*Journal of the Brazilian Computer Society, 2026, 32:1,* doi: 10.5753/jbcs.2026.5613


# *LearnVis*: Analyzing Higher Education Student Performance through Information Visualization Techniques

**Angélica Gomes Oliveira** [ **Federal University of Uberlândia** | ***angelicag_oliveira@hotmail.com*** ]
**Paulo Henrique Ribeiro Gabriel** [ **Federal University of Uberlândia** | ***phrg@ufu.br*** ]
**José Gustavo de Souza Paiva** [ **Federal University of Uberlândia** | ***gustavo@ufu.br*** ]

*Faculty of Computing, Universidade Federal de Uberlândia (UFU), Av. Joao Naves de Avila, 2121, 38408-100 Uberlandia, Brazil.*

**Received:** 28 February 2025 • **Accepted:** 06 October 2025 • **Published:** 02 April 2026

**Abstract** Understanding educational challenges in higher education requires a detailed analysis of the variables related to academic performance, such as grades, attendance, and student engagement. This analysis is crucial for identifying critical factors that affect learning and student retention. In this context, this study introduces *LearnVis*, a visual analytics system developed to analyze student performance in higher education. The system was designed to utilize data from university academic records, including grades, attendance, and engagement with specific topics. It offers a set of coordinated layouts that enable the analysis of both student groups and individuals. These layouts allow users to explore the structure of course modules, considering the topics they comprise and their sequence throughout the course. Additionally, the system facilitates the analysis of student behavior in each module, including their attendance in the topics covered, their respective grades, and the tracking of multiple attempts in specific modules, as well as their completion sequences. To evaluate its effectiveness, the *LearnVis* system was applied to a dataset of 1,490 students from the Computer Information Systems program at the Federal University of Uberlândia, covering the period from 2009 to 2019. The results demonstrate that the system provides valuable insights that contributes to improve academic performance and decrease student retention.

**Keywords:** Student Performance Analysis, Educational Data Analysis, Visual Analytics, Information Visualization, Multidimensional Projection

## 1 Introduction

Understanding the factors that influence students' academic performance is crucial for educational managers to create or adapt strategies and policies that will mitigate retention and dropout rates. According to the Brazilian National Institute for Educational Studies and Research Anísio Teixeira (Inep) and the Ministry of Education (MEC) [INEP, 2020], from 2010 to 2019, the cumulative students dropout rate in their first higher education courses was 59%. From these students, only 40% managed to graduate the same course they initially enrolled. In 2021, over 2 million students abandoned their courses, resulting in a dropout rate of 38.8% in private Higher Education Institutions (HEIs) and 9.4% (165,000 students) in public HEIs [Minhoto *et al*., 2023]. The quality of education in the country plays a fundamental role in socioeconomic development, impacting economic growth, job creation, and worker income [Portela, 2023]. These retentions affect graduation numbers, leading to a reduction in professionals across various sectors of society. Additionally, they impact university budgets due to the need of allocating more physical and human resources to accommodate additional classes. In 2022, a survey revealed a talent shortage faced by 75% of professionals globally [ManpowerGroup, 2022]. In the Brazilian context, the same survey indicated that the talent shortage exceeded the global average, reaching an index of 81%.

The importance of educational data for the continuous improvement of education and learning is widely discussed in research studies [Kuh *et al*., 2011; Susnjak *et al*., 2022], as it provides analytical tasks that guide expert decision-making. Educational institutions store vast amounts of data related to student performance, such as attendance, grades, module content, and completed curricula, which we refer in this work as **internal data**. Additionally, they collect demographic, social, and economic information, which we refer in this work as **external data**, that may also influence students' academic lives [Oqaidi *et al*., 2022; Gutierrez-Pachas *et al*., 2023].

It is possible to find several computational approaches to analyze academic performance, employing machine learning techniques in important tasks such as dropout prediction, grade prediction, and learning recommendations [Mushtaq and Khan, 2012; Yağcı, 2022; Pallathadka *et al*., 2023]. Visual Analytics strategies are also employed to enhance the analysis [Garcia-Zanabria *et al*., 2022; Zhang *et al*., 2022; Deng *et al*., 2019], transforming abstract and complex data into clear, intuitive, and meaningful visual representations [Munzner, 2014], used to assist the comprehension of particularities and intrinsic characteristics of the data. However, most of these works are based on external data, especially socioeconomic and demographic information related to students. Even when internal academic data is used, it is often presented in a summarized form or merely serves to complement visualizations based on external data collections [Garcia-Zanabria *et al*., 2022; Zhang *et al*., 2022; Deng *et al*., 2019], that, al-

though showing some correlation with student performance, may not represent an effective source of analysis. Often, a student's performance in a module can be impacted by his/her performance in past modules, or affected by a lack of content in these past modules, highlighting the importance of also considering internal data on academic life.

In this work, we present a visualization system called *LearnVis* for analyzing undergraduate academic performance. We employ information from students attendance and grades records, as well as information from course structure in terms of modules and their composing subjects. The system employs a set of coordinated layouts that allow users to explore the structure of course modules, both in terms of the topics they comprise and in terms of their sequence in the course. It also allows the analysis of students behavior in each module, reflected in their frequency in these module composing topics and their respective grades, as well as to track multiple students attempts in specific modules and their completion sequence. We believe that the use of students' internal data can provide a valuable analysis to coordinators and lecturers, among other educational professionals, which will benefit from the results of these analyses to formulate and evaluate policies aimed at improving academic performance. In summary, the main contributions of this work are:

I. a visual analytics system that supports the interactive exploration of data from any university academic records to analyze students' performance and its relationship with course modules' structure;
II. an anonymized data repository, publicly available, containing a rich set of information about students' performance in Computer Science courses from a Brazilian university;
III. a set of real-world case studies, employing the proposed system in the data repository from II., demonstrating the effectiveness and usability of our approach, as well as how it helps to guide education experts' decision making.

In this article, we employ our system on real data from the Faculty of Computing (FACOM) at the Federal University of Uberlândia (UFU), in Brazil, but it can be applied to data from any course from any educational institution, thus expanding the scope and relevance of the conducted analysis.

The structure of this work is organized as follows. Section 2 introduces the theoretical foundations related to education. Section 3 discusses the main related works in the literature. Section 4 presents the system requirements and the analytical tasks that guided the development of the proposed approach. Section 5 describes the dataset used in the system and the preprocessing steps applied. Section 6 details the proposed strategy and the implemented visualization system. Section 7 reports the case studies and provides a discussion of the main findings. Section 8 outlines the limitations of the study, and Section 9 presents the final remarks and directions for future work.

# 2 Education Fundamentals

The analysis of educational data is an essential tool for educational managers, as it allows the identification of patterns and insights that are difficult to detect through manual record analysis. To facilitate the understanding of the proposal and its related works, some fundamental concepts and the role of educational data, commonly employed in the literature, must be outlined. In an educational institution, **students** are central to the study. These students enroll in **courses**—structured programs comprising **modules** aimed at imparting knowledge and competencies in specific areas of study. Modules represent specific **subjects** or areas of study that revolve around **topics** taught by lecturers responsible for transmitting knowledge and developing skills. These modules are typically organized within **academic semesters**, dividing the academic year into distinct periods for teaching and learning. **Classes**, including lectures, group discussions, and practical activities, provide these contents to the students. Student performance is evaluated using **grades**, reflecting their success or failure in meeting the required standards. **Absences** from classes and activities, alongside **attendance** and participation, can negatively impact their performance.

Educational data plays a crucial role in analyzing and understanding student performance, supporting informed educational decisions. This data allows for continuous monitoring of academic progress and offers critical insights for educators and administrators to formulate effective policies and implement personalized interventions.

In this work, we categorize the available educational data into two categories: **external data** and **internal data**. External data includes demographic and socioeconomic information, such as age, gender, family income, parents' education level, and the student's social context, and although not directly related to academic performance, they can significantly influence it. Internal data, on the other hand, encompasses information directly associated to the student's academic journey, such as grades, attendance records, participation in curricular activities, course history, and assessments. These internal features are fundamental to understand the student's engagement and progress within the educational environment, as well as to comprehend how the course structure, in terms of sequence of modules, impacts their academic success [Realinho *et al*., 2022; Rahayu and Dong, 2023].

Layouts and visualization techniques are essential for organizing and interpreting this vast array of data, ensuring clear communication of insights. The analysis of educational data not only provides a robust foundation for understanding students' performance but is also essential for the continuous advancement of the educational system. The effective use of this data supports informed decision-making and promotes a proactive approach to ensuring academic success.

# 3 Related Work

## 3.1 Automatic Analysis of Educational Data

Several approaches employ machine learning techniques to analyze educational data, often to predict academic per-

formance and to identify the factors that influence student progress [Zaki and Meira, 2014]. Decision Trees and Random Forests are often used to predict grades and engagement with high accuracy [Hussain and Khan, 2023; Badal and Sungkur, 2023]. Random Forests are also used, as well as Neural Networks models, to analyze school dropouts factors, and to assist targeted interventions to prevent these dropouts, achieving satisfactory effectiveness results [Gutierrez-Pachas *et al*., 2023; Pachas *et al*., 2021; Martins *et al*., 2023]. Additionally, recommendation systems incorporating Natural Language Processing and Virtual Agents [Shahbazi and Byun, 2022; Ali *et al*., 2022] enhance the learning experience by suggesting courses and subjects based on student preferences.

Linear regression has also been widely applied in the educational context, particularly for predicting students' academic performance. Features such as attendance, previous grades, and socioeconomic characteristics are frequently used to model and forecast outcomes such as final grades or completion rates [Mengash, 2020; Mohd Arsad *et al*., 2014; Ogundele *et al*., 2024]. Mengash [2020] employed linear regression to support decision-making in university admission systems, while Ogundele *et al*. [2024] showed that regression models can effectively predict academic performance based on historical data.

While these machine learning models have demonstrated promising accuracy and effectiveness, they often rely primarily on external data, which may not fully capture the specific context of educational institutions and performance in different environments. Moreover, the opacity of more complex machine learning algorithms, such as Random Forests and Neural Networks, complicates the understanding of how input data is processed and decisions are made. This lack of transparency can lead to a trial-and-error approach in tuning models, increasing the complexity and time to develop effective solutions.

## 3.2 Information Visualization and Educational Visual Analytics

Information Visualization techniques transform abstract and complex data into clear and intuitive visual representations, making it easier to understand and analyze [Munzner, 2014]. This approach is particularly powerful when combined with Visual Analytics, which integrates visualization techniques with machine learning to facilitate interactive exploration and decision-making on complex data sets [Keim *et al*., 2008]. In this context, multidimensional projection techniques employs dimensionality reduction approaches to project high-dimensional data into a 2D/3D projection space. Examples of these techniques employ Principal Component Analysis (PCA), Multidimensional Scaling (MDS), and T-Distributed Stochastic Neighbor Embedding (T-SNE) to create interactive layouts, helping to reveal patterns, clusters, and relationships within the data by positioning similar data points close together and dissimilar ones further apart [Jolliffe, 2002; Cox and Cox, 2008; Van der Maaten and Hinton, 2008].

