LearnVis: Analyzing Higher Education Student Performance through Information Visualization Techniques

Angélica Gomes Oliveira

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2025

Angélica Gomes Oliveira

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

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: José Gustavo de Souza Paiva

Uberlândia
2025

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

O48
2025

Oliveira, Angélica Gomes, 1994-
LearnVis: Analyzing Higher Education Student Performance through Information Visualization Techniques [recurso eletrônico] / Angélica Gomes Oliveira. - 2025.

Orientador: José Gustavo de Souza Paiva.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2025.459
Inclui bibliografia.

1. Computação. I. Paiva, José Gustavo de Souza,1979-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. III. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação, 19/2025, PPGCO				
Data:	25 de Julho de 2025	Hora de início:	14:05	Hora de encerramento:	16:00
Matrícula do Discente:	12222CCP004				
Nome do Discente:	Angélica Gomes Oliveira				
Título do Trabalho:	LearnVis: Analyzing Higher Education Student Performance through Information Visualization Techniques				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Ciência de Dados				
Projeto de Pesquisa de vinculação:	Demanda Universal FAPEMIG - Processo: APQ-00150-21				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Rafael Dias Araújo - FACOM/UFU, Jorge Luis Poco Medina - FGV EMAp e José Gustavo de Souza Paiva - FACOM/UFU, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: Jorge Luis Poco Medina - Rio de Janeiro/RJ. Os outros membros da banca e a aluna participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. José Gustavo de Souza Paiva , apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação da Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir ao candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **José Gustavo de Souza Paiva**, **Professor(a) do Magistério Superior**, em 30/07/2025, às 15:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rafael Dias Araújo**, **Professor(a) do Magistério Superior**, em 30/07/2025, às 15:31, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Jorge Luis Poco Medina**, **Usuário Externo**, em 31/07/2025, às 10:55, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6488736** e o código CRC **60F0E970**.

Referência: Processo nº 23117.046337/2025-18 SEI nº 6488736

This work is dedicated to those who,
at some point in life, dared to dream — and persisted.

Agradecimentos

I would like to express my gratitude, first and foremost, to God, for granting me strength in moments of uncertainty and serenity in times of challenge. My husband, Luiz Miguel, for standing by my side with patience, love, and unconditional support throughout this journey. My parents, for their example of dedication, the values they have taught me, and for always believing in me. My advisor, Prof. Dr. José Gustavo, for his careful guidance, constant encouragement, and generous sharing of knowledge throughout this work. Prof. Dr. Paulo Henrique, for his support and valuable contributions. FAPEMIG, for funding Project APQ-00150-21, which enabled the development of this research. And the members of the examination committee, for their availability and for accepting to take part in this important moment of my journey. To all, my most sincere thanks.

"An investment in knowledge always pays the best interest."
(Benjamim Franklin)

Resumo

O sucesso acadêmico dos alunos no ensino superior é influenciado por diversos fatores, como desempenho nas avaliações, frequência às aulas e engajamento com os tópicos ministrados. Nesse contexto, este trabalho apresenta o ***LearnVis***, um sistema de visualização interativa desenvolvido para auxiliar profissionais da educação na análise detalhada do desempenho dos alunos ao longo do curso. O sistema permite explorar diferentes perspectivas, como padrões de frequência, múltiplas tentativas em um mesmo módulo, relação entre faltas e notas, e o impacto de tópicos específicos no sucesso ou fracasso dos alunos. A proposta foi desenvolvida a partir de requisitos e tarefas definidos em conjunto com especialistas da área educacional e aplicado a dados reais de 1.490 estudantes do curso de Sistemas de Informação da Universidade Federal de Uberlândia, abrangendo o período de 2009 a 2019. Por meio de estudos de caso, foi possível demonstrar que o ***LearnVis*** auxilia na identificação de padrões críticos, destacando situações que podem demandar intervenções, como módulos com altas taxas de reprovação ou grupos de alunos com comportamento de risco. Como contribuição, o sistema oferece suporte à tomada de decisão, auxiliando na criação de estratégias pedagógicas mais assertivas, no aprimoramento do currículo e no aumento da retenção dos alunos.

Palavras-chave: Análise de Desempenho de Alunos. Análise de Dados Educacionais. Análise Visual. Visualização da Informação. Projeção Multidimensional.

Angélica Gomes Oliveira

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2025

Abstract

Student success in higher education is influenced by several factors, such as academic performance, class attendance, and engagement with course topics. In this context, this work presents ***LearnVis***, an interactive visualization system designed to support education professionals in conducting an in-depth analysis of student performance throughout the academic program. The system enables the exploration of different perspectives, including attendance patterns, multiple attempts in the same module, the relationship between absences and grades, and the impact of specific topics on student success or failure. The system was developed based on requirements and tasks defined in collaboration with education experts and was applied to real data from 1,490 students enrolled in the Information Systems program at the Federal University of Uberlândia, covering the period from 2009 to 2019. Through case studies, the results demonstrate that ***LearnVis*** helps identify critical patterns, highlighting situations that may require interventions, such as modules with high failure rates or groups of students exhibiting at-risk behavior. As a contribution, the system provides decision-making support, helping to develop more effective pedagogical strategies, improve curriculum design, and enhance student retention.

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

List of Figures

List of Tables

Acronyms list

ANOVA Analysis of Variance

BISE Board of Intermediate and Secondary Education

CNN Convolutional Neural Network

C-index Concordance Index

CGPA Cumulative Grade Point Average

XAI Explainable Artificial Intelligence

ELRA E-learning Recommendation Architecture

FACOM Faculty of Computing

GRU Gated Recurrent Unit

IML Interpretable Machine Learning

Isomap Isometric Feature Mapping

KNN K-Nearest Neighbor

EaD Distance Learning

LMS Learning Management System

LSTM Long Short-Term Memory

LIME Local Interpretable Model-agnostic Explanations

LSP Least Square Projection

LAMP Local Affine Multidimensional Projection

MLP Multilayer Perceptron

ML Machine Learning

MSE Mean Squared Error

MDS Multidimensional Scaling

MOOCs Massive Open Online Courses

NLP Natural Language Processing

PI Permutation Importance

PDP Partial Dependence Plot

PCA Principal Component Analysis

AUC ROC Curve

SHAP SHapley Additive exPlanations

t-SNE T-Distributed Stochastic Neighbor Embedding

UFU Federal University of Uberlândia

Contents

Chapter 1

Introduction

Higher education institutions face several challenges, including high dropout and low graduation rates. The accumulated dropout rate in Brazil between 2018 and 2022 reached 57.2%, considering both on-campus and Distance Learning (EaD) programs in public and private institutions (SEMESP, 2024). In the same period, the accumulated graduation rate was only 25.4%, while the retention rate, that is, students who remained enrolled in the programs, was 17.4%. Recent studies highlight the severity of these issues and point out to several contributing factors, such as socioeconomic difficulties, absence of effective public policies for student retention, and issues related to students' academic performance (SILVA; SAMPAIO, 2022; FIOR et al., 2022; BRANDT; TEJEDO-ROMERO; ARAUJO, 2020). The quality of education in the country plays a fundamental role in socioeconomic development, impacting economic growth, job creation, and workers' income (PORTELA, 2023). High dropout and low graduation rates in higher education have generated significant impacts not only for educational institutions but also for the labor market and the economy as a whole. With fewer students completing their degrees, a shortage of qualified professionals emerges. In 2022, a global survey revealed that 75% of employers faced difficulties finding professionals with the necessary qualifications (MANPOWERGROUP, 2022). In Brazil, the same survey indicated that the talent shortage exceeded the global average, reaching 81%. According to the Leopold et al. (2025), it is estimated that, between 2025 and 2030, the structural transformation of the labor market will result in 78 million new job opportunities worldwide. However, this job creation will take place amid a growing gap between the skills demanded and those actually available in the labor market, with insufficient training and low qualification being identified as some of the main concerns among employers globally.

Students' academic performance has been identified as a central factor for retention and success in higher education. Research shows that students with learning difficulties, low performance in key modules, and a history of failures present a higher risk of dropping out or abandoning their undergraduate programs (SILVA; SAMPAIO, 2022; FIOR et al., 2022; BRANDT; TEJEDO-ROMERO; ARAUJO, 2020). These effects are even more notable in the early stages of the program, when students face academic adaptation challenges, prior learning gaps, and difficulties related to mastering fundamental competencies such as critical reading, interpretation, and academic writing (WILCOX; WINN; FYVIE-GAULD, 2005; TINTO, 2017; GOMES; SCHLICKMANN, 2024). Studies also show that emotional and psychological factors, such as low self-efficacy, academic insecurity and lack of motivation directly influence performance, compromising students' continuity in the program (FIOR et al., 2022).

Educational data refers to a broad set of information generated within the educational context, encompassing data related to students' academic performance, such as attendance records, grades, topics covered in the modules, and final outcomes, which we refer to in this study as **internal data**. Additionally, it includes demographic, social, and economic data, referred to as **external data**, which indirectly influence students' academic journeys (OQAIDI; AOUHASSI; MANSOURI, 2022; GUTIERREZ-PACHAS et al., 2023). Educational institutions store both types of data, enabling integrated analysis of students' background conditions (external data) and their performance throughout the program (internal data). The importance of collecting and analyzing educational data for the continuous improvement of education and learning has been widely discussed in the literature (KOCSIS; MOLNÁR, 2024; LIN et al., 2024; SUSNJAK; RAMASWAMI; MATHRANI, 2022), highlighting its role in supporting analytical tasks that guide decision-making processes for educational managers and experts.

Several computational approaches to analyze academic performance were proposed employing machine learning techniques in important tasks such as dropout prediction, grade prediction, and learning recommendations (MUSHTAQ; KHAN, 2012; YAĞCI, 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), assisting 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 datasets (GARCIA-ZANABRIA et al., 2022; ZHANG et al., 2022; DENG et al., 2019), that, although showing some correlation with student performance, may not represent an effective source of analysis, as they are not directly related to the educational process. 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.

1.1 Objectives and Contributions

This work presents a visual analysis methodology for academic data, employing advanced Information Visualization techniques on internal data about students' behavior in terms of class attendance and academic success or failure. The methodology employs a set of coordinated visual layouts that allow users to explore internal data such as attendance records, grades, and course structure, organized by modules and topics. It also enables the analysis of students' academic trajectories across semesters, tracking multiple attempts until the successful completion of each module. This approach allows for both individual and group-level analyses, facilitating the identification of patterns and trends among students with similar profiles. Our hypothesis is that the combination of internal academic data with Information Visualization and Visual Analytics techniques offers a more comprehensive and detailed perspective on students' performance, enabling the identification of behavioral patterns, critical points in the learning process, and supporting evidence-based decision-making to design and evaluate policies that enhance academic success.

We propose a visualization system called ***LearnVis***, designed to support the analysis of higher education students' behavior. This system was applied to real data from the Faculty of Computing (FACOM) at the Federal University of Uberlândia (UFU), in Brazil. It is important to highlight that the system can be applied to data from any undergraduate program in any educational institution, thereby broadening the scope and relevance of the analyses conducted in this research.

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 Computing and Informatics courses from a Brazilian university;

III. a set of evolution experiments using the data repository from II., demonstrating how the system highlights the phenomena present in the data, as well as how it helps to guide education experts' decision making.

The work described here was developed in collaboration with Prof. Dr. Paulo Henrique Ribeiro Gabriel for the submission of the following paper:

- Oliveira, A. G.; Paiva, J. G. S.; Gabriel, P. H. R. LearnVis: Analyzing higher education student performance through information visualization techniques. **Jornal of the Brazilian Computer Society. (Under review in 08/2025)**

1.2 Dissertation Organization

The dissertation is organized into five chapters. The chapters are structured as follows:

- In Chapter 2, we present the fundamentals of this work, including essential concepts about Education, Information Visualization, and a review of related works from the literature.
- In Chapter 3, we present the proposed approach, including the system requirements and the main tasks that guided the project design. This chapter also provides a description of the ***LearnVis*** system and its implementation details.
- In Chapter 4, we present the experimental results, including information about the dataset used and a set of evolution experiments.
- Finally, in Chapter 5, we summarize the main conclusions of this research, discuss its limitations, and propose directions for future work.

