used ontologies to analyze data. This innovative solution identifies specific dropout patterns, and uses implicit information, i.e., information discovered using inference algorithms, to understand student dropout data. To collect evidence of our approach's feasibility, an evaluation was carried out using data from a specific Brazilian university and analyzed by IS coordinators. We can cite as main contributions: a) The development of an ontology to analyze dropout issues; b) the development of EducAAr, capable of analyzing student's dropout factors; c) the Conduction of a feasibility study.

In future work, we aim to explore data from other IS courses and use ML techniques to support individual analyses and predict dropouts. A more comprehensive evaluation is also necessary, considering the proposal's impacts on the students' dropout rates.

Funding

This work was partially funded by UFJF/Brazil, CAPES/Brazil, CNPq/Brazil (grant: 307194/2022-1), and FAPEMIG/Brazil (grant: APQ-02685-17), (grant: APQ-02194-18).

[4] Early entry may be a factor, but it depends on a more direct analysis with dropout students; in the study, we focused on the more qualitative aspects of dropout.

References

Aguirre, C. E., & Pérez, J. C. (2020, October). Predictive data analysis techniques applied to dropping out of university studies. In 2020 XLVI Latin American Computing Conference (CLEI) (pp. 512-521). IEEE.

Ajoodha, R., Dukhan, S., & Jadhav, A. (2020, November). Data-driven student support for academic success by developing student skill profiles. In 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC) (pp. 1-8). IEEE.

da Cruz, R. C., Juliano, R. C., Monteiro Souza, F. C., & Correa Souza, A. C. (2023, May). A Score approach to identify the risk of students dropout: an experiment with Information Systems Course. In Proceedings of the XIX Brazilian Symposium on Information Systems (pp. 120-127).

da Silva, Claiton. A holistic profile ontology for undergraduate students. 2021. https://bdm.unb.br/bitstream/10483/31241/1/2021_ClaitonCustodioDaSilva_tcc.pdf. accessed in: ago. 2023. (in Portuguese)

da Silva, L. M., Dias, L. P. S., Rigo, S., Barbosa, J. L. V., Leithardt, D. R., & Leithardt, V. R. Q. (2021). A literature review on intelligent services applied to distance learning. Education Sciences, 11(11), 666.

De Oliveira, P., da Silva, G., Dourado, R., & Rodrigues, R. L. (2021). Linking Engagement Profiles to Academic Performance Through SNA and Cluster Analysis on Discussion Forum Data. In LALA (pp. 39-47).

El-Rady, Alla Abd . An Ontological Model to Predict Dropout Students Using Machine Learning Techniques . Arab Academy for Science , Technology & Maritime Transport. Egypt, 2020. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9096743.

Fernández-López, M., Gómez-Pérez, A. and Juristo, N. (1997). “Methontology: from ontological art towards ontological engineering”. Spring Symposium Series, 1997. Facultad de Informática (UPM).

General Data Protection Law (LGPD) - Brazil**.** *Effective Date.* Brasília, DF: Presidency of the Republic, [year of publication]. Available at: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., García-Castelán, R. M. G., & Escobar-Castillejos, D. (2022). Learning Analytics to Determine Profile Dimensions of Students Associated with Their Academic Performance. Applied Sciences, 12(20), 10560.

Gruber, Thomas R. . A Translation Approach to Portable Ontology Specifications . 1993. https://tomgruber.org/writing/ontolingua-kaj1993.pdf.

MAPA do Ensino superior no Brasil. Brasil, 8 maio 2024. Disponível em: https://static.poder360.com.br/2024/05/mapa-do-ensino-superior-no-brasil-202.pdf. Acesso em: 5 jun. 2024.

Ozkaya I., "Application of Large Language Models to Software Engineering Tasks: Opportunities, Risks, and Implications," in IEEE Software, vol. 40, no. 3, pp. 4-8, May-June 2023, doi: 10.1109/MS.2023.3248401.

Peffers, Ken et al. A design science research methodology for information systems research. Journal of management information systems, v. 24, n. 3, p. 45-77, 2007.

Priyambada, S. A., Er, M., Yahya, B. N., & Usagawa, T. (2021). Profile-based cluster evolution analysis: Identification of migration patterns for understanding student learning behavior. IEEE Access, 9, 101718-101728.

Saqr, M., López-Pernas, S., Helske, S., & Hrastinski, S. (2023). The longitudinal association between engagement and achievement varies by time, students' profiles, and achievement state: A full program study. Computers & Education, 199, 104787.

Vinker, E., & Rubinstein, A. (2022, March). Mining code submissions to elucidate disengagement in a computer science MOOC. In LAK22: 12th international learning analytics and knowledge conference (pp. 142-151).

Wieringa, Roel. Design science as nested problem solving. In: Proceedings of the 4th international conference on design science research in information systems and technology. 2009. p. 1-12.

World Wide Web Consortium. (2012). OWL 2 Web Ontology Language Document Overview. W3C Recommendation. Retrieved from https://www.w3.org/TR/owl2-syntax/