In the educational context, Information Visualization techniques create layouts that allow educators and analysts to visually identify patterns, explore similar patterns, detect outliers, and understand the relationships between various academic and demographic features. This not only facilitates the interpretation of complex data but also improves the communication of insights, assisting the implementation of effective policies and interventions to improve student performance and educational outcomes.

Recent studies have applied these techniques in educational contexts, demonstrating their ability to reveal critical insights into student engagement and performance. Garcia-Zanabria *et al*. [2022] propose the *SDA-Vis*, a visualization system that explores counterfactual scenarios to understand student dropout, using data from more than 11,000 students in Latin American universities. The system integrates machine learning techniques to simulate how changes in student patterns—such as attendance, participation, and academic performance—might alter final outcomes in their courses. Zhang *et al*. [2022] also explore counterfactual scenarios in a system called *DropoutVis*, which employs a CNN-LSTM model to predict the risk of dropout in massive open online courses (MOOCs) with 79,186 students. The system presents five main visualizations to analyze engagement-related characteristics that contribute to dropout, such as temporal factors and perturbation effects, allowing targeted interventions.

A variety of interactive layouts are used to visually explore strategic information from educational data. Deng *et al*. [2019] present the system called *PerformanceVis* to analyze the performance of a chemistry course with 949 students. The system employs several layouts including *Sankey* diagrams and Parallel Coordinates to track grade variations, identify performance trends, and explore correlations between exam questions and homework assignments. Etemadpour *et al*. [2020] combine data visualization and machine learning to evaluate educational from two data repositories. The system creates predictive models to estimate future grades based on previous performance and attendance and offers interactive layouts that explore the relationship between grades, attendance, and additional features such as gender and parents' educational level.

Tsung *et al*. [2022] propose *BlockLens*, a system designed to analyze coding exercises of elementary and middle school students in block-based programming. It presents visualizations for exercise selection, student distribution, path summaries, and sequences, helping instructors understand student engagement and performance. Similarly, Goulden *et al*. [2019] present *CCVis* to explore clickstream data and investigate student behavior patterns. *CCVis* uses higher-order networks and structural identity classification to analyze behavior patterns and relate them to student performance in an introductory course with over two thousand students. The system offers four coordinated views: behavior pattern, behavior analysis, clickstream comparison, and grade distribution. Chen *et al*. [2024] developed *StuGPTViz*, to help instructors examine interactions between students and ChatGPT in data visualization courses. *StuGPTViz* collects and categorizes conversations based on cognitive levels and response quality, providing detailed visualizations of interaction patterns. This system facilitates multi-level analyses, allowing instructors to gain insights into students' use of ChatGPT, identify areas for pedagogical intervention, and evaluate the effectiveness of ChatGPT in education. Validated through expert interviews and case studies, *StuGPTViz* demonstrates its potential

to enhance conversation analysis and support personalized education with AI.

The aforementioned systems highlight the value of combining data visualization and machine learning to uncover key insights about student engagement, performance, and risk factors. However, many existing systems focus on specific aspects of educational analysis, such as dropout prediction or performance tracking, without integrating a comprehensive view of the curriculum or allowing for detailed analysis of student engagement in individual topics. Furthermore, many existing approaches rely primarily on external data. Even among those that use internal data, it is often handled in a summarized manner or merely used to complement visualizations based on external datasets. *LearnVis* aims to provide a detailed view of internal data focusing in the distribution of attendance/absence in a module, to enhance the comprehension of how student engagement with specific topics within each module impacts overall performance. This topic-level visual analysis bridges the gap between general performance trends and detailed curricular insights, supporting a more accurate understanding of the factors behind student success or failure.

# 4 System Requirements

The main objective of *LearnVis* is to assist educational experts in exploring students performance in course modules, and to comprehend how their behavior is related to their success or failure in these modules. We first conducted an analysis of existing Data Visualization literature in visual strategies focused on educational analysis to identify research gaps [Mandinach and Abrams, 2022; Martins *et al*., 2019]. We then held iterative discussions with domain experts, including two coordinators from the Computer Science programs at the university from which the data was collected and two lecturers from those programs with experience in educational management tasks, who are also co-authors of this work. These meetings were crucial to identify the most critical student analysis tasks, serving as the foundation to guide the requirements definition, as well as to conduct the experiments to evaluate our proposal. From these discussions, we were able to identify the following requirements that our system should address:

**R1-Utilize available internal data from student records to analyze their performance in course modules.** The system should integrate and use the internal academic data (grades, attendance, attempts, etc.) available at the institution to provide an overview of student performance throughout the course modules. This data serves as the foundation for all subsequent analyses;

**R2-Explore and evaluate the impact of the curriculum structure on student performance.** The system should provide the analysis of the relationship between the course structure and student performance. This includes the possibility of identifying modules that positively or negatively affect academic progress;

**R3-Identify critical module sequences and/or topics for student success or failure.** The system should be able to provide the analysis of the students' progression through the course as well as to provide information, when possible, about the impact of specific topic enrollment sequence in the success or failure in subsequent topics;

**R4-Provide the analysis of student performance in multiple attempts of the same module.** The system should provide a detailed analysis of multiples enrollments on a specific module, allowing the analysis of different behavior in these attempts and the impact of them on their performances;

**R5-Provide complementary information for performance analysis**. The system should offer additional insights that enable users to access specific data about individual students or groups of students. This includes detailed performance metrics and attendance patterns in particular topics, supporting a deeper understanding of factors that influence student outcomes.

These requirements were used to guide the distillment of the following analytical tasks to guide the framework design:

**T1-Analyze the performance of individuals and student groups in course modules, considering their attendance and obtained grades.** The system shall provide a detailed analysis of the performance of individual students or groups, taking into account their grades and class attendance. This will help identify performance patterns and areas that need attention (**R1, R4, R5**);

**T2-Track students' multiple attempts in course modules.** The system shall allow to track students' multiple attempts in specific modules, showing how their performance evolved in each attempt, facilitating the analysis of progress and challenges (**R4**);

**T3-Investigate how module topics impact students' performance in these modules.** It is important for educators to investigate how the topics covered in each module impact students' performance, helping to identify which topics are critical for academic success or failure (**R2, R3**);

**T4-Identification of at-risk groups.** Users must be able to identify students who share similar behavioral or academic patterns, facilitating the creation of intervention strategies targeted at at-risk groups (**R1, R3**).

# 5 Data Understanding and Preprocessing

We utilized a data repository provided by the Information Technology department at the UFU, which anonymized all students' personal information. The dataset includes records from 1,490 students enrolled in the Computer Information Systems course at the Faculty of Computing at UFU, who entered their courses between 2009 and 2019. All the features represent **internal data** related to students' academic life, including the modules they enrolled in, their grades in each module (on a scale from 0 to 100), attendance per class, and the topics covered in the modules. According to the General Rules of the Undergraduate Council (CONGRAD) at the UFU, in Chapter II – On Evaluation, Art. 127, for a student to succeed, it is necessary to achieve at least 60 academic performance points and 75% attendance in academic activities [Conselho de Graduação da Universidade Federal de Uberlândia, 2022]. Students are considered as **failed** if they scored less than 60 or as **dropped out** if they did not complete the module. Given the requirements of our analysis, several preprocessing steps were applied to the original data, which

are described below:

**Data Collection:** The dataset covers a 10-year period from 2009 to 2019, containing information such as modules, grades, attendance, and curriculum topics. The data was initially structured as shown in Table 1, and the preprocessing is described as following.

**Data Cleaning:** The data cleaning process involved addressing missing values, particularly in cases where the instructor did not specify the topic covered in a class (TOPIC_DESCRIPTION). To fill these gaps, we leveraged data from the 20 available semesters (spanning 10 years). Since the topics were recorded in free-text fields by different instructors, there was no uniformity in the descriptions, which posed a challenge in standardizing the content. We first created a standardized topic table for each module, based on information from the curriculum structure of the module, and the various topic descriptions informed by instructors were then manually mapped to the corresponding standardized topic. In cases where the topic information was missing, we inferred the standardized topic using data from previous semesters, considering the typical order in which topics were taught. This ensured that missing data were aligned with the general curriculum structure of the module.

**Features Engineering:** To enhance the dataset and support the visualizations and analyses required by *LearnVis*, several new features were engineered. `TOPIC_ID` and `TOPIC_NAME` replaced the original `TOPIC_DESCRIPTION` field, providing a standardized way to reference topics covered in each module, for all 40 modules. In our experiments, we manually processed data from only the first 20 modules of the Information Systems course. We intend to perform the analysis of the remaining modules as a future work. The `ATTENDANCE_FLAG` features was created to standardize attendance tracking, where 0 represents absence, 1 represents presence, and 2 indicates that the topic was not covered. This flag was derived from the combination of `LECTURE_DATE` and `TOPIC_ID`,allowing to track whether students attended each specific class session and topic. Additionally, the `REPEAT_COUNT` features was introduced to track how many times a student attempted the same module. By analyzing the number of occurrences of a student's `ENROLLMENT_ID` in each module across different semesters, this features provided insights into student multiple module attempts.

# 6 *LearnVis*: System Description

In this section, we present a new visual analysis system called *LearnVis*, which assists instructors in exploring, analyzing, and understanding student behavior patterns. We refer to student behavior as the representation of their sequence of attendance and absence across all topics covered in a module. We consider a student to have dropped out from a module when there is no record of approval for this student within all his/her attempts in this module. Figure 1 shows the main interface of the system considering a specific module, with no students selection performed. Users are able to use the filters located at the top of the interface to perform the analysis for specific modules and classes, as well as to choose students from specific absence and ranges (Figure 1 - F). Each layout is detailed as follows.

**(A) Student Behavior View (T1, T2, T4):** the objective of this view is to explore the behavior of groups of students, regarding academic performance, as well as students with specific behavior. We applied a t-SNE dimensionality reduction technique to map each student to a 2D layout. The resulting layout ideally groups students with similar behavior, and position students with non-similar behavior far from each other. Our student representation results in a potential high-dimensional data. We thus decided to use t-SNE due to its ability to capture non-linear relationships among instances, adequately unveiling the complex distributions and patterns in this space. This view is displayed as a scatterplot, in which each circle in the layout represents an individual student. Students who failed are shown in red, while those who succeeded are shown in green. The layout considers all semesters in which the module was offered, allowing the same student to appear multiple times, one for each attempt, potentially placed in different positions of the layout, depending on his/her behavior. Users can select individual or groups of students for more detailed analysis. When a group is selected, all other views are automatically updated to reflect this selection. The selected students are highlighted in brown, whereas the remaining ones are colored gray, as shown in Figure 2. Additionally, it is possible to identify students who performed multiple attempts in a module. In a selection containing one student in this situation, the corresponding circles will be uniformly colored and numbered according to the number of attempts. If a student succeeded in one of the attempts, the correspondent circle is marked with a star glyph, indicating their success. A stacked bar chart complements the analysis by presenting the proportion of students who succeeded (in green) and failed (in red). When a user selects a specific group of students, it is updated to reflect the success and failure rates within that selection. If no selection is made, it displays the overall distribution for the entire dataset. This stacked bar chart provides a quick and clear overview of overall performance distribution, serving as a summary of the success and failure rates among students;