Chapter 2

Fundamentals

Understanding student performance throughout the course is essential for evaluating the effectiveness of the curriculum structure and the topics covered in the modules. Academic progression can be influenced by various factors, including course organization, interdependence between modules, and how students engage with topics over time. Academic records provide a rich source of data that can be explored to identify patterns of success and challenges faced by students. Different analytical approaches have been applied for this purpose, ranging from traditional statistical methods to machine learning techniques (VENKATESAN; KARMEGAM; MAPPILLAIRAJU, 2024; BELLAJ et al., 2024). However, the complexity and volume of these data often make direct interpretation challenging, and Information Visualization techniques may facilitate analysis and knowledge extraction. By transforming raw data into interactive visual representations, the produced layouts enable the identification of patterns in a more intuitive way, assisting in the development of strategies to improve academic performance. Given this scenario, this study proposes a visual analysis system aimed at providing a better understanding of the factors that impact student learning and academic trajectories. This chapter presents the fundamental concepts necessary for understanding this research, as well as reviewing related works that motivated its proposal.

2.1 Basic Concepts

To facilitate the understanding of the proposal, as well as the works related to it, some basic concepts employed in the literature are presented below:

- **Student:** an individual enrolled in an educational institution, which in the context of this work is a university, representing one of the main subjects of study in the educational context;

- **Course:** a structured and organized program of modules designed to provide knowledge and competencies in a specific area of study. It consists of a sequence of lectures, practical activities, and other forms of instruction;

- **Course period:** refers to a time unit that represents each sequential stage of the curriculum structure. Each period groups the modules planned to be taken at that stage, reflecting the student's academic progression;

- **Academic semester:** a specific course period within the academic year in which teaching and learning activities take place in an educational institution. Typically, an academic semester is a division of the academic year into two sections, during which the modules are scheduled;

- **Module:** a set of topics related to a specific skill or competency within the context of a particular course;

- **Topic:** subject matter related to a specific competence, presented to students to convey information, knowledge, and concepts related to a particular subject or module;

- **Lecturer:** a professional responsible for transmitting knowledge in classes and/or practical activities, guiding learning, and facilitating the development of skills and competencies in students;

- **Class:** a unit of instruction in which specific topic is delivered to students in various formats, including lectures, group discussions, practical activities, demonstrations, or a combination of several methodologies;

- **Absence:** the lack of participation or attendance of a student in classes, activities, or educational events;

- **Grade:** a numerical or descriptive assessment given to a student based on their performance in evaluations, tests, assignments, projects, or other academic tasks. Grades are important for assessing student progress and guiding academic decisions, including determining passing, failing, or recommending subsequent educational levels;

- **Success/Failure:** success refers to achieving the minimum performance required in a module, encompassing both grades and attendance, as defined by the guidelines of the educational institution. Conversely, fail occurs when a student does not meet the institution's criteria, meaning they have not attained the minimum performance in a module in terms of grades and attendance;

- **Dropout:** the abandonment of a course by students before completing all the required modules;

- **Data collection:** a set of multiple instances or related data records. This data collection can be organized and represented in various formats, such as tables, spreadsheets, databases, or digital files. In the educational context, it can be a student database containing information such as demographic data, entry method, current student status, class attendance, grades in modules, among others.

- **Instance:** an object of study in a given context that contains a set of features of interest for analysis, represented by its features. In the educational context, an instance may contain information about a student, such as grades obtained in specific modules, attendance records, history of absences or participation in extracurricular activities;

- **Feature (dimension):** a characteristic or property that describes an object or entity of interest. In the educational context, features are used to represent the different variables or characteristics observed and/or measured in students, subjects, or other relevant elements. These features can be used to categorize, classify, or segment the data;

- **Layouts:** the visual organization of elements in a graphical representation, using Information Visualization techniques. They play a crucial role in the clarity and understanding of data, influencing how information is perceived and interpreted by the user.

2.2 Educational Data

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. Table 1 presents some commonly stored information in educational data, which is used in the corresponding analyses. In this context, several studies have explored the use of internal and external educational data in automated analysis systems. These approaches leverage academic records to identify patterns, predict student performance, and support decision-making in educational settings. The following sections will detail some of these studies, highlighting their methodologies and contributions.

Table 1 – Example of educational data representation.

Representation	Feature	Type
Internal - Academic	Topic	Nominal
	Course	
	Module	
	Dropout Type	
	Admission Type	
	Final Result	
	Class Data	Temporal
	Data Admission	
	Attendance	Scalar
	Grades	
	Course period	Ordinal
	Academic Semester	
External - Demographic	Color/Race	Nominal
	Marital Status	
	Gender	
	Nationality	
	Age	Discrete
External - Socio-Economic	Parents' occupation	Nominal
	Parent qualification	

Source: Prepared by the Author (2025).

2.3 Automatic Analysis of Educational Data

Automatic data analysis is a process that involves computational techniques such as Machine Learning (ML) and data mining to build predictive models (ZAKI; MEIRA, 2014). These models can be used to make predictions about students' future performance or to identify factors that influence their academic progress. The applications of automatic data analysis in the educational field are diverse and encompass several important tasks, including grade prediction, dropout prediction, and recommendation.

- Grade Prediction: Grade prediction refers to the early estimation of a student's future scores using historical data, such as assessments, participation, and other performance indicators. Studies in this area have applied approaches ranging from recurrent neural networks (GHAZVINI; SHAREF; SIDI, 2024; LAKSHMI; MAHESWARAN, 2024) to decision trees and ensemble methods (HUSSAIN; KHAN, 2023; BADAL; SUNGKUR, 2023).

- Dropout Prediction: Dropout prediction focuses on identifying students at risk of leaving their studies before completion, enabling the development of preventive strategies to improve retention and academic success. Research has explored logistic regression, neural networks, random forests, and interpretable models to address this challenge (RABELO; ZáRATE, 2025; NAGY; MOLONTAY, 2024; GUTIERREZ-PACHAS et al., 2023; PACHAS et al., 2021; MARTINS et al., 2023).

- Recommendation: A recommendation system suggests relevant items based on user preferences. In the educational context, it assists students and educators in providing a personalized learning experience by adapting to each student's preferences, skills, and progress. Approaches in the literature include clustering, feature selection, semantic analysis, and agent-based systems (ARCINAS et al., 2025; RAO; WANG; LI, 2024; SHAHBAZI; BYUN, 2022; ALI et al., 2022).

2.4 Information Visualization

Information visualization is a process that transforms abstract and complex data into clear, intuitive, and meaningful visual representations (MUNZNER, 2014). Card, Mackinlay e Shneiderman (1999) introduced a visualization pipeline, which can be described according to Figure 1.

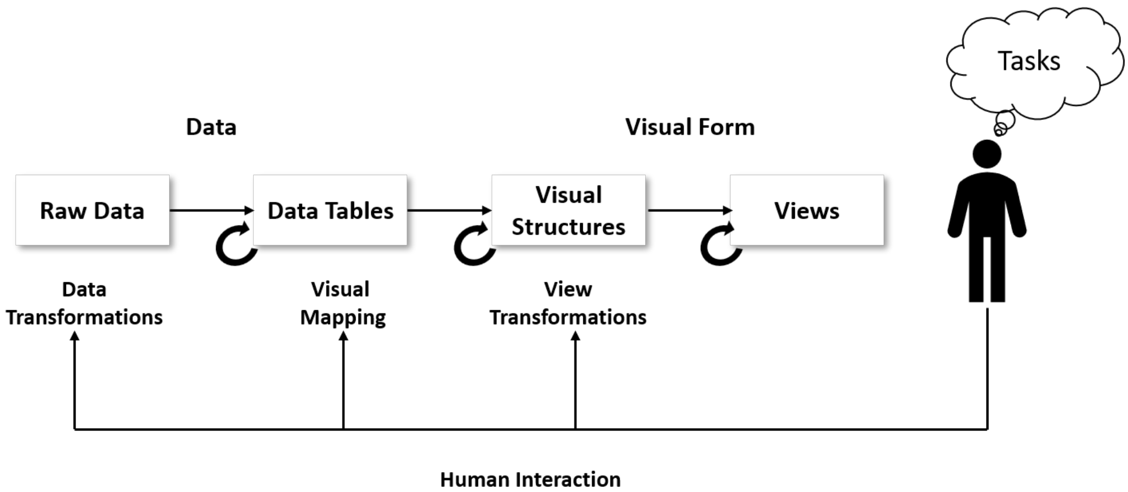


Figure 1 – Example of an Information Visualization Process (Image adapted from: (CARD; MACKINLAY; SHNEIDERMAN, 1999)).

The process begins with the provision of raw data, which contains information or records related to a specific context in its original form, without processing or formatting. Next, the raw data undergoes a transformation process to become standardized and organized into a data table, which will be interpreted by visual analysis techniques. At this stage, the data is processed in various ways, including handling missing values, rescaling/normalizing data, and other tasks. The data tables are then mapped into visual structures using metaphors that relate to the tasks to be performed and the context in question. These visual structures are organized based on one or several aspects of the analysis, forming interactive views that users explore executing specific tasks.

There are various techniques in Information Visualization, which can be categorized based on how layouts are constructed and information is presented. Among these categories, feature-based techniques stand out, employing visual properties such as colors, shapes, sizes, and textures to encode different data features. These techniques aim to analyze how features relate to one another to determine the behavior of instances. The heatmap technique is an example of an feature-based approach, offering a clear and direct representation of data. This method maps a color gradient to a range of measured values, organizing them into a matrix where each cell is assigned a color corresponding to its specific data value. Figure 2 shows an example of a heatmap used to understand students' academic performance in colleges (ETEMADPOUR et al., 2020). The horizontal axis represents absence values, while the vertical axis represents grades in the Human-Computer Interaction course. The lower the color intensity of a cell, the higher the percentage of students with absences. In this example, approximately 28.57% of students with zero absences scored around 70% as their final grade. This layout enables the understanding of how attendance may influence students' grades in each course.

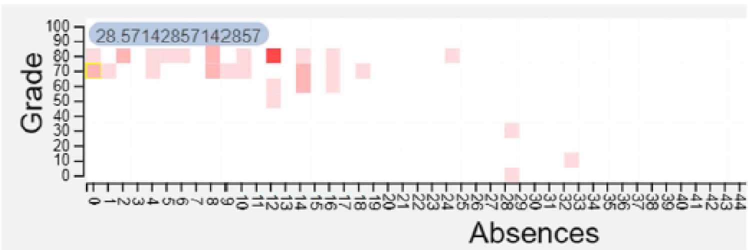


Figure 2 – *Heatmap*: Attendance vs. Grade for the module *Human-Computer Interaction (HCI)* (Image sourced from: (ETEMADPOUR et al., 2020)).

Another feature-based technique is the scatterplot matrix (CLEVELAND, 1993). It displays the relationship between multiple pairs of variables using scatterplots and correlation coefficients. This matrix is organized into a grid of cells, where each cell represents the relationship between two specific variables. In the cells of the main diagonal, scatterplots or histograms show the individual distribution of each variable. In the off-diagonal cells, scatterplots illustrate how two variables relate to each other. Correlation coefficients can be calculated and displayed to indicate the strength and direction of the linear relationship between the variables. Figure 3 presents an example of a scatterplot matrix for an air quality dataset in China (LIU et al., 2023). In this visualization, histograms on the main diagonal represent the distribution of each feature. However, histograms can sometimes appear irregular and vary depending on bin size. To address this, the study incorporates a kernel density curve, which smooths the distribution of the data, providing a clearer representation.

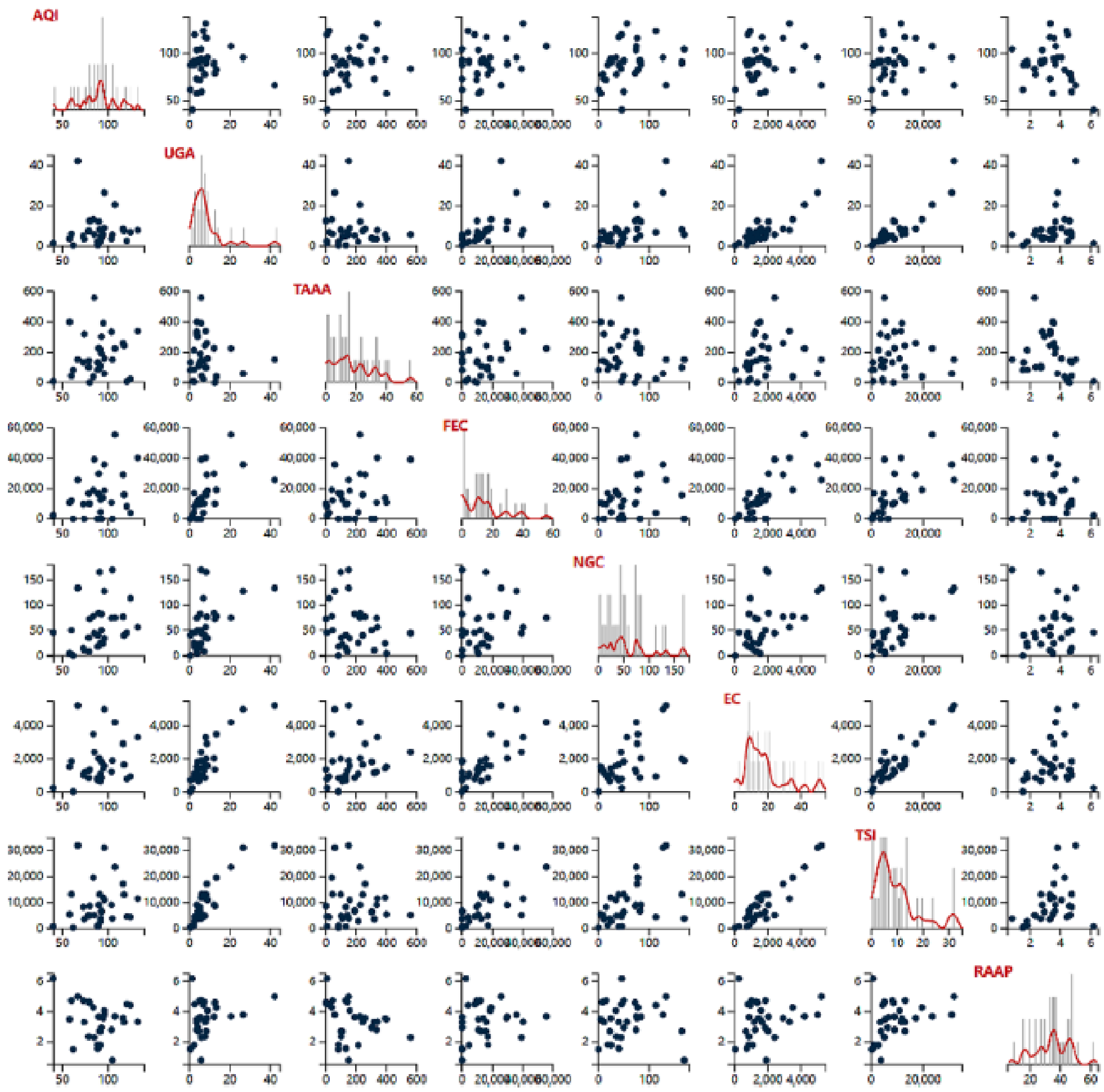


Figure 3 – Scatterplot Matrix of an Air Quality Dataset from China (Image sourced from: (LIU et al., 2023)).

Several other feature-based visualization techniques, such as Pixel Bar Charts (KEIM et al., 2002), Sankey diagrams (RIEHMANN; HANFLER; FROEHLICH, 2005), Radar Charts (MAYR, 1877), Treemaps (JOHNSON; SHNEIDERMAN, 1998) can also be employed. For a comprehensive survey of techniques, the reader can consult (LIU et al., 2014).

Another widely used class of visualization techniques in the literature includes point-placement techniques, in which a set of points is arranged in a two-dimensional or three-dimensional space to highlight the positioning of instances within the visualization space, ideally representing the relationships present in the original space where these instances are represented. The goal is to position similar instances closer together, while dissimilar instances farther apart, allowing for clear distinctions between them according to a specific aspect (PAULOVICH et al., 2008). We present as follows some point-placement techniques.

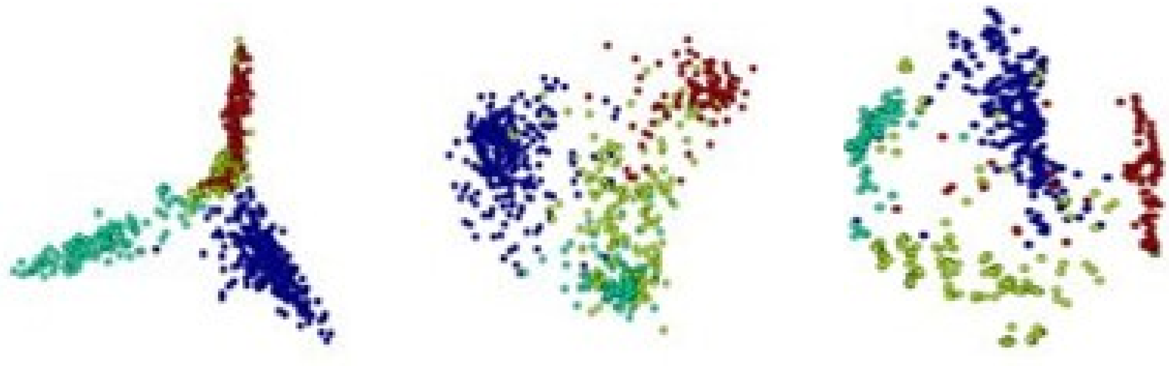
Multidimensional Scaling (MDS) techniques aim to transform high-dimensional data into a visual representation in a two- or three-dimensional space. This is achieved such that the distances between points in the reduced space reflect the distances or similarities between observations in the original space. The procedure involves projecting the points into a lower-dimensional space while ensuring that the Euclidean distance between any two points in this space is equivalent to the dissimilarity function value between the same points in the original high-dimensional space, applied to all pairs of points (COX; COX, 2008). In classical MDS, the dissimilarity function corresponds precisely to the Euclidean distance between points, and the method seeks to find additional coordinates in extra dimensions to preserve the distances between points (BORG; GROENEN, 2005).

PCA (JOLLIFFE, 2002) performs dimensionality reduction by combining the information from all dimensions into new axes called principal components, which capture the most significant variance in the original data. The principal components are linear combinations of the original variables, with the first principal component containing the greatest variance, followed by the second principal component, and so on. Figure 4a illustrates the application of the PCA technique to a document collection which includes 680 documents in four distinct topics, with an imbalance in the number of documents across labels. Each document is represented by a word list containing 1,423 terms or dimensions.

Isomap (TENENBAUM; SILVA; LANGFORD, 2000) is a nonlinear dimensionality reduction technique based on classical MDS. In this method, original distances are replaced with geodesic distances calculated on a nearest-neighbor graph. This aims to obtain a global estimate of the optimal solution for maintaining distances. To construct the weighted nearest-neighbor graph, the dataset is used, where distances between pairs of points are considered as edge weights. The distance between two points is derived from the shortest path within this graph. Figure 4b illustrates the application of the Isomap technique to the same dataset as Figure 4a.

LSP (PAULOVICH et al., 2008) is defined as a multidimensional projection method that starts with the initial projection of a sample of the original dataset, referred to as control points. This initial projection determines the positioning of the remaining points by solving a linear system constructed based on the neighborhood relationships of these points in their original space and the Cartesian coordinates of the control points in the projection space, aiming to preserve the similarity relationships among points as defined by a multidimensional metric. Results demonstrate that the technique can produce clusters of points by degree of similarity in a two-dimensional space. LSP is a modern technique notable for its efficiency and high accuracy, as confirmed by objective quality measures. Figure 4c shows an example of the application of the LSP technique to the same dataset as Figures 4a and 4b.

The LAMP is a multidimensional projection technique designed to map high-dimensional data onto a lower-dimensional visual space while preserving local relationships between



(a) (b) (c)

Figure 4 – Examples of Point-Based Visualization Techniques: **(a)** PCA, **(b)** Isomap, **(c)** LSP. Images sourced from (ETEMADPOUR et al., 2014).

data points. Based on orthogonal mapping principles, LAMP leverages control points to guide the transformation process, enabling precise local adjustments that can be dynamically refined based on user input (JOIA et al., 2011). Figure 5 shows the visualization created from the data in the Segmentation collection[1], which contains information on 2,100 manually segmented images.

Figure 5 – LAMP (Images sourced from: (JOIA et al., 2011)).

The t-SNE technique (MAATEN; HINTON, 2008) aims to perform nonlinear dimensionality reduction while preserving the similarity relationships between the data. In this process, it creates a probability distribution in a high-dimensional space that reflects the similarities among the original data. Then, a similar distribution is constructed in a lower-dimensional space, with parameter adjustments to maintain the similarity relationships. Figure 6 illustrates the application of this technique to the cytometry dataset.

[1] Available at http://archive.ics.uci.edu/dataset/50/image+segmentation

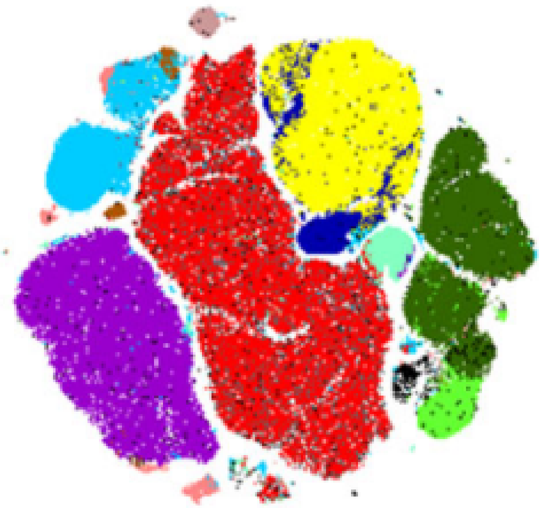


Figure 6 – t-SNE (Images sourced from: (MAATEN; HINTON, 2008)).

To interpret the layouts generated by the visualization techniques presented earlier, the analyst must adopt a meticulous and systematic approach, carefully observing the formation of groups or clusters of points in the space. The proximity between groups indicates a higher similarity between the corresponding elements in the original data. The analyst should also evaluate the density of these clusters as well as the distances that separate them. It is essential to pay attention to visual trends, such as the distribution of points along an axis, which may indicate variations in specific features. The identification of outliers, i.e., isolated points, also plays an important role, as it may signal special cases or errors in the data. Regarding the analysis of colors, shapes, or other representations used to encode variables, the analyst has the opportunity to identify patterns that suggest the magnitude of relationships between different features.

2.5 Visual Analysis of Educational Data

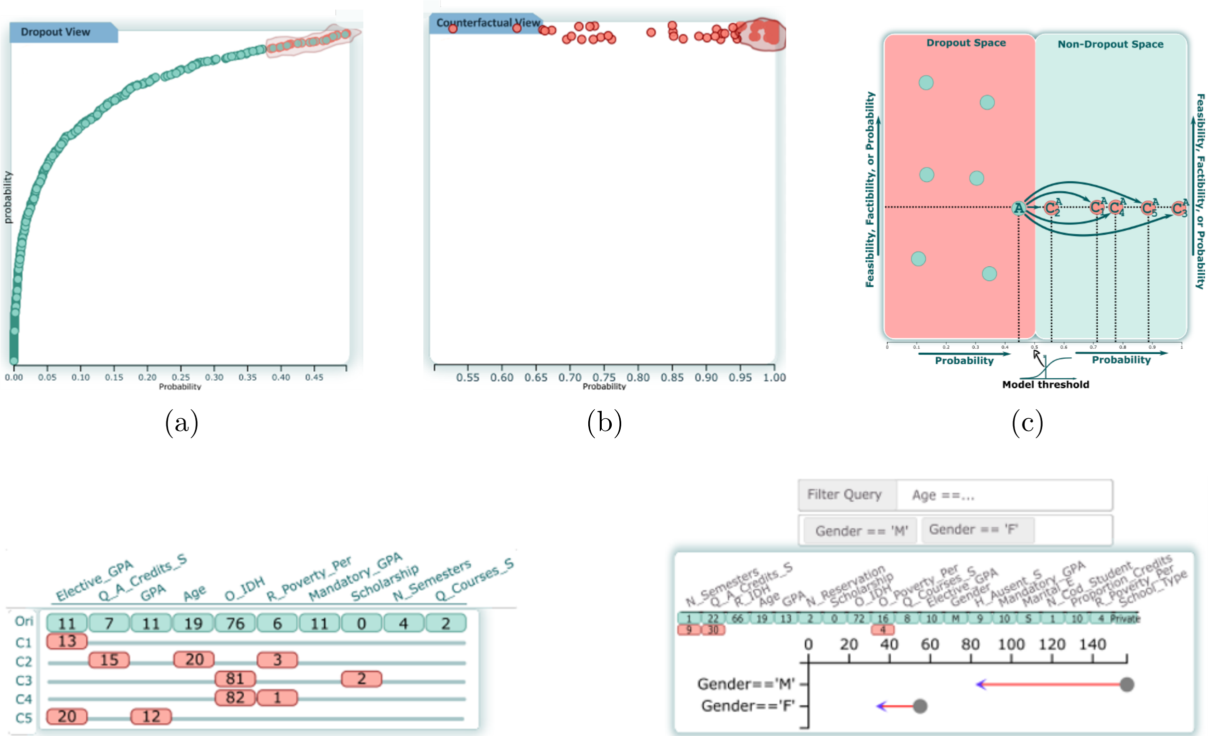
In this section, studies applying Information Visualization techniques in the educational context are presented, serving as motivation for the development of this work.