**(B) Grades Distribution View (T1, T4):** represents the distribution of students' grades in the module. A light purple curve represents the overall grade distribution in the module, and when a group of students is selected, their grade distribution is displayed on top of the overall distribution chart in dark purple. The vertical dark purple lines indicate the average grades of the selected group, while the dotted line in light purple represents the overall average grades of all students in the module. A rectangular shaded area is also drawn around each vertical line to represent the standard deviation of the grades. The shaded region in light purple corresponds to the overall standard deviation for all students in the module, while the dark purple shaded area reflects the standard deviation within the selected group. This visual encoding helps to highlight the variability of grades and supports comparison between the selected group and the overall population. When hovering the mouse over the vertical lines, a tooltip appears, displaying the exact value of the average grades for both the selected group and the overall student population;

**(C) Absences vs. Grades View (T1, T4):** displays a scatterplot where students' grades are plotted on the horizontal

**Table 1.** Description of the features used in *LearnVis*.

| Type | Feature | Description | Example 01 | Example 02 |
|---|---|---|---|---|
| Numeric | CLASS_ID | Class identification | 13 | 11 |
| | ENROLLMENT_ID | Masked person identifier | 111 | 123 |
| | NUMBER_OF_ABSENCES | Number of absences | 23 | 12 |
| | NUMBER_OF_YEAR | The year the module was taught | 2021 | 2022 |
| | **ATTENDANCE_FLAG** | Attendance identifier for topics | 1 | 3 |
| | **REPEAT_COUNT** | Number of repetitions in the module | 2 | 0 |
| | **TOPIC_ID** | Topic identifier | 8 | 5 |
| Nominal | FINAL_STATUS | Student's final status (Succeeded, Failed or Dropped Out) | Failed | Succeeded |
| | MODULE_CODE | Code of module | GSI002 | GSI019 |
| | MODULE_NAME | Name of module | Introduction to Computer Programming | Information Systems Career |
| | PERIOD_NAME | Course period description | First | Third |
| | TOPIC_DESCRIPTION | Topic description | Function | Switch |
| | **TOPIC_NAME** | Module topic based on the curriculum guide | Program Modularization | Information Technology Career |
| Scalar | FINAL_GRADE | Student's final grade (0 to 100) | 43 | 78 |
| Temporal | LECTURE_DATE | Lecture date | 02/04/2021 | 02/08/2022 |

axis and the number of absences on the vertical axis. The purpose of this layout is to analyze the correlation between the number of absences and final grades, helping to determine whether the final grade is influenced by the number of absences. The occurrence of one or several students sharing the same grade and absences are mapped to a circle, whose color maps success (green) or failure (red), and whose color intensity maps the number of students in this specific situation. Higher intensities are mapped to higher occurrences, and vice versa. When hovering the mouse over a circle, a tooltip displays the number of students represented by the circle, their corresponding grades, and the number of absences;

**(D) Student Frequency View (T1, T3, T4):** represents student attendance on module topic, as well as if that topic was taught in a specific semester. Each row corresponds to a student, whose enrollment number is colored in green if that student succeeded in that module, in red if he/she failed. Each column represents the topics covered on each semester for the module, colored in dark blue if the student attended that topic, in light blue if the student was absent, and in gray if that topic was not taught for a specific class/semester. The last column indicates each student's class using a unique color. The system offers two versions of this view: one focused on each individual student and another that summarizes the information by class. In the class view, each row represents a class, the columns correspond to the topics covered, and each rectangle displays the proportion of students present (in dark blue) versus absent (in light blue) for that topic. When hovering over the rectangles, a tooltip shows the exact proportion value, and in both versions, hovering over the column numbers displays a tooltip with a detailed description of the topic;

**(E) Enrollment Progress View (T1, T2, T4):** illustrates the students' trajectory in terms of modules enrollment sequence. Each bar height maps the proportion of students in each situation. The first bar represents in blue all the students enrolled in the module, and subsequent bars represent in green the students who succeeded the module, in red those who failed, and in orange those who dropped out. It is possible to move the bars to reposition them as needed, and when hovering over the transitions between bars, a tooltip shows the percentage of students that moved from one state to another. Zooming can be applied to individual bars to increase the clarity of those with shorter lengths.

**Implementation Details:** *LearnVis* was implemented using HTML, CSS, and JavaScript. To perform dimensionality reduction, we used the *Manifold*[1] Python library, especifically the t-SNE (t-Distributed Stochastic Neighbor Embedding) technique. Data querying, cleaning, and filtering were carried out using *SQL*[2] and *Pandas*[3] Python library. Finally, all the layouts were developed using the *D3.js*[4] library. This system implementation, along with a user manual and a demonstration video, is available at: `https://github.com/angelicago/LearnVis_System.git`.

# 7 Case Studies

This section presents the application of *LearnVis* to a set of case studies representing distinct scenarios on our data repository. For one of these cases, a demonstration video was produced and is available in the repository. During the meetings described in Section 4, the experts reported the

[1] `https://scikit-learn.org/stable/modules/manifold.html`
[2] `https://learn.microsoft.com/en-us/sql/?view=sql-server-ver17`
[3] `https://pandas.pydata.org/`
[4] `https://d3js.org/`

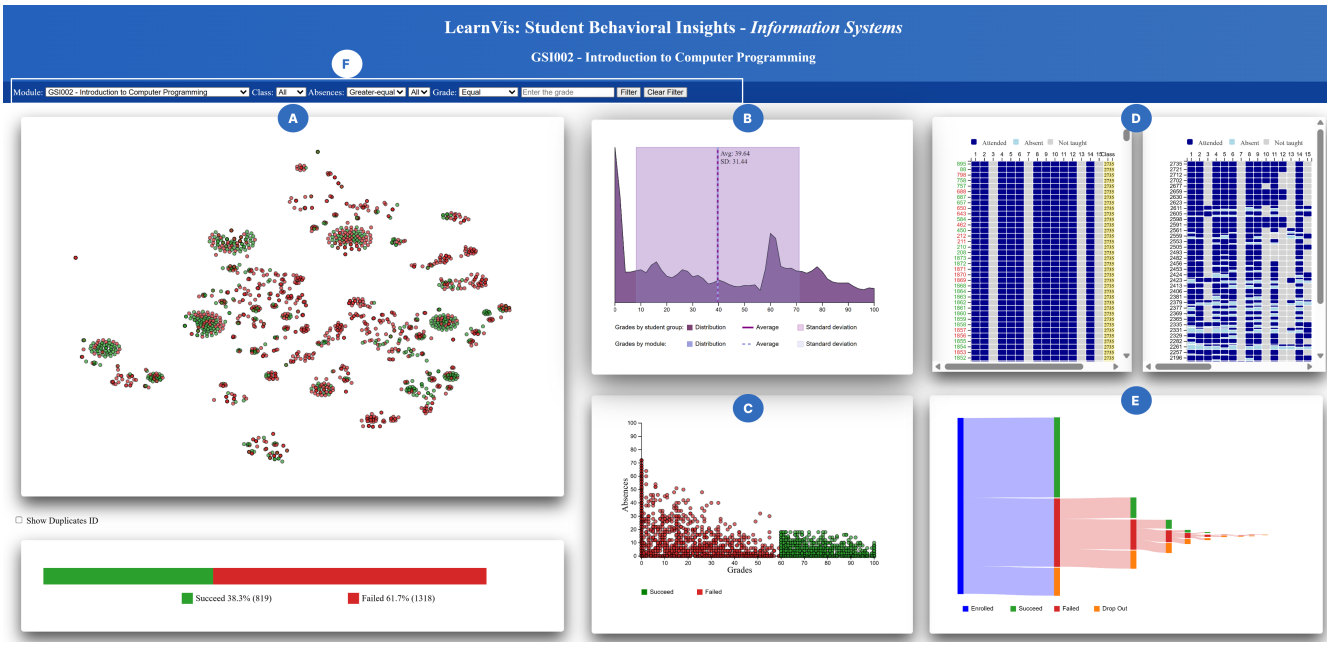


**Figure 1.** *LearnVis* main interface showing all coordinated views: **A** - Student Behavior View, **B** - Grades Distribution View, **C** - Absences vs. Grades View, **D** - Student Frequency View, **E** - Enrollment Progress View, and **F** - Filters.

importance of performing analyses on modules with high retention rates. They also suggested the analysis of different stages of the students' academic journeys, to explore how students behave in initial modules - when they are faced with a new study university's routine, and subsequent modules - in which they are already habituated with this routine. Finally, they reported how important it is to identify topics in the modules that are more challenging in terms of approval. All these reports helped us to define a set of analyses that were used to validate our proposal, which we present in the following sections.

## 7.1 GSI002 - Introduction to Computer Programming

This module is offered in the first period of the course, when students are still getting adapted to the academic routine and have not yet developed full maturity in their learning process, and historically presents a high failure rate. Figure 1 presents the layout produced from this module. The layout confirms the high number of failure, in this case 61.7% (Figure 1 - A). Figure 1 - B shows that the average student grade is 39.6, which is approximately 20 below the minimum approval score of 60.0, indicating substantial difficulties in succeed the course content. One can also notice, by looking at the Enrollment Progress View (Figure 1 - E), that many students perform several attempts on this module, and that a significative number of students drops out.

By analyzing a group of students who failed (Figure 2), it becomes clear that all of them obtained very low grades, as shown in the Grades Distribution View (Figure 3). Most of these students scored between 0 and 10, with an average grade of 2.28. All of them were enrolled in the same class, and the majority did not attend the topics covered in this module (Figure 4). These students failed on their first attempt, and despite multiple attempts, they eventually dropped out (Figure 5).

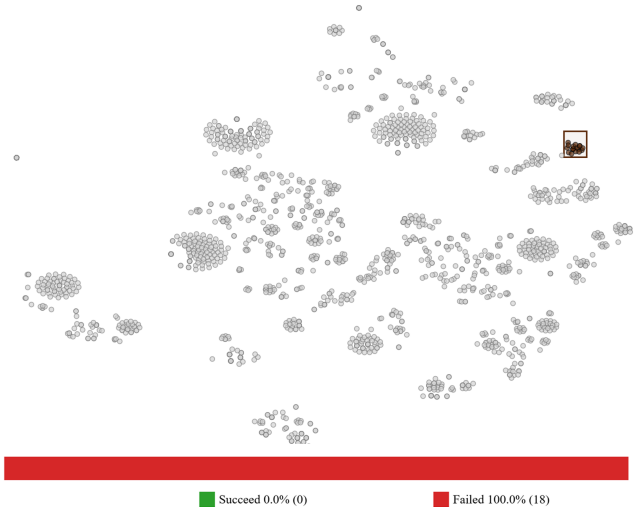


**Figure 2.** Selection of a group on Student Behavior View composed only by students who failed.

There are other students with similarly low grades — that is, grades insufficient for approval, with the majority of them having scored zero (Figure 6), which were placed in different groups. The difference between these students and the initially selected (Figure 4) group is that, in these groups, students attended at least one common topic, such as *Variables, Data type and Program Basic Structure* (2), *Arrays and Matrices* (6), *Program Modularization* (8), and *Final Exams, Final Activities* (15).

It is noticeable that most students in these groups dropped out after a few attempts. Only one student (ID 3169) succeeded in the fourth attempt, in which he/she attended more topics compared to previous attempts (Figure 7). This attendance pattern, indicating different engagement behavior, justified the creation of these new groups in the layout.