Several studies use the concept of counterfactual analysis as a way to represent alternative scenarios generated as consequences of different choices or events, examining what could have occurred if the conditions had been altered. This allows researchers to analyze and compare actual outcomes with hypothetical results, helping to understand the importance of certain events or factors in a given situation.

Garcia-Zanabria et al. (2022) propose a visualization system called SDA-Vis for analyzing student dropout based on counterfactual exploration. The system considers academic, social, and economic variables from a real dataset encompassing over 11,000 students across 12 different undergraduate programs in Latin American universities. It provides users with the opportunity to explore hypothetical scenarios for students who dropped out, where, in these scenarios, dropout does not occur. This enables a better understanding of the possible changes that could have kept the students in the educational institution.

SDA-Vis offers various visualizations, including the **Student Projection Visualization (SP)** (Figure 7a). On the vertical axis, the analyst can choose among three different metrics—viability, feasibility, or probability—to analyze a specific student or a group of students.

After selecting a group of interest in the SP layout (Figure 7a), the **Counterfactual Projection Visualization (PC)** presents all counterfactuals associated with that choice, as illustrated in Figure 7b.



(a) (b) (c)

(d) (e)

Figure 7 – SDA-Vis views: **(a)** Student Projection view, **(b)** Counterfactual Projection view, **(c)** Representation of students and their counterfactual projections in the probabilistic space, **(d)** Counterfactual Exploration view, and **(e)** Impact View. Images adapted from: (GARCIA-ZANABRIA et al., 2022).

The representation in Figure 7c describes how the data is distributed across these two layouts. The left side corresponds to the SP view, while the right side corresponds to the CP view, separated by a line denoting the decision boundary at 0.5. The range of dropout values varies from 0 to 0.49 in the SP view and from 0.5 to 1 in the CP view, depending on the chosen metric. When selecting a student or group of dropout students, five alternative scenarios (C1-C5) are displayed, in which these students would pass, highlighting the necessary changes in their original features to achieve this result (Figure 7d). For example, in the case of C5, it is sufficient to improve the ELECTIVE_GPA and GPA to 20 and 12, respectively, to prevent the student from dropping out of college. Additionally, the **Impact view** enables visualizing the effect of applying these alternative realities to groups of students (Figure 7e).

It is possible to select certain counterfactuals in the **Counterfactual Exploration view** (Figure 7d) and assess how these changes might affect each group. For instance, considering male students, the dropout rate is reduced from 150 to 80 students, resulting in a reduction of 70 dropout cases. In the case of female students, the technique prevents approximately 20 dropout cases. These layouts can assist in identifying viable solutions to guide educational institutions in making decisions aimed at reducing dropout rates, demonstrating how applying counterfactual strategies can positively impact student retention.

Similarly, Zhang et al. (2022) developed an interactive system called *DropoutVis* to visualize student behavior and dropout, using counterfactual explanations on a real dataset encompassing 79,186 students enrolled in 39 courses offered by a Chinese platform for Massive Open Online Courses (MOOCs). Moreover, the system employs a CNN-LSTM model to predict students' dropout risk. The analysis focuses on nine behavioral characteristics displayed by students on the platform, such as answering questions, watching videos, accessing course materials, and consulting Wikipedia[2], in addition to participating in forum discussions, among other activities. The tool provides five visualizations, as shown in (Figure 8). **The temporal circle view** (Figure 8a) presents factors that may contribute to student dropout. Courses last for 5 weeks, visualized in the layout as a ring with five layers, representing different periods of the course from the inside out. Each ring layer contains nine color-coded arcs to represent specific behavioral characteristics, such as video (purple), discussion (gray), and navigation (light blue). The length of the arc represents the degree of influence of the dropout factor: the longer the arc, the greater its impact on dropout occurrence, and vice-versa. By selecting a group, the dropout reason is displayed for an initial evaluation of the causes.

The **windmill view** (Figure 8b) is used to show how the model perturbs different characteristics when searching for counterfactual explanations, including the frequency and extent of the perturbation. It consists of two parts: the inner layer uses a radial line chart to display the frequency of behavioral characteristic perturbations. There are a total of five radial axes representing the course time analyzed in a clockwise direction, with each line representing a behavioral characteristic. The outer layer is represented by an area chart designed to analyze the average range of perturbation for each behavioral characteristic over time. The horizontal axis represents time, and the vertical axis shows the range of perturbations. As this visual representation reveals the particular context in which behavioral characteristics are perturbed, educators can examine potential causes for dropout across various student groups.

[2] https://www.wikipedia.org/

The **feature temporal comparison view** (Figure 8c) presents students' original behavioral characteristics and the distribution of behavioral characteristics in counterfactual explanations across different time series. The layout consists of nine histograms, each representing a behavioral characteristic, where the horizontal axis shows the range of the behavioral characteristic's values, and the vertical axis represents its frequency. The red bars indicate the original behavioral characteristic, while the green bars represent the behavioral characteristic's values in each counterfactual explanation.

The **counterfactual explanation view** (Figure 8d) aims to indicate the direction and extent of modifications in the original behavioral characteristics during the search for counterfactual explanations. This allows instructors to develop targeted interventions for students with diverse behavioral characteristics. This layout presents nine axes corresponding to each behavioral characteristic, with four green bars indicating the value ranges on the right. The change in behavioral characteristics between different groups through counterfactual explanation is represented by arcs with arrows, illustrating the transition from original values to the counterfactual instance values. If the original and counterfactual values are in the same group, they are represented by a circle, and the circle's size encodes the frequency of the characteristic's perturbation within that group.

The **instance view** (Figure 8e) allows exploration of correlations between behavioral characteristics across different weeks at the individual level. This view presents a parallel coordinates chart with nine axes corresponding to the nine behavioral characteristics, where each curve represents a student's counterfactual instance. Users can select the range of different behavioral characteristics to explore a small number of students of interest and analyze their relationships with other behavioral characteristics based on the selected constraints.

Deng et al. (2019) developed *PerformanceVis* to analyze the performance of 949 students during the fall semester of 2018 in a chemistry course. The system provides the analysis of students' admissions, course performance data, homework tasks and exam question formats.

PerformanceVis includes four main visualizations. The **Overall Exam Grade Path (OEGP)** is represented through a *Sankey* diagram (Figure 9a), which illustrates the variation in grade distribution (from A to F) throughout the semester, considering four different types of exams. A flow between two nodes incorporates students who received the corresponding grades in both exams. The height of this flow is proportional to the number of students it contains. This layout allows users to quickly observe which grades are likely or unlikely to occur, considering a specific behavioral pattern, motivating students to modify their learning patterns to behaviors that are more likely to correlate with better grades.

The **Detailed Exam Grade Path (DEGP)** helps identify student(s) who exhibited interesting changes, showing their demographic and academic characteristics in more

Figure 8 – DropoutVis interface: (a) Temporal Circle View, (b) Windmill View, (c) Feature Temporal Comparison View, (d) Counterfactual Explanation View, and (e) Instance View. Images sourced from: (ZHANG et al., 2022).

details employing a parallel coordinates layout (Figure 9b). To represent the autocorrelation between exam questions (i.e., the correlation between any two questions within the same exam), a force-directed graph was used in the Detailed Exam Item Analysis (DEIA) layout (Figure 9c). Each node of the graph corresponds to an exam question, and the thickness of the links indicates the strength of the correlation: thicker lines represent stronger correlations between nodes. The color of each node indicates its level of difficulty or topic. In each visualization, nodes with stronger correlations are grouped, indicating similar or related topics. This representation helps instructors validate exam designs, identifying questions that should not be included in the set.

The **Overall Exam & Homework Analysis (OEHA)** technique (Figure 9d) uses a collapsible radial tree to simultaneously visualize the correlations between exam and homework questions. In the representation, there are distinct levels: the broadest level includes questions from a specific exam displayed in the DEIA. The subsequent level displays questions related to this exam in comparison to three others. The most detailed level shows homework questions, revealing connections between exam and homework questions. The technique also applies colors to represent levels of difficulty or topics. DEIA and OEHA layouts are coordinated to each other.

BlockLens (TSUNG et al., 2022) is a visual analysis system designed to help instructors and platform owners to understand the behavior of primary and secondary school students as they solve basic coding exercises using block-based programming (*Scratch*). The system allows grouping students based on their progress and performance in exercises, identifying

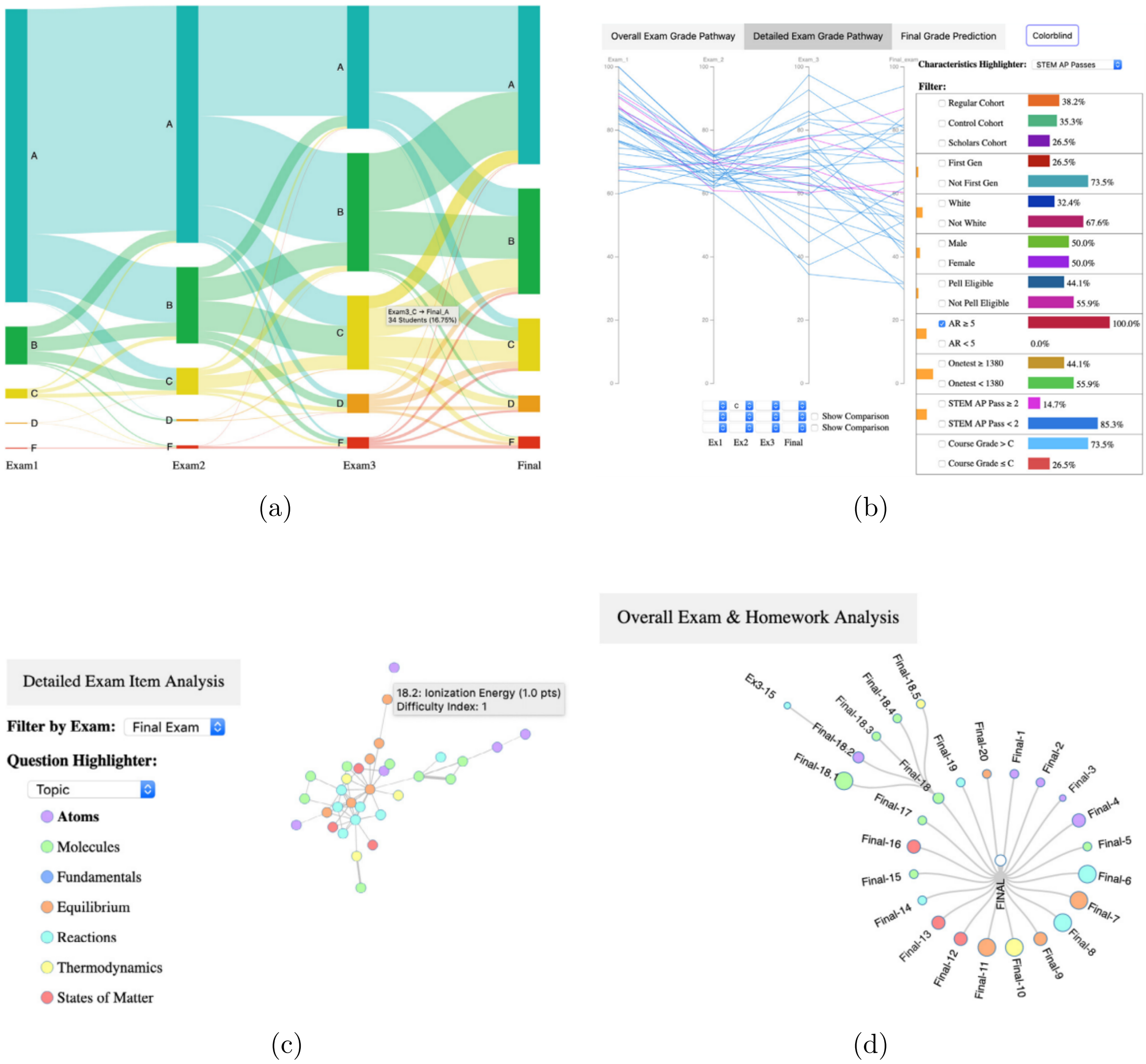


Figure 9 – PerformanceVis techniques: **(a)** OEGP, **(b)** DEGP, **(c)** DEIA, **(d)** OEHA. Images sourced from (DENG et al., 2019).

common patterns in exercise solving, and providing analyses at different levels of detail. It offers a comprehensive overview of all students, as well as a detailed analysis of an individual student's behavior and performance through snapshots that classify success (*checkpoints*) or failure (*warning signs*) during exercise solving.

The system provides four visualizations. The **Exercise Selection View** (Figure 10a) enables users to select an exercise for interaction in other visualizations. Three key statistics are provided for each exercise: the success rate, displayed as a progress bar; the average number of steps, represented by rectangles where each rectangle symbolizes one step; and the average elapsed time, shown as a clock-inspired unit visualization.

The **Student View** (Figure 10b) illustrates how students are distributed based on performance metrics such as *checkpoints* and *warning signs*. The Student *Checkpoint-Warning* Chart is a unique visualization that helps identify outlier students within a group, based on the number of identified snapshots. The chart plots the count of *warning* *signs* (y-axis) against the count of *checkpoints* (x-axis), separating students who correctly solved the exercise (represented by green points) from those who did not (represented by orange points). The size of each point encodes the number of steps taken by the student. The chart allows the selection of a single or multiple students. Below the *Checkpoint-Warning* Chart, two distribution graphs are included: one for the elapsed time from the start to submission and another for the number of completed steps. Both distributions use Kernel Density Estimation curves to compare students who solved the exercise correctly with those who did not.

The **Path Summary View** (Figure 10c) is a Sankey-based representation depicting how students gradually build their code. Each node represents a snapshot of the code, and the connections show transitions from one step to another. The width of the connections reflects the frequency of transitions. Snapshots are organized in columns based on the number of blocks. Meanwhile, other snapshots are shown compactly, displaying only the number of blocks they contain. The type of each snapshot is indicated by the fill and outline of the rectangle representing the code area. Green snapshots are marked as *checkpoints*, orange snapshots as *warning signs*, and gray snapshots represent other types. Each snapshot visualization (except "others") incorporates a step count distribution bar. This bar is a horizontally stacked bar chart that illustrates the proportion of times each snapshot appeared in each quartile of steps within a student's snapshot sequence.

Finally, the **Sequence View** (Figure 10d) presents complete sequences of snapshots, as well as the events connecting each snapshot, for any student selected in the Student Visualization (Figure 10b) or the Path Summary Visualization (Figure 10c). The objective is to display the complete problem-solving paths of individual students, rather than merely summarizing multiple students' data.

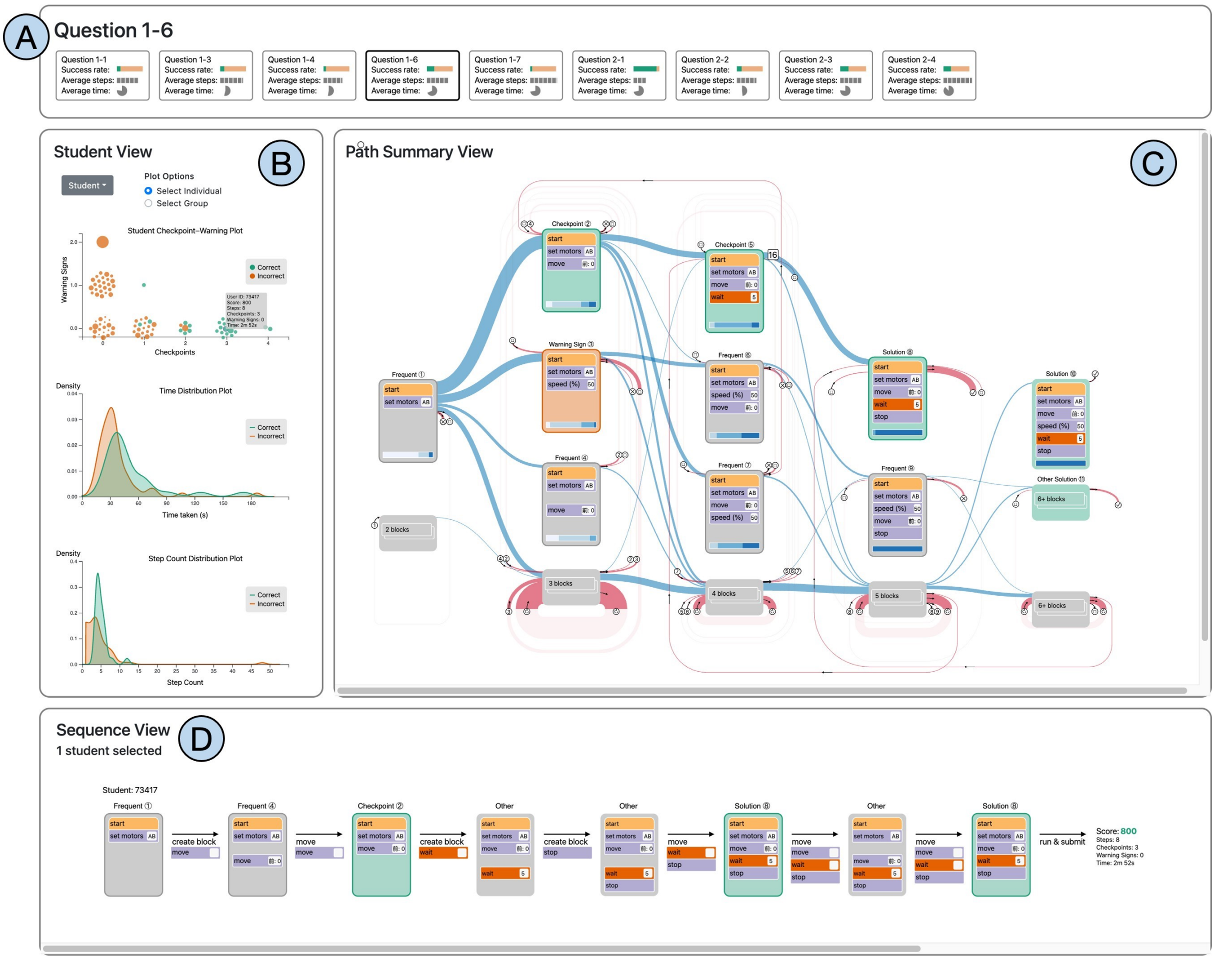


Figure 10 – *BlockLens* interface: (a) Question Selection View, (b) Student View, (c) Path Summary View, and (d) Sequence View. Images sourced from: (TSUNG et al., 2022).

LearningViz is an interactive learning analytics dashboard developed by Pei et al. (2024) to help instructors identify student performance across students and classes. It contains three, listed as follows:

1. The **Student Overall Performance Analysis Module** (Figure 11 (a)), which provides a comprehensive understanding of students' learning in the course from three perspectives:

   ❏ A bar chart that displays the grade distribution in the course, showing an overview of whether students' learning outcomes align with initial expectations;

   ❏ A tree diagram (Course Structure Analysis), which presents the course structure with students average performance per week and per assessment activity to identify factors affecting performance, as well as to highlight the most critical topics and weeks;

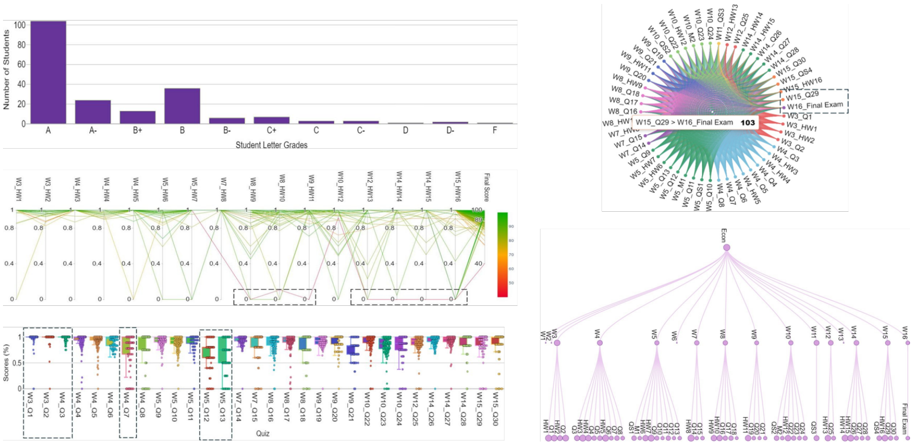
- A parallel categories diagram (Performance Pathway Analysis) which shows how each student's performance evolves over the weeks, considering different types of assessments.

2. The **Group Performance Analysis Module** (Figure 11 (b)) examines performance gaps across different groups and identifies factors contributing to these gaps in two dimensions:

   - Performance Group Analysis categorizes students into different performance groups, such as the *Thriving Group*, *nonThriving Group*, and *DFW Group*. This approach utilizes density estimation within learning performance groups to identify student performance patterns and estimate their final grades;

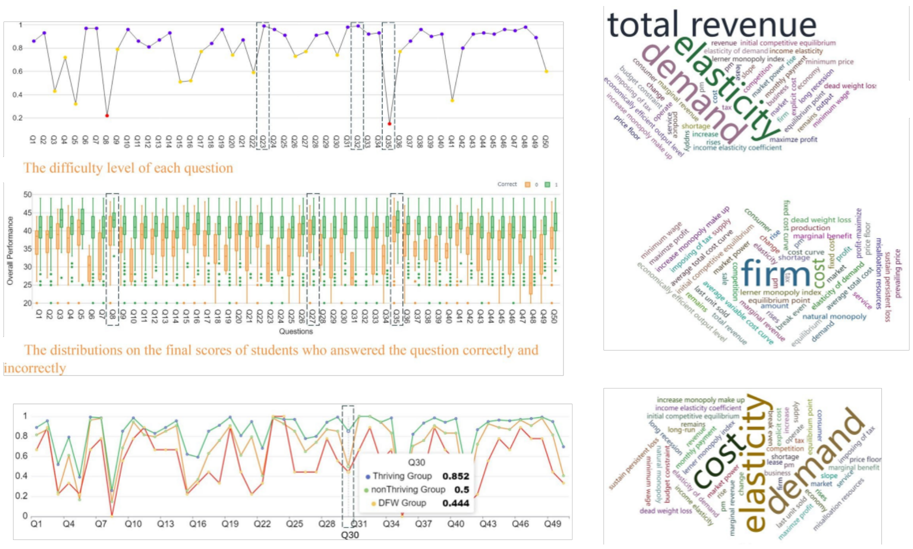
   - Assignment Performance Gap Analysis, a heatmap used to show differences in average performance between groups for each assessment activity, allowing instructors to identify the most challenging topics and evaluate the effectiveness of the activities in discriminating between groups. Line charts complement this analysis by showing the evolution of performance gaps between groups across different assessment.

3. The **Final Exam Item Analysis Module** (Figure 11 (c)) evaluates the quality of exam questions and identifies strategies for closing performance gaps. Based on the understanding of performance differences among student groups, this module provides a detailed analysis of each group's performance on each final exam question. The goal is to help instructors identify which questions were either too easy or too difficult, by analyzing the distribution of final grades for students who answered each question correctly or incorrectly. The line chart at the top represents the questions numbered from 1 to 50 on the x-axis and the performance scores on these questions on the y-axis, allowing instructors to identify which questions had the highest and lowest performance. The chart below shows the distribution of final performance scores for students who answered each question correctly and incorrectly, helping instructors identify questions that had little impact on overall final exam performance. The line chart at the bottom illustrates the performance gaps for each question across the three performance groups, enabling instructors to identify which questions contribute most significantly to performance gaps between different groups. Additionally, a word cloud presents the keywords associated with the 10 most difficult questions for each group, helping instructors design better instructional materials to address performance gaps.



(a)

(b)



(c)

Figure 11 – *LearningViz* views: (a) overview of the overall performance analysis module, (b) analysis of different performance groups, (c) final exam item analysis and performance gaps targeted interventions. Images sourced from: (PEI et al., 2024).

To analyze the interaction patterns of 48 students with ChatGPT in a master's-level data visualization course over one semester, Chen et al. (2025) developed a visual analysis system called *StuGPTViz*, which tracks and compares temporal patterns in students' prompts and the quality of ChatGPT's responses across different scales. *StuGPTViz* is structured into three main components.

The **Filter View** (Figure 12 (A)) serves as the starting point for analysis, allowing instructors to filter students and tasks of interest. This view provides an overview of learning activities and student profiles through interactive visualizations, making it easier to analyze grade distribution, task difficulty, and students' prior experience.

The **Pattern View** (Figure 12 (B)) allows for an in-depth analysis of interaction patterns between students and ChatGPT. This view provides both a macro-level summary of interaction trends and a detailed catalog of specific patterns, enabling comparisons across groups and task types. Additionally, it includes a visualization of student-ChatGPT interaction flows, helping to analyze prompt-response dynamics and key quality metrics.