This insight provided by the layout may represent an im-

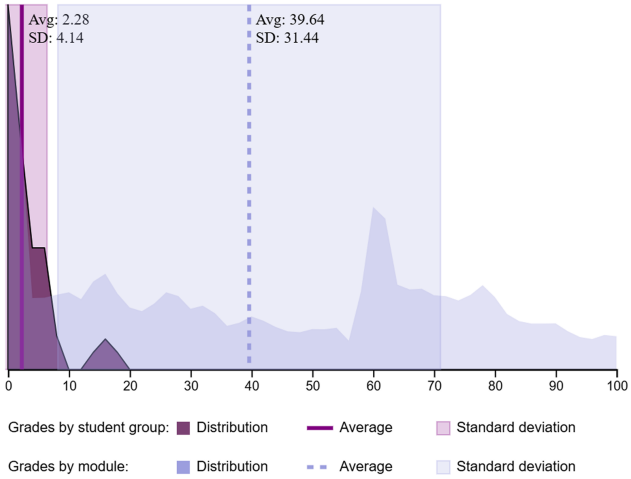


**Figure 3.** Selection of a group on Grades Distribution View composed only by students who failed.

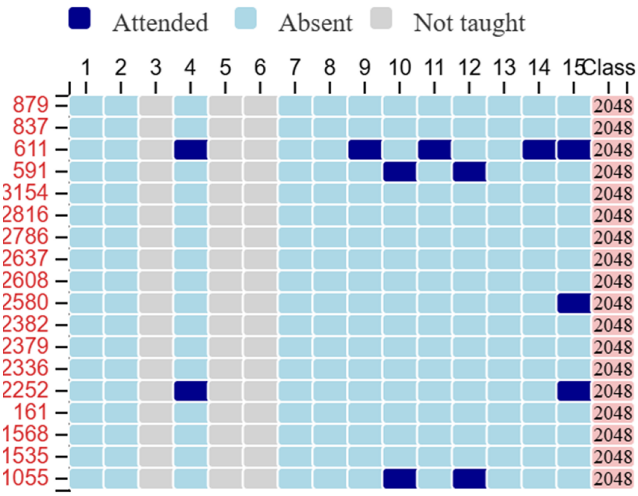


**Figure 4.** Selection of a group on Student Frequency View composed only by students who failed.

portant opportunity for educational experts. By identifying students with low grade and exploring their attendance patterns in a visual and easily interpretable way, educators are able to implement targeted interventions in the upcoming semesters. Grouping students with similar behavior allows experts to focus on those at risk, potentially designing specific strategies, such as personalized tutoring or efforts to increase engagement. The layout's ability to reveal these patterns may thus contribute significantly to decision making.

Analyzing another selected group (Figure 8) who were present in all the topics taught (Figure 9), it can be noticed that 57.9% of students succeeded. The Enrollment Progress View shows that most of the students who succeeded did so on their first attempt (Figure 10). It also shows that few attempts were necessary for succeeding on this module, and that few of the students dropped out. Their grades range from 20.0 to 100.0, and many students scored between 70.0 and 80.0 (Figure 11), indicating a higher average grade compared to the overall one.

Four students from this group succeeded in subsequent attempts, as shown in Figure 12. One notices that all of them attended all topics in each attempt (Figures 13 (a) to (d)), except for student 1585, who missed the *Program Modularization* topic (8) but still managed to succeed. In the group where these students were before succeeding (Figure 8), the approval rate was below 60.0%. It is interesting to observe that the students migrated to groups with even higher approval rates in subsequent attempts (Figure 12). For example, students 2540 and 2548 succeeded in a group with a 68.0% approval rate, student 1585 in a group with 69.2%, and student 2499 in a group with 77.2%. This repositioning to groups with higher engagement and attendance suggests that a more collaborative environment, combined with more consistent class participation, may have been a decisive factor in these students' success.

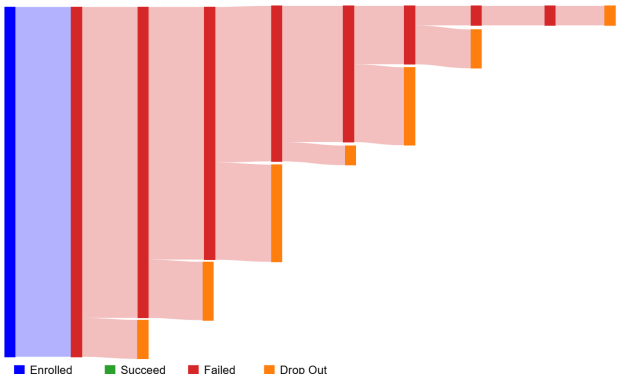


**Figure 5.** Selection of a group on Enrollment Progress View composed only by students who failed.

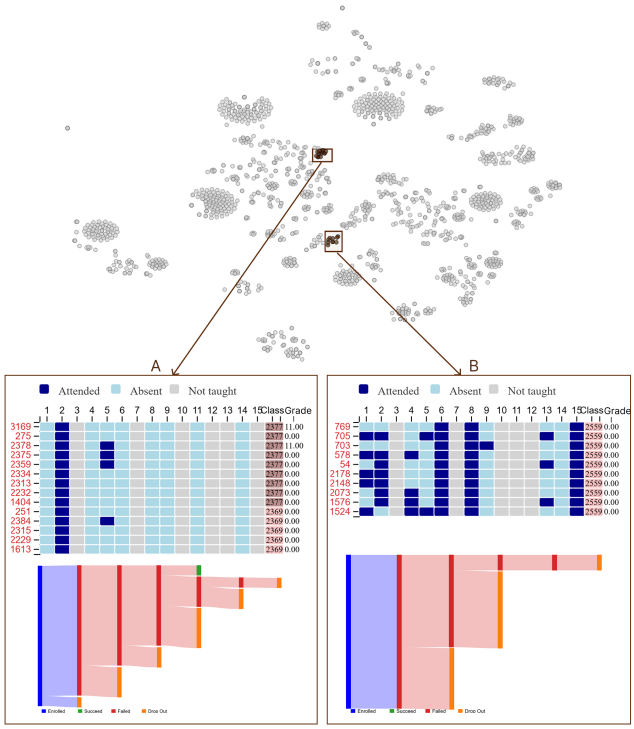


**Figure 6.** Group of students with low grades.

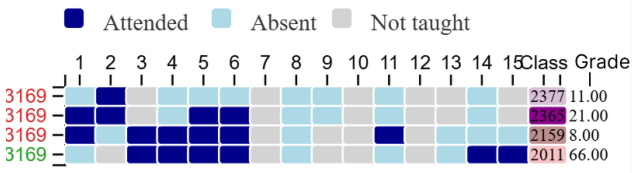


**Figure 7.** Student 3169's behavior across module attempts.

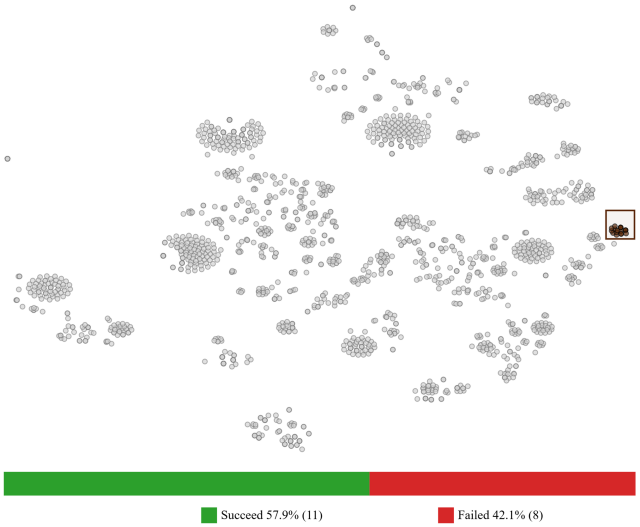


**Figure 8.** Selection of a group containing students that succeeded and students that failed - Student Behavior View.

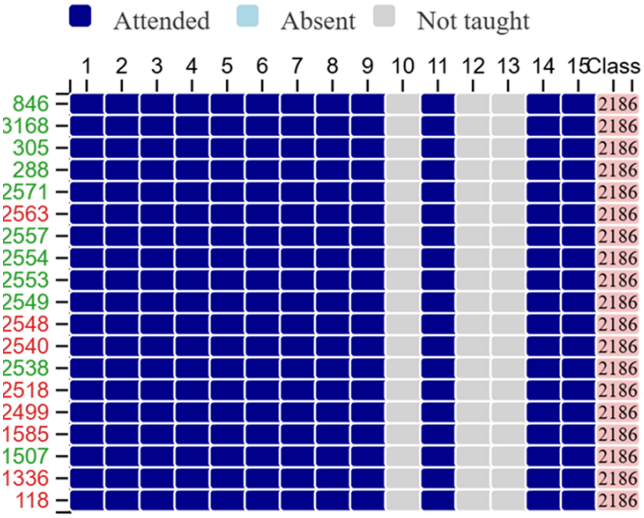


**Figure 9.** Selection of a group containing students that succeeded and students that failed - Student Frequency View.

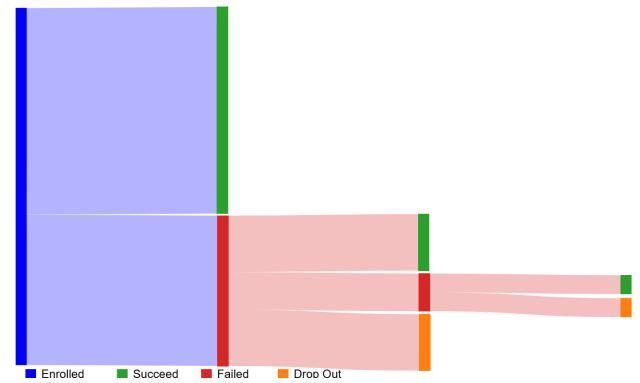


**Figure 10.** Selection of a group containing students that succeeded and students that failed - Enrollment Progress View.

## 7.2 GSI009 - Information Systems Career

This module is offered in the second period of the course and has a history of more students succeeding than failing. Figure 14 shows the layout produced from this module showing a higher proportion of students who succeeded (78.1%) compared to those who failed (21.9%), with an average grade of approximately 69.2 (Figure 15).

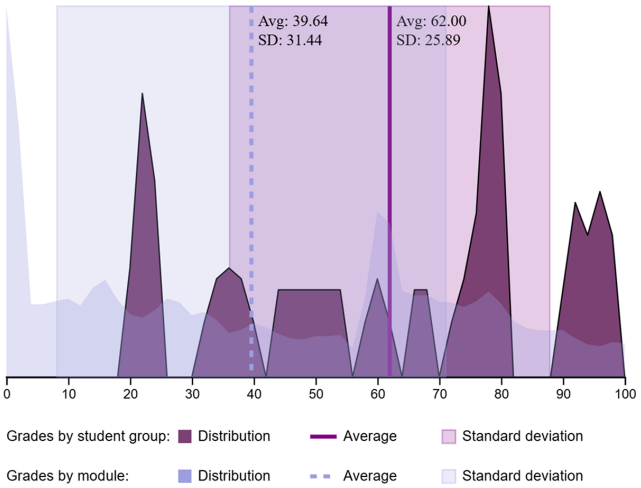


**Figure 11.** Selection of a group containing students that succeeded and students that failed - Grades Distribution View.