The **Detail View** (Figure 12 (C)) allows instructors to explore individual student responses and task details. By selecting a student and a task, they can examine task descriptions and review raw student-ChatGPT interactions. This view helps analyze how interaction patterns influence learning outcomes, offering flexible filters to compare different student groups or task types for a more comprehensive analysis.

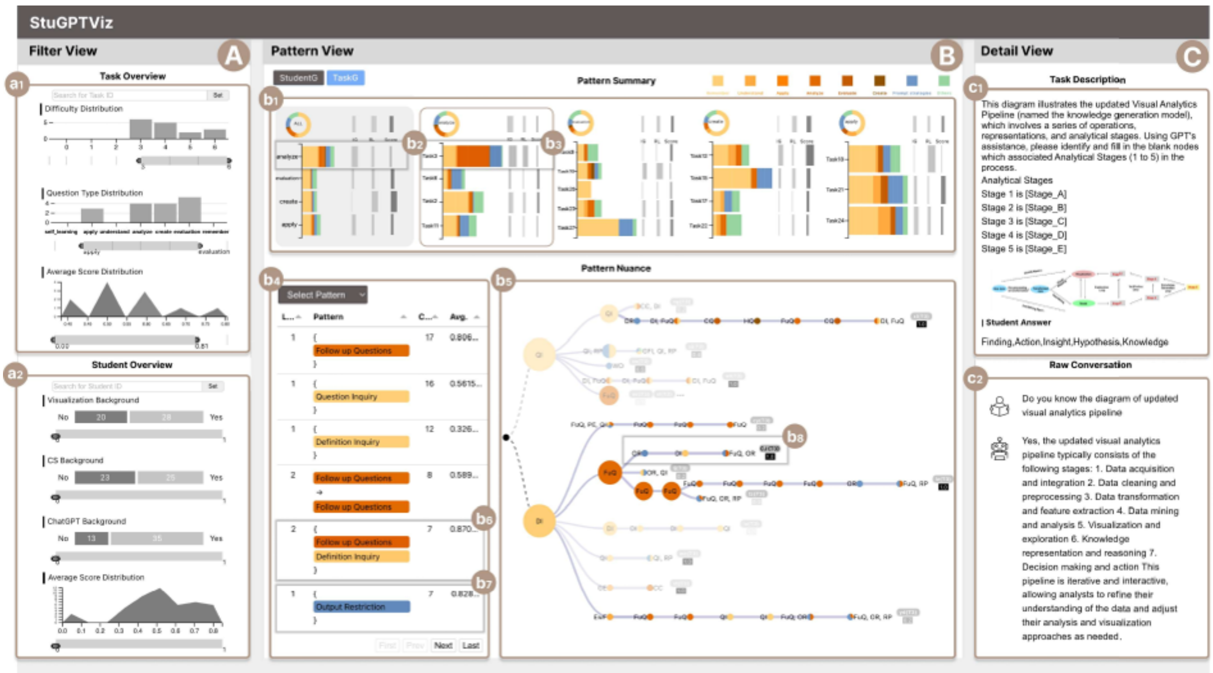


Figure 12 – *StuGPTViz* interface: (A) The Filter View. (B) The Pattern View. (C) The Detail View. Images sourced from: (CHEN et al., 2025).

2.6 Final Considerations

Studies applying Information Visualization techniques in education highlight the potential of data-driven approaches to understanding student behavior, performance, and risk factors. Despite significant advancements, challenges remain in enhancing the visual analysis of educational data to provide deeper insights into student engagement and learning outcomes.

One of the main challenges lies in the complexity of educational data, which encompasses a wide range of variables, including academic performance, attendance, and socioeconomic factors, as well as subjective elements like motivation and engagement with specific topics. Many existing systems focus on isolated aspects, such as dropout prediction or performance tracking, often relying on external or summarized internal data. This limitation restricts a comprehensive understanding of how students interact with the curriculum over time.

Our approach addresses this gap by integrating detailed internal data—such as academic records, attendance, and performance across multiple semesters—into a visual analysis framework. By enabling a topic-level exploration of student engagement, our method provides a more granular view of how specific curriculum components influence overall learning outcomes. Moreover, examining relationships between course modules and shared topics can uncover underexplored patterns, leading to more precise and insightful analyses that support data-driven improvements in educational systems.

Chapter 3

Proposal

We believe that understanding students' performance in course modules throughout their academic journey can provide strategic insights for improving education and supporting more effective decision-making in educational settings. In this sense, Information Visualization techniques can serve as an essential tool to facilitate this process, enabling the interactive exploration of academic data and the identification of relevant patterns that help explain students success or even dropout. Student performance is influenced by several factors, such as class attendance, the topics covered in each module, the impact of multiple attempts in the same module, and the relationship between these aspects and final grades.

To ensure that the analysis is efficient and relevant to education experts, we first conducted an analysis of existing Data Visualization literature in visual strategies focused on educational analysis to identify research gaps (MANDINACH; 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. 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 discussion, 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 modules' structure, in terms of topics, on student performance. The system should provide the analysis of the relationship between the module's topics and student performance, including the possibility of identifying topics that positively or negatively affect academic progress;

R3-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 such behavior on their performances;
R4-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 distille the following analytical tasks, that guided the framework design:
T1-Analyze the performance of individuals and groups of students in course modules, considering their attendance and obtained grades. The system must provide a detailed analysis of individual students' or groups' performance, taking into account their grades and class attendance. This facilitated the identification of performance patterns and areas requiring attention (**R1, R3, R4**).
T2-Track students' multiple attempts in course modules. It is interesting to track multiple attempts in specific modules, displaying how student performance evolved in each attempt and facilitating the analysis of progress and challenges (**R3**).
T3 Investigate how module topics impact students' performance in these modules. The system must provide the investigation of the impact of module topics on students' performance, helping to identify which topics were critical for academic success or failure (**R2**).
T4-Identification of at-risk groups. This task must assist in the identification of students who shared similar behavioral or academic patterns, facilitating the development of intervention strategies targeted at at-risk groups (**R1, R2**).

Figure 13 presents the architectural diagram of ***LearnVis***. The data is obtained from a set of specific queries built on the original dataset. These queries feed the Query Data Table, which organizes the information required for each system visualization. In the case of the *Student Behavior View*, the data resulting from its corresponding query undergoes a dimensionality reduction technique before generating the final layout. The other visualizations use their specific queries directly to compose their respective layouts. Each of these visualizations is detailed in the following sections, which also describe how the defined tasks were addressed to provide an effective visual analysis tailored to the needs of education experts.

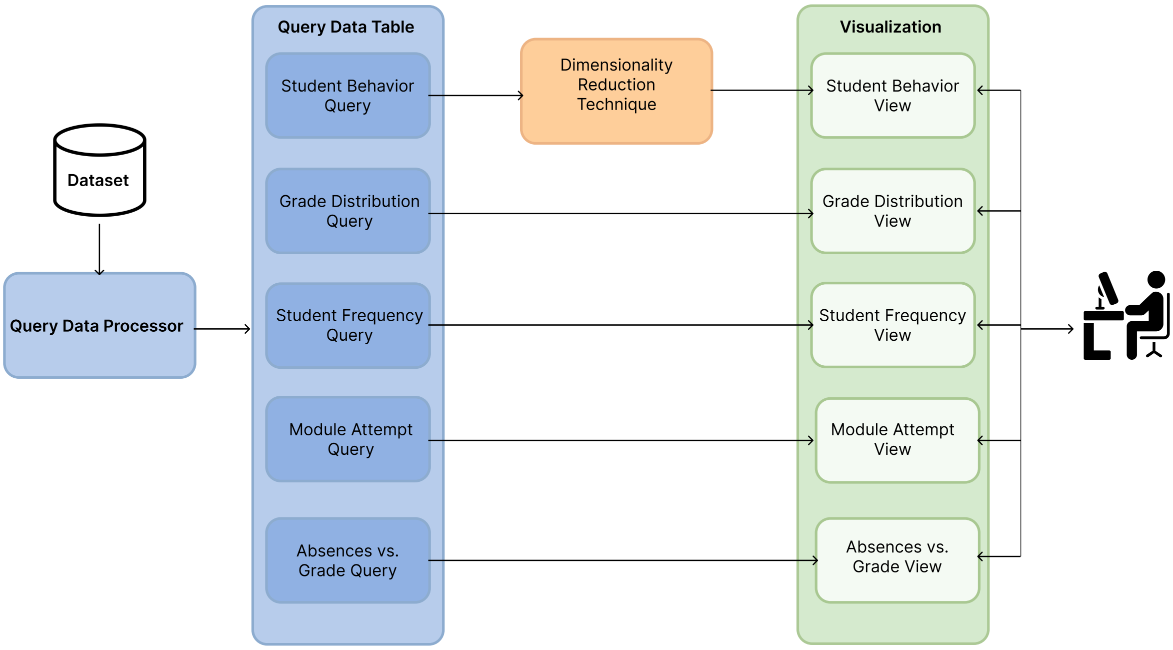


Figure 13 – ***LearnVis*** architectural diagram.

3.1 *LearnVis*: System Description

In this section, we present a new visual analysis system called ***LearnVis***, which assists instructors to explore, analyze, and understand 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 14 shows the main interface of the system considering a specific module, with no selected students. 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 14F). Each layout is detailed as follows.

3.1.1 Student Behavior View

The Student Behavior View (Figure 14A) addresses tasks T1, T2 and T4, and was designed to identify similar behavior patterns in student groups regarding academic performance, as well as students with specific or strategic behaviors. The data view used by the layout is shown in Table 2.

Each student is represented by their sequence of attendance and absences for all topics covered in a module, encoded in the `ATTENDANCE_FLAG` feature, organized into a binary vector of k positions, where k corresponds to the number of topics in the module. In this vector, the value 0 indicates absence, and the value 1 indicates attendance. The t-SNE dimensionality reduction technique was applied due to its ability to capture

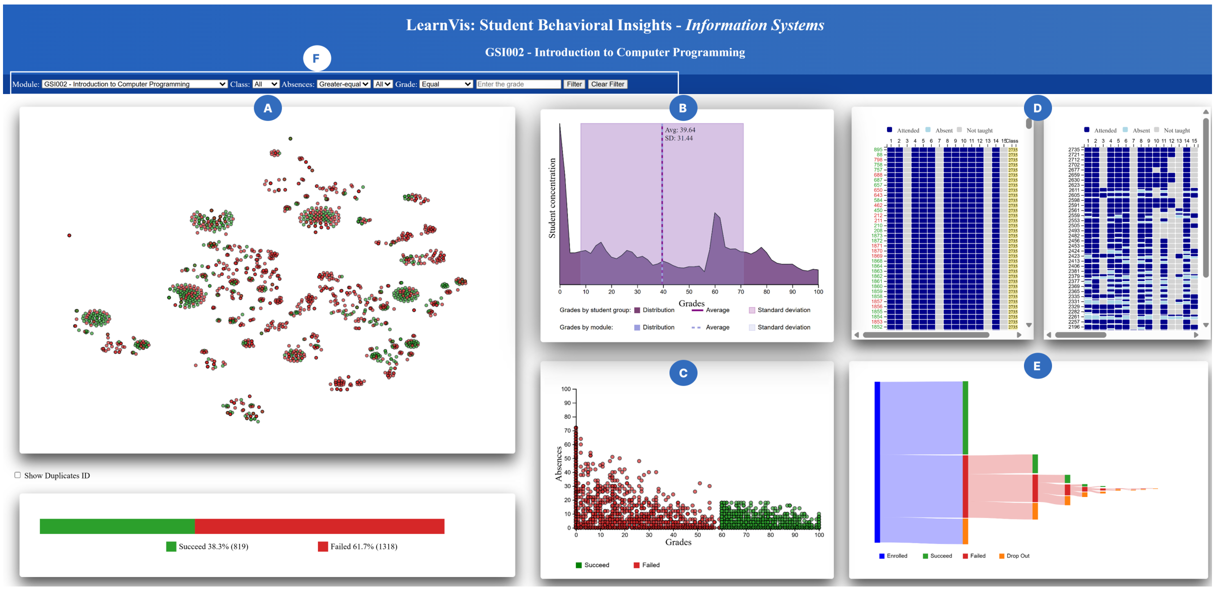


Figure 14 – ***LearnVis*** main screen, with a set of coordinated views. (a) Student Behavior View, (b) Grades Distribution View, (c) Absences vs. Grades View, (d) Student Frequency View, (e) Module Attempt View and (f) Filters.

Table 2 – Data view - Student Behavior View

ENROLLMENT_ID	integer
CLASS_ID	integer
MODULE_CODE	varchar
FINAL_GRADE	integer
FINAL_STATUS (SUCCEEDED OR FAILED)	varchar
REPEAT_COUNT	integer
TOPIC_ID	integer
ATTENDANCE_FLAG	integer

non-linear relationships between instances, enabling the mapping of each student into a two-dimensional layout. Prior to projection, the features `ENROLLMENT_ID`, `CLASS_ID`, `MODULE_CODE`, `FINAL_GRADE`, `FINAL_STATUS`, `REPEAT_COUNT`, and `TOPIC_ID` were removed, as they were not relevant to the dimensionality reduction process. The resulting layout places students with similar attendance patterns close to each other, while students with different behaviors are positioned farther apart in the projected space. The algorithm was configured to generate two components (`n_components=2`) with a fixed random seed (`random_state=42`) to ensure reproducibility. The projection used the default t-SNE distance metric, Euclidean distance, applied directly to the frequency-related variables. No explicit normalization or scaling was performed, as the variables already consisted of discrete values indicating attendance or absence. The resulting two-dimensional embedding was then merged back with the original metadata fields, allowing the visualization to retain information about each student's identifier, class, course code, final status, final grade, and number of attempts, thus enabling richer contextual analyses.

This visualization is presented as a scatterplot, in which each circle in the layout repre-

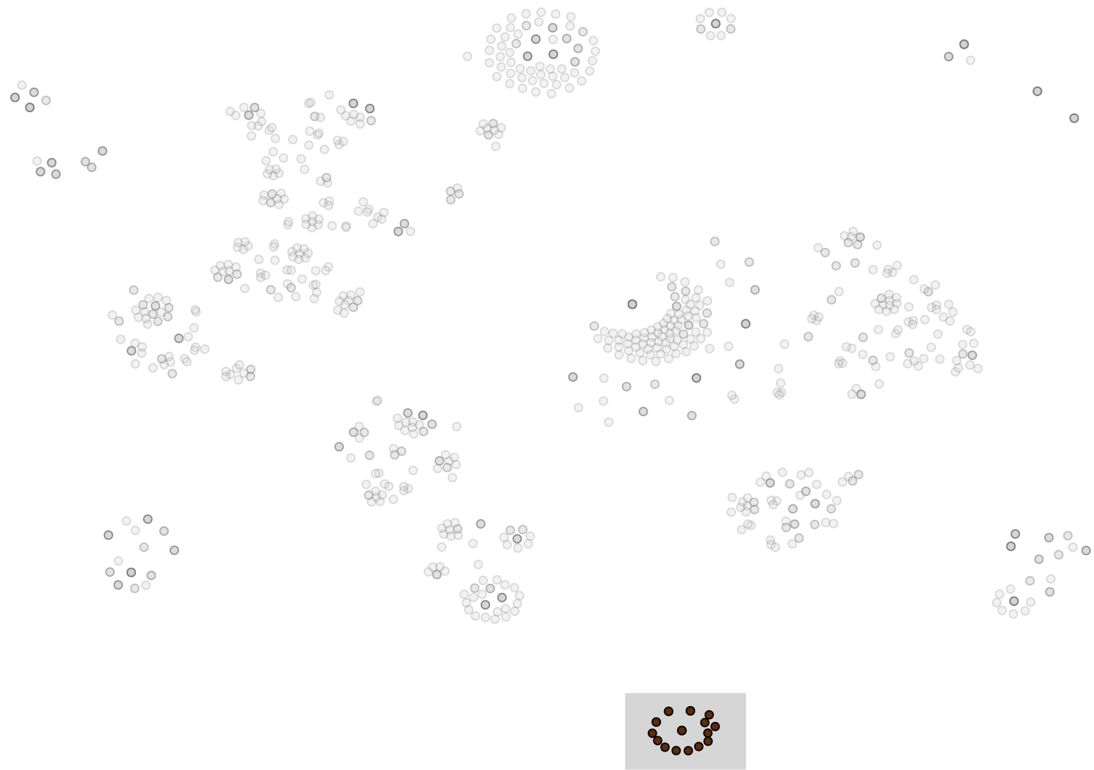


Figure 15 – Student Behavior View - Selected group of students.

sents an individual student. Students who failed are shown in red, while those who succeed 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 positioned in different areas of the layout depending on their different behavior. Users can select individual or groups of students for more detailed analysis. When a selection is performed, all other views are automatically updated to reflect this selection. The selected students are highlighted in brown, whereas the remaining ones are colored gray (Figure 15). Additionally, it is possible to identify students who made multiple attempts in a module. To do this, a checkbox located below the layout can be selected to highlight these students. If a selection includes a student in this situation, the corresponding circles will be uniformly colored and numbered according to the number of attempts. When a student succeeds in the module, the circle is replaced by a star-shaped glyph to facilitate his/her identification (Figure 16). Moreover, when a use selects a specific group of students, a stacked bar chart is displayed to show the success and failure rates to reflect the information in 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.

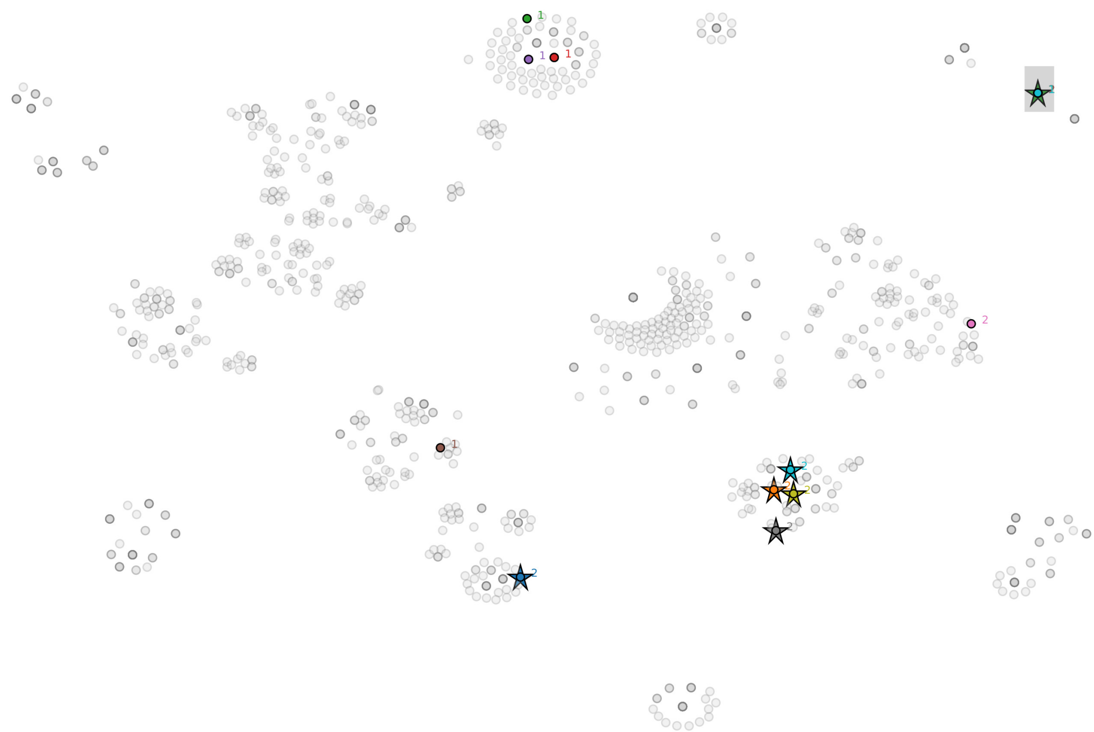


Figure 16 – Student Behavior View - Visualization of students' attempts.

3.1.2 Grades Distribution View

The Grade Distribution View (Figure 14B) supports tasks T1 and T4. The data view adopted for this layout is described in Table 3. This layout applies Kernel Density Estimation (KDE) (SILVERMAN, 2018) technique to represent how student grades are distributed within a specific module. This method produces a smooth curve shown in Figure 17, that summarizes the frequency of grades across a continuous range, facilitating the identification of trends and concentration patterns in student performance. In this visualization, the horizontal axis represents the grade values (from 0 to 100), while the vertical axis indicates the relative concentration of students for each grade. A light purple curve represents the overall grade distribution in the module. When a group of students is selected, their grade distribution is displayed on top of the overall distribution curve in dark purple. Vertical dark purple lines indicate the average grades of the selected group, while a light purple dotted line represents the overall average for all students in the module. A rectangular shaded area is drawn around each vertical line to represent the standard deviation. The light purple region corresponds to the overall standard deviation, while the dark purple region reflects the deviation within the selected group.

This visual encoding highlights grade variability and supports comparisons between the selected group and the broader population. When hovering over the vertical lines, a tooltip displays the exact average and standard deviation of the grades for both the selected group and the overall cohort.

Table 3 – Data view - Grade Distribution View

ENROLLMENT_ID	integer
CLASS_ID	integer
MODULE_CODE	varchar
FINAL_GRADE	integer
FINAL_STATUS (SUCCEEDED OR FAILED)	varchar

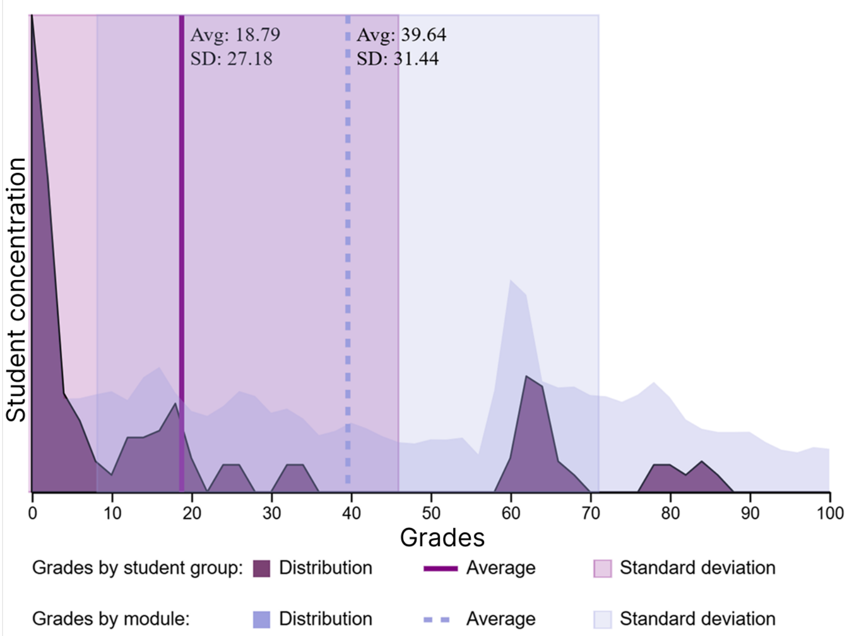


Figure 17 – Grades Distribution View with selected students.

3.1.3 Absences vs. Grades View

The Absences vs. Grades View (Figure 14C) contributes to tasks T1 and T4, and its goal is to analyze the correlation between the number of absences and final grade. Table 4 presents the data view used in this layout.

Table 4 – Data view - Absences vs. Grades View

ENROLLMENT_ID	integer
CLASS_ID	integer
MODULE_CODE	varchar
FINAL_GRADE	integer
FINAL_STATUS (SUCCEEDED OR FAILED)	varchar
NUMBER_OF_ABSENCES	integer

The layout, shown in Figure 18, displays a scatter plot in which the horizontal axis maps the grades and the vertical axis maps the 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.

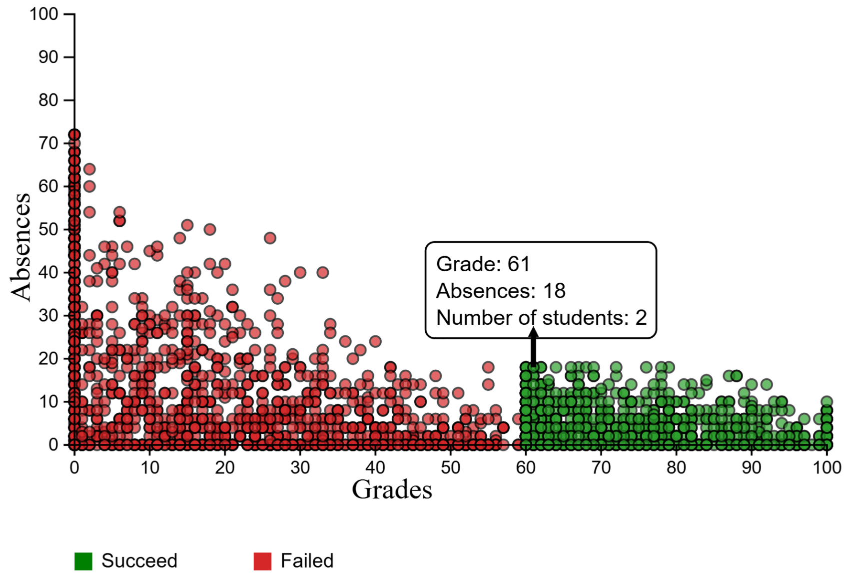


Figure 18 – Grades Distribution View with tooltip.