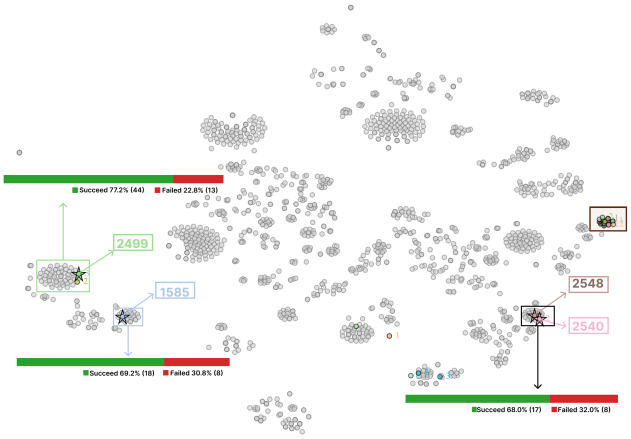


**Figure 12.** Analysis of students who succeeded after multiple attempts.

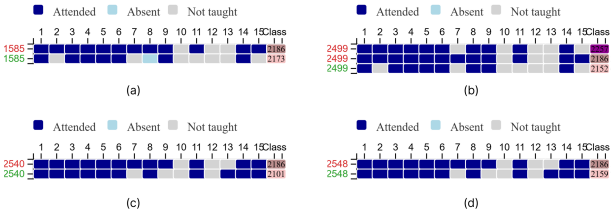


**Figure 13.** Student Frequency View - a) student 1585, b) student 2499, c) student 2540 and d) student 2548.

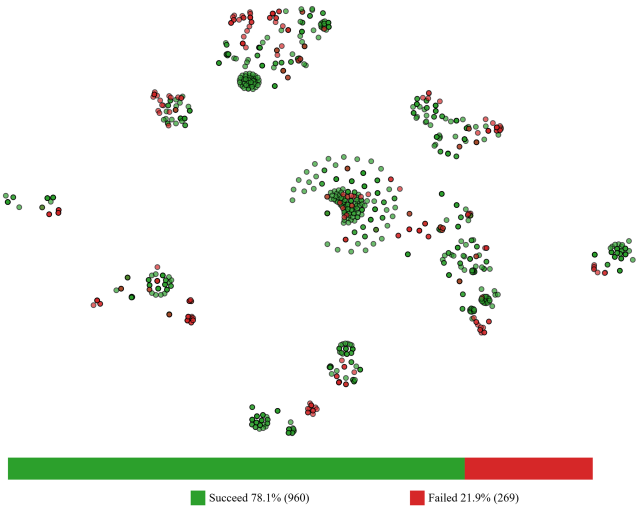


**Figure 14.** Overview of student performance in the GSI009 module - Student Behavior View.

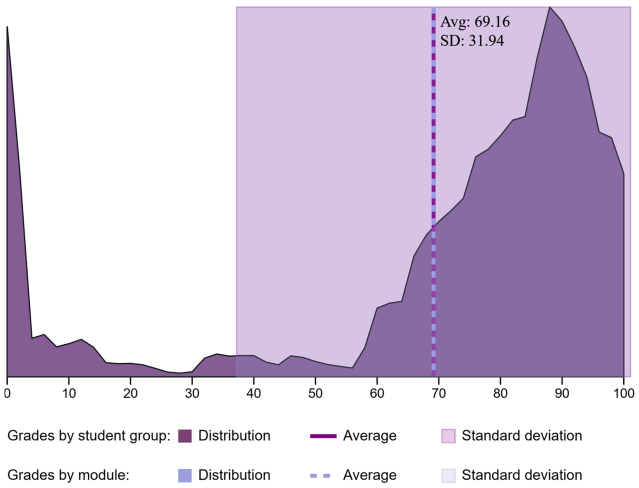


**Figure 15.** Overview of student performance in the GSI009 module - Grades Distribution View.

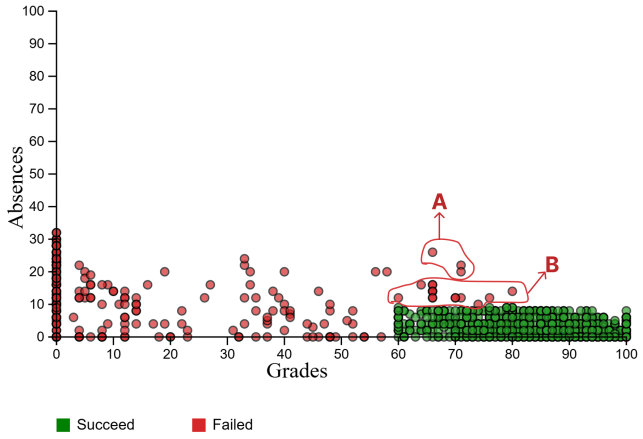


**Figure 17.** Overview of student performance in the GSI009 module - Absences vs. Grades View.

Many students received a grade of 90.0. Most of the students succeeded on their first attempt (Figure 16), and some students, despite receiving approval grades, failed due to insufficient attendance, as marked in A in Figure 17. An interesting case appears in marker B (Figure 17), involving students who had grades high enough to succeed and fewer than 25% absences, yet were still marked as failed. This may indicate an error in the dataset, possibly due to incorrect data entry. This example highlights how the system can be useful in uncovering inconsistencies in the records. This unusual pattern highlights the layout's ability to reveal crucial insights that could easily go unnoticed. An educational expert could further investigate these cases, exploring possible causes, such as internal and external factors influencing student participation, as well as how attendance policies are being applied. The analyses of this layout may lead to the development of effective strategies to reduce such occurrences in the future.

The analysis of a group of students who failed (Figure 18) shows an average grade of approximately 5.7 (Figure 19). This group had a significant number of absences in almost all topics (Figure 20), and most of the students dropped out the module (Figure 21), except for one student (ID 2449) (Figure 22). This student failed the first attempt receiving a grade of 6.0 (Figure 23) and missed the topics *Ethics in Computer Science* (2), *Intellectual Property* (4), *Information Technology Professional Profile* (7), and *Information Technology Career* (10). In the second attempt the student received a grade of 90.0, and missed only the topic *Ethics in Computer Science* (2). The layout suggests that students who attended these topics may have been more likely to succeed, potentially guiding further analysis by the course coordinator to identify which topics are most critical to academic success.

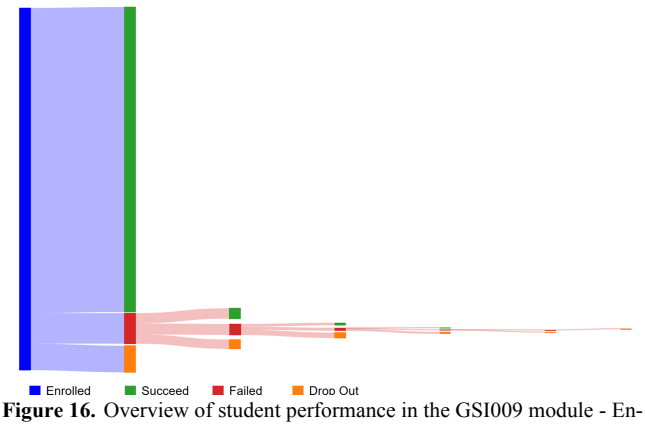


**Figure 16.** Overview of student performance in the GSI009 module - Enrollment Progress View.

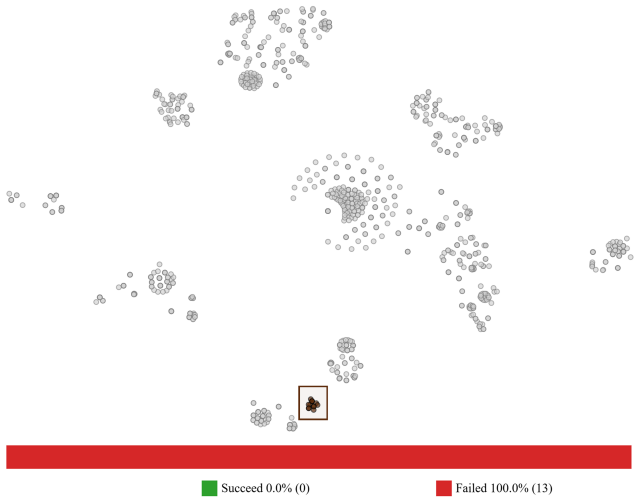


**Figure 18.** Selection composed only by students who failed - Student Behavior View.

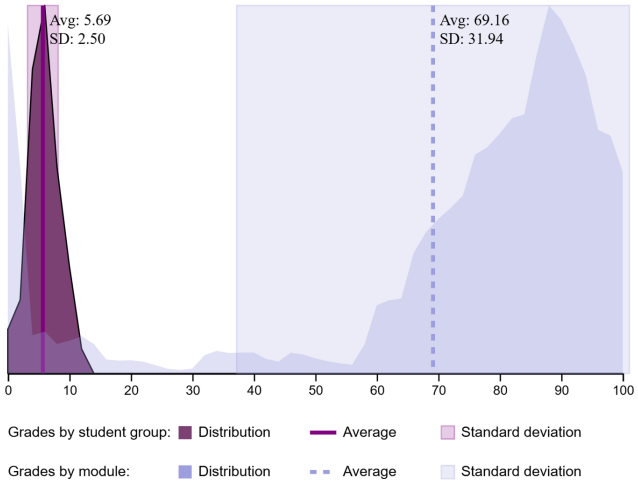


**Figure 19.** Selection composed only by students who failed - Grades Distribution View.

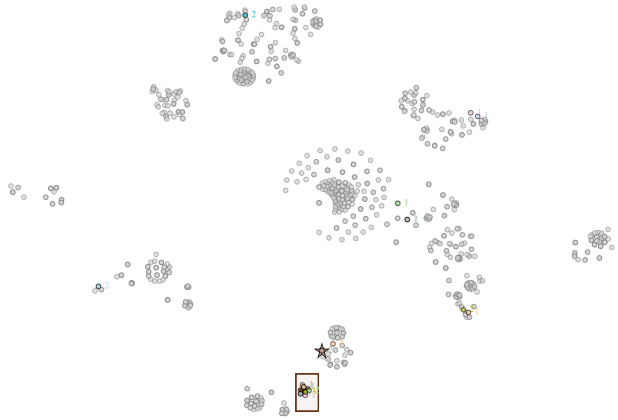

**Figure 22.** Analysis of Student ID 2449 - Student Behavior View.

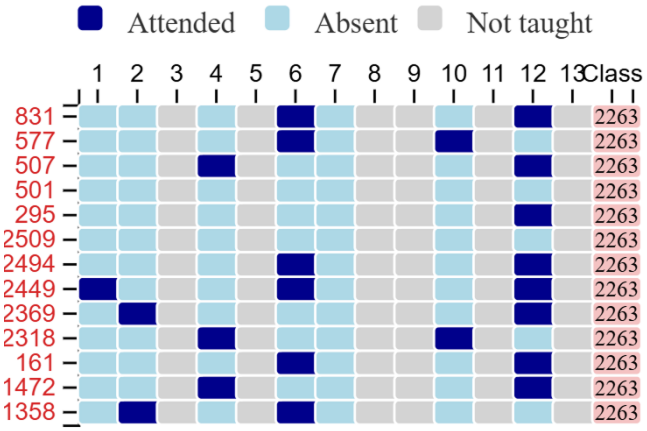


**Figure 20.** Selection composed only by students who failed - Student Frequency View.

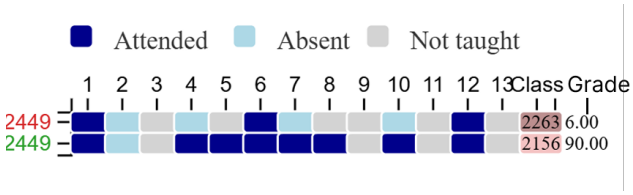


**Figure 23.** Analysis of Student ID 2449 - Student Frequency View.

Selecting a group of students who succeeded (Figure 24), it is possible to notice that their grades are of 89.4 (Figure 25), while the overall average is of 69.1. All of them attended all topics (Figure 26). Most succeeded on the first attempt (Figure 27), except for one student (ID 843), who succeeded on the second attempt (Figure 28). This student missed most of the classes in his/her first failed attempt, obtaining a grade of 0.0. He/she then received a grade of 72.0 on his/her second attempt, attending all the topics (Figure 29).

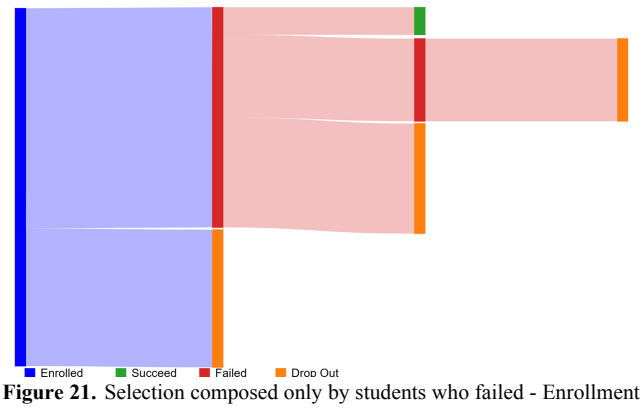


**Figure 21.** Selection composed only by students who failed - Enrollment Progress View.

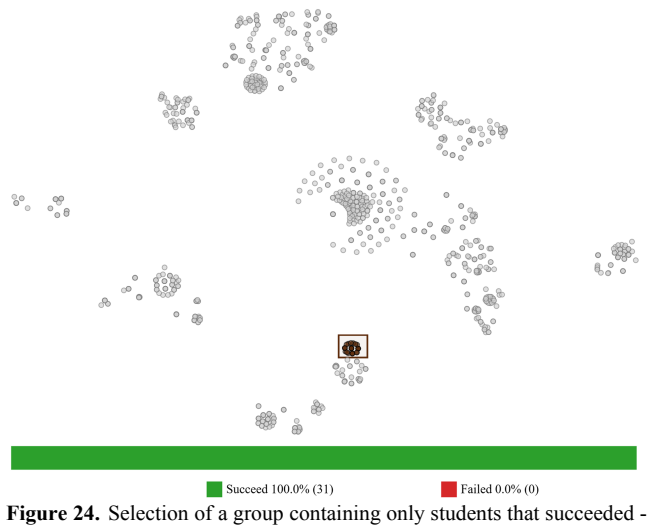


**Figure 24.** Selection of a group containing only students that succeeded - Student Behavior View.

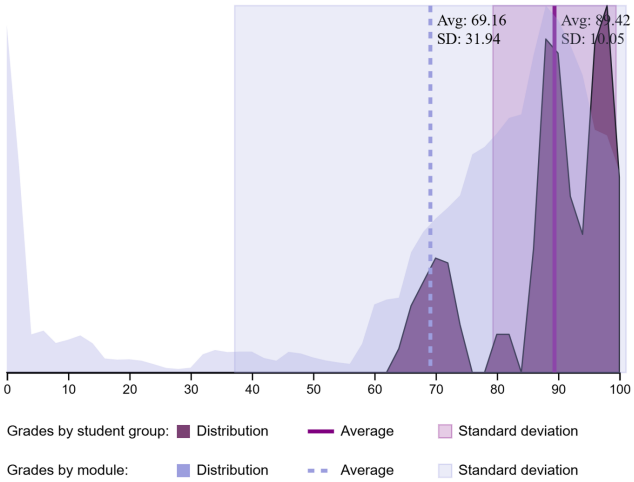


**Figure 25.** Selection of a group containing only students that succeeded - Grades Distribution View.

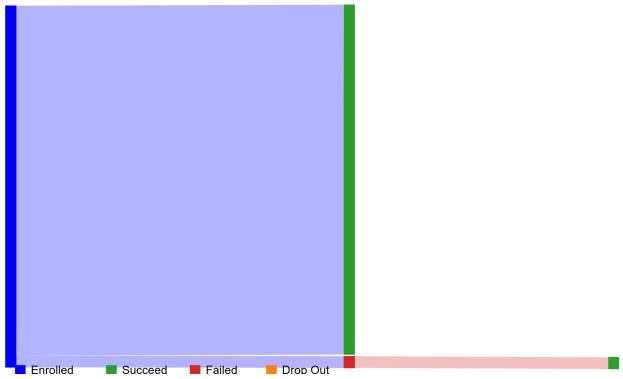


**Figure 27.** Selection of a group containing only students that succeeded - Enrollment Progress View.

**Figure 28.** Analysis of Student ID 843 - Student Behavior View.

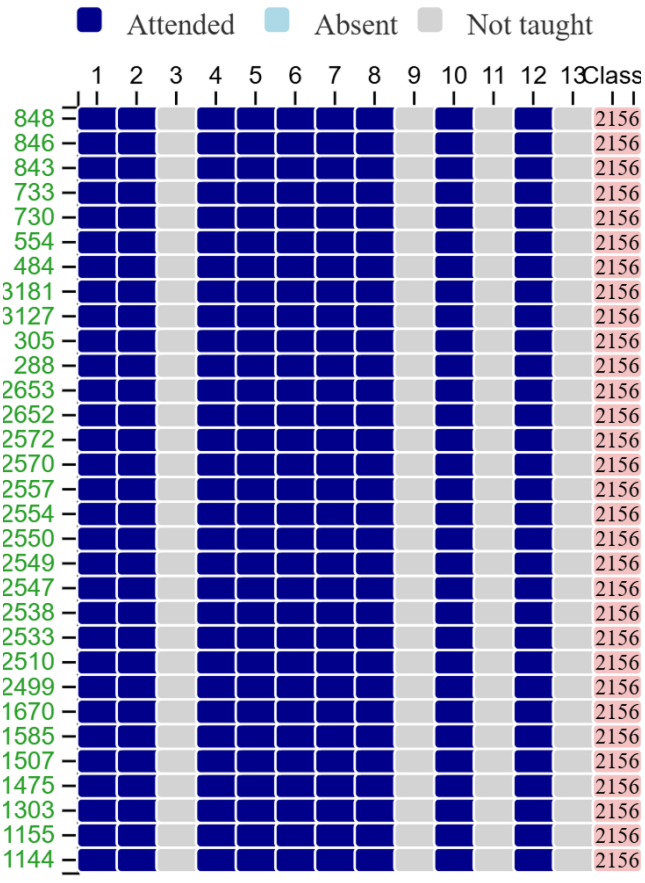


**Figure 26.** Selection of a group containing only students that succeeded - Student Frequency View.

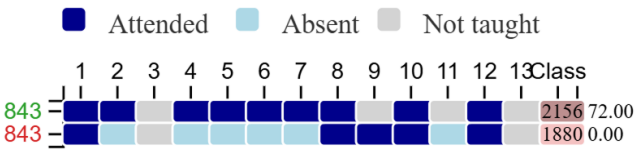


**Figure 29.** Analysis of Student ID 843 - Student Frequency View.

The analysis of this module suggests that students do not face significant difficulties on succeeding. The layout is in conformity with the students feelings about this module, in terms of approval challenge. The layout highlights that students with regular attendance tend to perform better, providing insights that may support further exploration of attendance-related trends in this module.

## 7.3 GSI019 - Web Programming Module

This module is offered in the fourth period of the course and has a history of more students succeeding than failing. The layout in Figure 30 shows that 58.9% of the students were approved, and that a significant portion of the students obtained a grade between 60.0 and 90.0 (Figure 31), but one also notices a considerable number of students which obtained a grade of 0.0. Figure 32 shows that students with more absences tend to have lower grades, suggesting a possible relationship between attendance and performance that may warrant further investigation. One also observes that most

of the students succeeded on their first attempt (Figure 33). Some of the students who failed required up to four attempts until being approved.

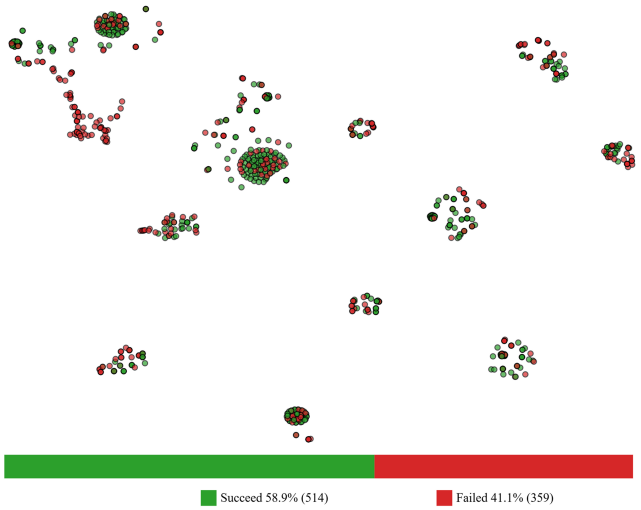


**Figure 30.** Overview of student performance in the GSI019 module - Student Behavior View.

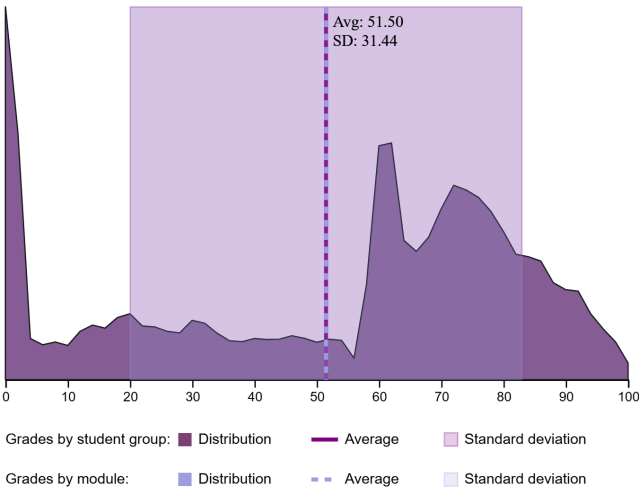


**Figure 31.** Overview of student performance in the GSI019 module - Grades Distribution View.

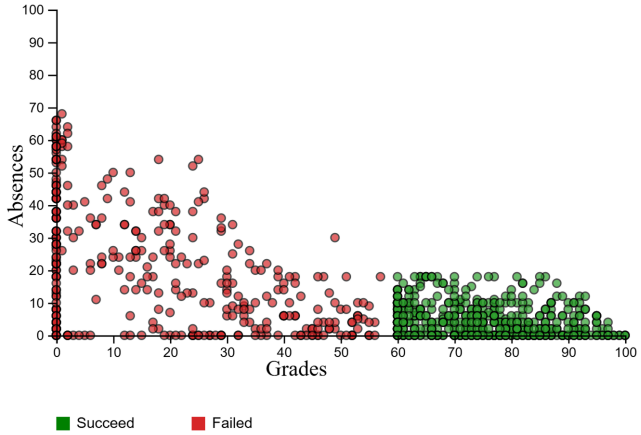


**Figure 32.** Overview of student performance in the GSI019 module - Absences vs. Grades View.

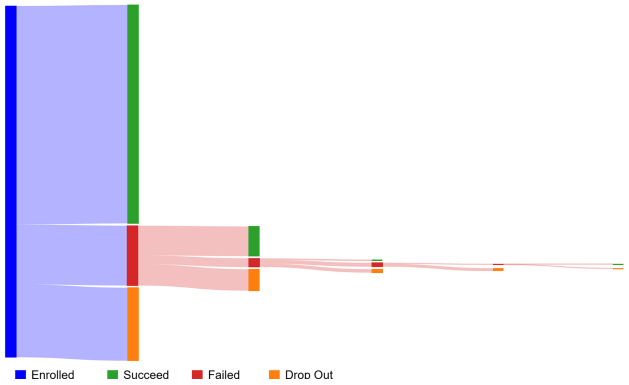


**Figure 33.** Overview of student performance in the GSI019 module - Enrollment Progress View.

When selecting the group highlighted in Figure 34, it is possible to notice that the overall performance in the module was positive and the approval rate was 70.5%. Figure 35 shows a higher concentration of grades between 60 and 90, confirming their approval. All students from this group attended all topics (Figure 36), and most of them were approved in their first attempt (Figure 37).

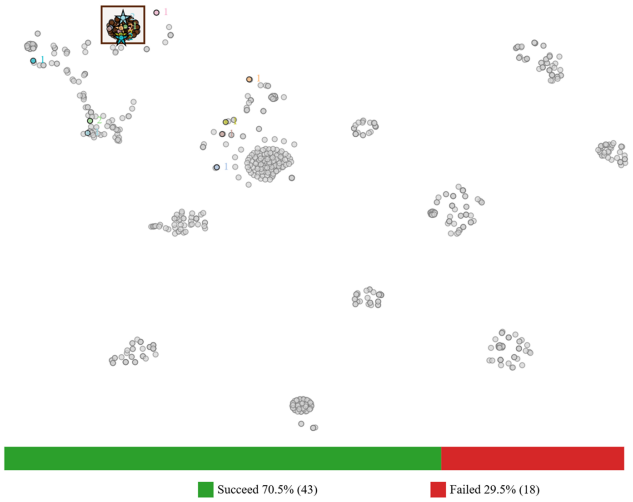


**Figure 34.** Selection of a group of students in the GSI019 module - Student Behavior View.

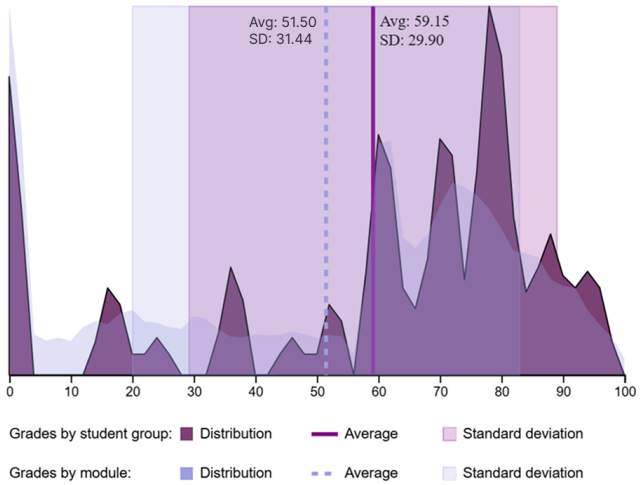


**Figure 35.** Selection of a group of students in the GSI019 module - Grades Distribution View.

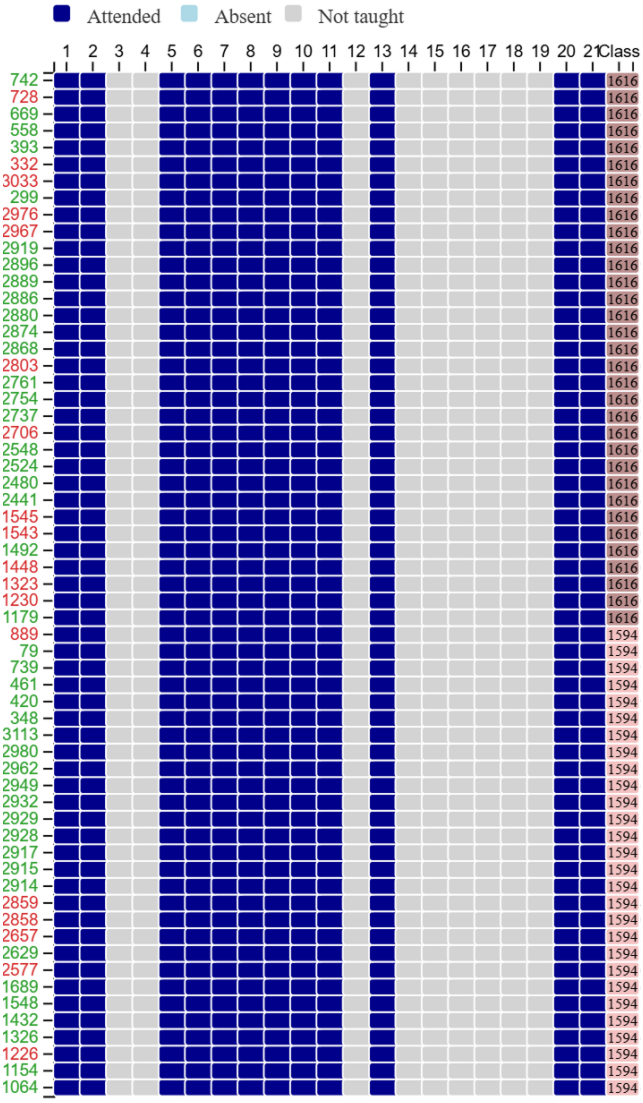


**Figure 36.** Selection of a group of students in the GSI019 module - Student Frequency View.

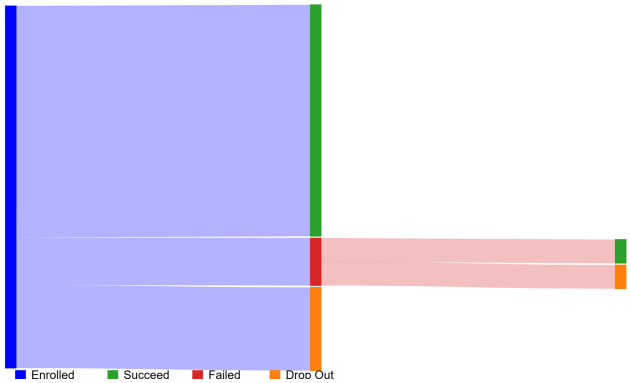


**Figure 37.** Selection of a group of students in the GSI019 module - Enrollment Progress View.

However, there is a group of students who either dropped out or failed, with no more than two attempts at the module. Four of these students succeeded in their second attempt. As shown in Figure 38, they all missed some topics during their first attempt. In contrast, during their second attempt, they regularly attended all topics. This shift in attendance behavior aligns with their improved performance and may help identify

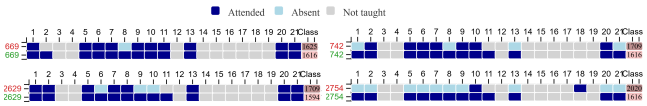


**Figure 38.** Selection of a group of students in the GSI019 module.

participation patterns that influence academic success.

In summary, the analysis of the GSI019 module reveals patterns that suggest a potential relationship between attendance and academic performance. A significant number of students succeeded on their first attempt, particularly those with more consistent attendance. Conversely, several students who initially failed were later approved in subsequent attempts where they attended more classes and missed fewer topics. These patterns may provide valuable insights for educational experts, helping guide further investigation into how attendance could be related to improved outcomes. Such findings may also support the design of policies aimed at fostering student engagement and participation throughout the course.

# 8 Discussion and Limitations

We developed *LearnVis* based on iterative discussions with educational experts, guided by the requirements and analytical tasks described in Section 4. The system was designed to provide a comprehensive understanding of student performance throughout the course, offering insights into how class attendance, engagement with specific topics, and multiple enrollment attempts influence academic outcomes. The case studies demonstrated that the system can support course coordinators and instructors in identifying student behavior patterns and exploring potential points of intervention. During the system development and evaluation process, we identified some limitations, that we discuss in this section.

**Complexity of Data Standardization:** The process of standardizing data, especially when it comes to topic descriptions provided by instructors, can lead to inaccuracies, especially if manual or inferred mapping is required. Inconsistencies in how topics are described across semesters can impact the reliability of the insights drawn from the data. *LearnVis*, however, offers tools that highlight discrepancies in topic descriptions and assist in their standardization, contributing to reduce these inconsistencies. For example, *LearnVis* can be used to identify recurring patterns in topic entries, provide recommendations for aligning descriptions to a consistent standard, and visually represent variations across semesters. These features simplify the standardization process, making it more efficient, while also improving the quality and reliability of the analyses generated from the data;

**Scalability Challenges:** Although *LearnVis* is effective for analyzing educational data in moderately sized datasets, its scalability is limited when handling datasets containing a large historical period, or a large number of students, due to the potential high computational cost to generate the layouts, especially the ones that employ dimensionality reduction techniques, and due to the impact on the interaction with the layouts. In these scenarios, we believe that criterious sampling approaches may reduce the data volume and allow the exploration of data in a

reasonable way. It is also possible to employ interaction functionalities to provide summarization, hierarchization, or even multiple time resolutions, also reducing the amount of displayed information.

**Reliability of Attendance Data:** An additional limitation concerns the reliability and consistency of the attendance data. In some cases, attendance tracking may be inconsistent or imprecise, as not all instructors record absences uniformly. For example, some instructors may log only the total number of absences rather than recording attendance per class or topic. This limitation may affect the accuracy of analyses that rely on attendance patterns to infer student engagement or performance. Therefore, any interpretation involving attendance data should be made with caution and, when possible, validated with complementary information.

# 9 Conclusion

This study presented *LearnVis*, a visualization system designed to support the analysis of student performance across course modules. We illustrated the application of the system using data from students enrolled in the Computer Science program at a Brazilian university. The results demonstrate the effectiveness of our approach in highlighting critical aspects such as academic performance, attendance patterns, differences across multiple attempts, and the correlation between attendance and grades. We believe that *LearnVis* enables educators to quickly identify critical intervention points—such as modules with high failure rates, low engagement levels, or attendance behaviors that negatively affect academic outcomes. As such, it proves to be a valuable tool for guiding the development of precise strategic actions aimed at improving academic performance, increasing student retention, and supporting curriculum planning in a more efficient and informed manner.

Future work include developing a student-focused module that allows students to visualize their performance using data from previous semesters and identify strategies for improvement, thereby encouraging self-regulated learning. Additionally, integrating machine learning techniques into the system could significantly enhance its analytical capabilities. These techniques may provide additional analyses, such as exploring counterfactual scenarios and creating predictive models, offering deeper insights and supporting more effective educational strategies. We also intend to combine external data sources, such as socioeconomic background and participation in extracurricular activities with internal academic records to enable a more comprehensive understanding of student performance and other external influences that impact the learning process. Another promising direction is to enhance user interaction by incorporating additional visual information directly into the layouts, including cluster-level summaries that display average performance, pass rates, or attendance statistics. Expanding the dataset itself could also provide richer analyses. For instance, tracking which instructor taught each module may uncover relevant correlations related to specific teaching strategies. Finally, we intend to apply and validate the system in collaboration with coordinators from other higher education courses.

# Declarations

## Acknowledgements

The authors thank the Directorate of Academic Administration and Control (DIRAC) at the Federal University of Uberlandia (UFU) for providing the entire data used in this research.

## Funding

This research was supported by National Council for Scientific and Technological Development (CNPq), Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) and Research Support Foundation of the State of Minas Gerais (FAPEMIG) (grant APQ-00150-21).

## Authors' Contributions

All authors were involved in every stage of the work and contributed equally to its completion.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

The dataset used in this study is publicly available for research purposes at Mendeley Data. It contains comprehensive information on academic performance, attendance, and module details from the Faculty of Computing at UFU, enabling further analyses and applications in educational data research Oliveira *et al*. [2025].

# References

Ali, S., Hafeez, Y., Humayun, M., Jamail, N. S. M., Aqib, M., and Nawaz, A. (2022). Enabling recommendation system architecture in virtualized environment for e-learning. *Egyptian Informatics Journal*, 23(1):33–45. DOI: 10.1016/j.eij.2021.05.003.

Badal, Y. T. and Sungkur, R. K. (2023). Predictive modelling and analytics of students' grades using machine learning algorithms. *Education and Information Technologies*, 28(3):3027–3057. DOI: 10.1007/s10639-022-11299-8.

Chen, Z., Wang, J., Xia, M., Shigyo, K., Liu, D., Zhang, R., and Qu, H. (2024). StuGPTViz: A visual analytics approach to understand student-ChatGPT interactions. *arXiv preprint arXiv:2407.12423*. DOI: 10.48550/arXiv.2407.12423.

Conselho de Graduação da Universidade Federal de Uberlândia (2022). Normas gerais da graduação da universidade federal de uberlândia. Available at: `http://www.reitoria.ufu.br/Resolucoes/ataCONGRAD-2022-46.pdf`. Accessed on November 17, 2024. In Portuguese.

Cox, T. F. and Cox, M. A. A. (2008). Multidimensional scaling. In houh Chen, C., Härdle, W., and Unwin, A., editors, *Handbook of Data Visualization*, pages 315–347. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-33037-0_14.

Deng, H., Wang, X., Guo, Z., Decker, A., Duan, X., Wang, C., Ambrose, G. A., and Abbott, K. (2019). PerformanceVis: Visual analytics of student performance data from an introductory chemistry course. *Visual Informatics*, 3(4):166–176. DOI: 10.1016/j.visinf.2019.10.004.

Etemadpour, R., Zhu, Y., Zhao, Q., Hu, Y., Chen, B., Sharier, M. A., Zheng, S., and S Paiva, J. G. (2020). Role of absence in academic success: an analysis using visualization tools. *Smart Learning Environments*, 7(2):1–25. DOI: 10.1186/s40561-019-0112-3.

Garcia-Zanabria, G., Gutierrez-Pachas, D. A., Camara-Chavez, G., Poco, J., and Gomez-Nieto, E. (2022). SDA-Vis: A visualization system for student dropout analysis based on counterfactual exploration. *Applied Sciences*, 12(12). DOI: 10.3390/app12125785.

Goulden, M. C., Gronda, E., Yang, Y., Zhang, Z., Tao, J., Wang, C., Duan, X., Ambrose, G. A., Abbott, K., and Miller, P. (2019). CCVis: Visual analytics of student online learning behaviors using course clickstream data. *Electronic Imaging*, 31:681–1–681–11. DOI: 10.2352/ISSN.2470-1173.2019.1.VDA-681.

Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G., and Gomez-Nieto, E. (2023). Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the latin american context. *Education Sciences*, 13(2):1–19. DOI: 10.3390/educsci13020154.

Hussain, S. and Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of Data Science*, 10(3):637–655. DOI: 10.1007/s40745-021-00341-0.

INEP (2020). Em dez anos, 40% dos que iniciaram um curso o concluíram. Available at: `https://www.gov.br/inep/pt-br/assuntos/noticias/censo-da-educacao-superior/em-dez-anos-40-dos-que-iniciaram-um-curso-o-concluiram`. Accessed on July 17, 2024. In Portuguese.

Jolliffe, I. T. (2002). *Principal Component Analysis for Special Types of Data*, chapter 8, pages 338–372. Springer Series in Statistics. Springer, New York, 2 edition. DOI: 10.1007/b98835.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In Kerren, A., Stasko, J. T., Fekete, J.-D., and North, C., editors, *Information Visualization*, pages 154–175. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-70956-5_7.

Kuh, G. D., Kinzie, J., Schuh, J. H., and Whitt, E. J. (2011). *Student success in college: Creating conditions that matter*. John Wiley & Sons, USA. Book.

Mandinach, E. B. and Abrams, L. M. (2022). Data literacy and learning analytics. In Lang, C., Siemens, G., Wise, A. F., Gasevic, D., and Merceron, A., editors, *The handbook of learning analytics*, chapter 19, pages 196–204. SoLAR, Canada. DOI: 10.18608/hla22.019.

ManpowerGroup (2022). Escassez de talentos no brasil e no mundo: quem detém o talento, detém o futuro. Available at: `https://blog.manpowergroup.com.br/escassez-de-talentos-no-brasil-e-no-mundo-quem-detem-o-talento-detem-o-futuro`. Accessed on July 17, 2024. In Portuguese.

Martins, M. V., Baptista, L., Machado, J., and Realinho, V. (2023). Multi-class phased prediction of academic performance and dropout in higher education. *Applied Sciences*, 13(8):1–15. DOI: 10.3390/app13084702.

Martins, R. M., Berge, E., Milrad, M., and Masiello, I. (2019). Visual learning analytics of multidimensional student behavior in self-regulated learning. In Scheffel, M., Broisin, J., Pammer-Schindler, V., Ioannou, A., and Schneider, J., editors, *Transforming Learning with Meaningful Technologies*, pages 737–741. Springer. DOI: 10.1007/978-3-030-29736-7_78.

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, 8:55462–55470. DOI: 10.1109/ACCESS.2020.2981905.

Minhoto, M. A., Smaili, S., and Arantes, P. (2023). 2,3 milhões abandonaram curso superior em 2021. Blog *Sou Ciência* – Folha de São Paulo. Available at: `https://www1.folha.uol.com.br/blogs/sou-ciencia/2023/02/23-milhoes-abandonaram-curso-superior-em-2021.shtml`. Accessed on May 14, 2023. In Portuguese.

Mohd Arsad, P., Buniyamin, N., and Ab Manan, J.-l. (2014). Neural network and linear regression methods for prediction of students' academic achievement. In *2014 IEEE Global Engineering Education Conference (EDUCON)*, pages 916–921. DOI: 10.1109/EDUCON.2014.6826206.

Munzner, T. (2014). *Visualization Analysis and Design*. CRC Press, New York. DOI: 10.1145/3721241.3733989.

Mushtaq, I. and Khan, S. N. (2012). Factors affecting studentsa academic performance. *Global Journal of Management and Business Research*, 12(9):17–22. Available at:`https://journalofbusiness.org/index.php/GJMBR/article/view/100221`.

Ogundele, I. M., Taiwo, O., Babalola, A. E., and Ayeni, O. C. (2024). Prediction of student academic performance based on machine learning model. In *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, pages 1–11. DOI: 10.1109/SEB4SDG60871.2024.10629703.

Oliveira, A., de Souza Paiva, J. G., and Ribeiro Gabriel, P. H. (2025). Analyzing higher education student performance through information visualization techniques. Available at: `https://data.mendeley.com/datasets/vggp27d2t5/2`. Accessed on January 15, 2025.

Oqaidi, K., Aouhassi, S., and Mansouri, K. (2022). Towards a students' dropout prediction model in higher education institutions using machine learning algorithms. *International Journal of Emerging Technologies in Learning (Online)*, 17(18):103–117. DOI: 10.3991/ijet.v17i18.25567.

Pachas, D. A. G., Garcia-Zanabria, G., Cuadros-Vargas, A. J., Camara-Chavez, G., Poco, J., and Gomez-Nieto, E. (2021). A comparative study of WHO and WHEN prediction approaches for early identification of university students at dropout risk. In *2021 XLVII Latin American Computing Conference (CLEI)*, pages 1–10. IEEE. DOI: 10.1109/CLEI53233.2021.9640119.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., and Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*, 80(3):3782–3785. DOI: 10.1016/j.matpr.2021.07.382.

Portela, A. (2023). Qualidade da educação está associada a maiores taxas de crescimento, revela estudo. Portal FGV. Available at: `https://portal.fgv.br/noticias/qualidade-educacao-esta-associada-maiores-taxas-crescimento-revela-estudo`. Accessed on May 14, 2023. In Portuguese.

Rahayu, A. P. and Dong, Y. (2023). The relationship of extracurricular activities with students' character education and influencing factors: A systematic literature review. *AL-ISHLAH: Jurnal Pendidikan*, 15(1):459–474. DOI: 10.35445/alishlah.v15i1.2968.

Realinho, V., Machado, J., Baptista, L., and Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, 7(11):1–17. DOI: 10.3390/data7110146.

Shahbazi, Z. and Byun, Y.-C. (2022). Agent-based recommendation in e-learning environment using knowledge discovery and machine learning approaches. *Mathematics*, 10(7):1192. DOI: 10.3390/math10071192.

Susnjak, T., Ramaswami, G. S., and Mathrani, A. (2022). Learning analytics dashboard: a tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, 19(1):12. DOI: 10.1186/s41239-021-00313-7.

Tsung, S., Wei, H., Li, H., Wang, Y., Xia, M., and Qu, H. (2022). Blocklens: Visual analytics of student coding behaviors in block-based programming environments. In *Proceedings of the Ninth ACM Conference on Learning @ Scale*, L@S '22, page 299–303, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3491140.3528298.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11). Available at:`http://jmlr.org/papers/v9/vandermaaten08a.html`.

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(11):1–19. DOI: 10.1186/s40561-022-00192-z.

Zaki, M. J. and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, Cambridge, MA. Book.

Zhang, H., Dong, J., Lv, C., Lin, Y., and Bai, J. (2022). Visual analytics of potential dropout behavior patterns in online learning based on counterfactual explanation. *Journal of Visualization*, 26:723–741. DOI: 10.1007/s12650-022-00899-8.