Hovering over a circle triggers a tooltip displaying the number of students represented by the circle, their corresponding grades, and the number of absences.

3.1.4 Student Frequency View

The Student Frequency View (Figure 14D) helps address tasks T1, T3 and T4 and represents student attendance on module topics, as well as if that topic was taught in a specific semester. The data view used by the layout is structured as Table 5.

Table 5 – Data view - Student Frequency View

ENROLLMENT_ID	integer
CLASS_ID	integer
MODULE_CODE	varchar
FINAL_GRADE	integer
FINAL_STATUS (SUCCEEDED OR FAILED)	varchar
TOPIC_ID	integer
TOPIC_NAME	varchar
ATTENDANCE_FLAG	integer

The resulting layout is shown in Figures 19 and 20, and organizes the selected students in a table, in which each row represents 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 (Figure 19) and another that summarizes the information by class

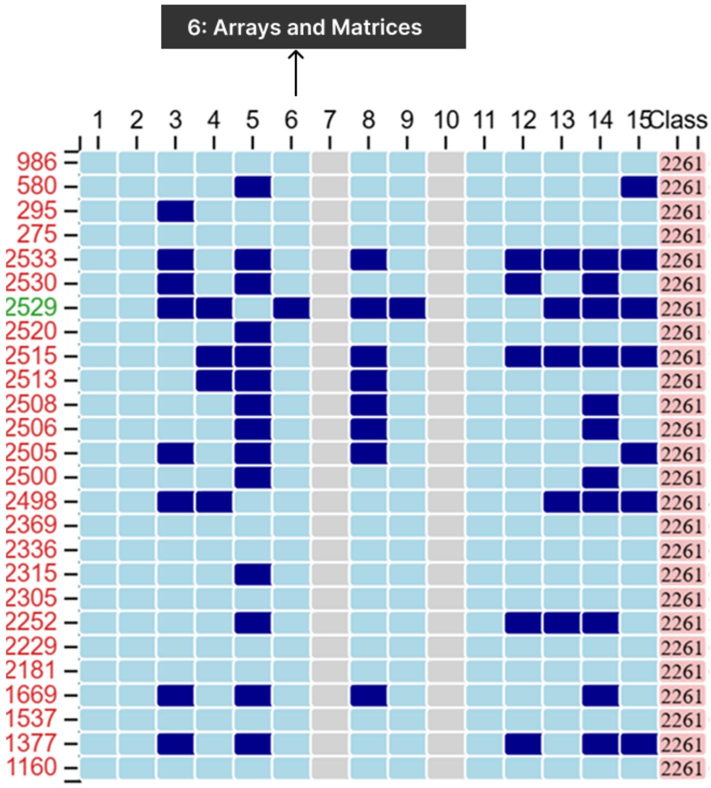


Figure 19 – Student Frequency View - shows individual student behavior.

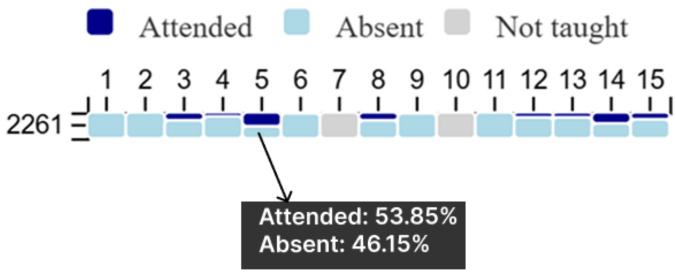


Figure 20 – Student Frequency View - spresents a summary by class.

(Figure 20). 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. In both versions, when hovering over the column numbers that represent the topics, the system displays a tooltip with a detailed description of the topic.

3.1.5 Module Attempt View

Module Attempt View (Figure 14E) contributes to tasks T1, T2 and T4 and illustrates the students' trajectory in terms of repeated enrollment attempts in the same module. The data view used by this layout is shown in Table 6.

Table 6 – Data view - Module Attempt View

ENROLLMENT_ID	integer
CLASS_ID	integer
MODULE_CODE	varchar
FINAL_STATUS (SUCCEEDED OR FAILED)	varchar

The layout uses a Sankey diagram (RIEHMANN; HANFLER; FROEHLICH, 2005), shown in Figure 21, where 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. When hovering over the transitions between bars, a tooltip shows the percentage and number 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.

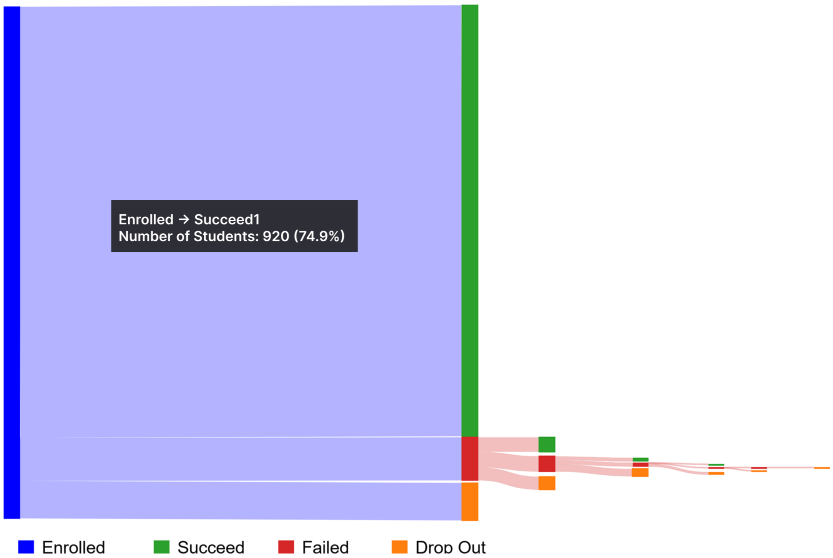


Figure 21 – Module Attempt View with tooltip

In summary, the layouts implemented in ***LearnVis*** were designed to meet the system requirements and support the execution of the defined analytical tasks. Table 7 presents the relationship between each requirement, the corresponding tasks, and the layouts that address them, providing a clear overview of how the system's components align with its objectives.

Table 7 – Relationship between system tasks, requirements, and ***LearnVis*** views

Tasks	Requirements	Views
T1 – Analyse the performance of individuals and groups of students in course modules, considering their attendance and obtained grades	**R1** - Utilize available internal data from student records to analyze their performance in course modules, **R3** - Provide the analysis of student performance in multiple attempts of the same module and **R4** - Provide complementary information for performance analysis	Student Frequency, Grades Distribution, Absences vs. Grades, Student Frequency, Module Attempt
T2 – Track student's multiple attempts in course modules	**R3** - Provide the analysis of student performance in multiple attempts of the same module	Student Behavior, Student Frequency, Module Attempt
T3 – Investigate how module topics impact students' performance in these modules	**R2** - Explore and evaluate the impact of modules' structure, in terms of topics, on student performance	Student Frequency
T4 – Identification of at-risk groups	**R1** - Utilize available internal data from student records to analyze their performance in course modules and **R2** - Explore and evaluate the impact of modules' structure, in terms of topics, on student performance	Student Behavior, Grades Distribution, Absences vs. Grades, Student Frequency, Enrollment Progress

3.2 Implementation Details

The development of ***LearnVis*** involved multiple technologies to enable efficient extraction, processing, and visualization of academic data. As illustrated in Figure 13, the system was structured into several components: the dataset; the t-SNE generation; and finally, the creation of the system's visual layouts. Below, we detail each of these phases and the technologies used.

3.2.1 Dataset

Once stored in the *SQL*[1] Server database, SQL queries were used to extract relevant data subsets, perform aggregations, and structure the information for analysis. These

[1] <https://learn.microsoft.com/en-us/sql/?view=sql-server-ver17>

queries are executed only once, and the resulting data is subsequently loaded into the system to ensure fast and responsive interaction. This preprocessing strategy significantly improves performance by reducing computational overhead at runtime. To complement the data processing, we used the *Pandas*[2] library in Python, an open-source tool widely adopted for data analysis and manipulation of structures such as DataFrames (MCKINNEY et al., 2010).

3.2.2 t-SNE

For the creation of the *Student Behavior View* layout, after generating the data view 3.2.1, we applied dimensionality reduction using the *Manifold*[3] library, specifically the t-SNE method. t-SNE is a non-linear machine learning technique widely used for visualizing high-dimensional data in lower-dimensional spaces while preserving the similarity relationships between data points (MAATEN; HINTON, 2008).

3.2.3 Visualization

With the data structured and processed, we developed interactive visualizations using HTML, CSS, and JavaScript. The core visualization engine was implemented with the *D3.js*[4] library. D3.js is a powerful JavaScript library for manipulating data-driven documents, enabling the creation of dynamic and interactive charts with support for data-driven transformations (BOSTOCK; OGIEVETSKY; HEER, 2011).

[2] <https://pandas.pydata.org/>
[3] <https://scikit-learn.org/stable/modules/manifold.html>
[4] <https://d3js.org/>

Chapter 4

Experimental Results

In this chapter, we present the results derived from the application of the proposed visual analysis methodology through the ***LearnVis*** system, considering different scenarios related to academic data exploration. Initially, we describe the procedures for data acquisition and preprocessing, which were essential to prepare the dataset for analysis. Subsequently, we discuss the results based on a set of evolution experiments, highlighting how each layout contributes to the analytical process.

4.1 Data Understanding and Preprocessing

We utilized a data repository provided by the Information Technology department at UFU, in which all students' personal information were anonymized. It is important to note that, although the presented data have undergone anonymization steps, it reflects real academic trajectories, allowing relevant inferences about the learning process. 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 (Resolution 46/2022 – CONGRAD) at the UFU, in Chapter II – On Evaluation, Art. 127, for a student to succeed, it is necessary to achieve a grade of at least 60 academic performance points and 75% attendance in academic activities, being considered as failed otherwise. 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 8, and the preprocessing is described as following.

Feature Engineering: To enhance the dataset and support the visualizations and analyses required by ***LearnVis***, several new features were engineered. The attributes

Table 8 – Description of the initial features used in ***LearnVis*** (original variable names translated from Portuguese to English for clarity).

Type	Initial features	Description
Numeric	CLASS_ID	Class identification
	ENROLLMENT_ID	Masked person identifier
	NUMBER_OF_ABSENCES	Number of absences
	NUMBER_OF_YEAR	The year the module was taught
	PERIOD_CODE	Course period identification
	TOPIC_ID	Topic identification
Nominal	COURSE_CODE	Course identification
	FINAL_STATUS	Student's final status (Succeeded, Failed)
	MODULE_CODE	Code of module
	MODULE_NAME	Name of module
	PERIOD_NAME	Course period description
	TOPIC_DESCRIPTION	Topic description
Scalar	FINAL_GRADE	Student's final grade (0 to 100)
Temporal	LECTURE_DATE	Lecture date

`TOPIC_ID` and `TOPIC_NAME` were extracted from the original `TOPIC_DESCRIPTION` field and are now part of the `TOPIC` table. These fields provide a standardized way to reference the topics covered in each module, enabling more precise analyses of student engagement with specific topics and facilitating the identification of possible correlations between attendance and performance. The attributes `ATTENDANCE_FLAG` and `REPEAT_COUNT` were generated in derived views to support analytical tasks. The `ATTENDANCE_FLAG` feature was created to standardize attendance tracking: 0 represents absence, 1 represents attendance, and 2 indicates that the topic was not taught. This flag was derived from the combination of `LECTURE_DATE` and `TOPIC_ID`, allowing the system to track whether students attended each specific class session and topic. The `REPEAT_COUNT` feature 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 feature provides insights into repeated attempts. Table 9 presents a summary of the new features added to the dataset, complementing the original fields shown in Table 8.

Table 9 – Description of the new features used in ***LearnVis***.

Type	New Features	Description
Numeric	ATTENDANCE_FLAG	Attendance identifier for topics
	REPEAT_COUNT	Number of repetitions in the module
	TOPIC_ID	Topic identifier
Nominal	TOPIC_NAME	Topic description

Data Restructuring: To address the lack of standardization in topic descriptions (`TOPIC_DESCRIPTION`), we created a standardized topic table for each module. This restructuring was based on the official curriculum structure and involved manually mapping the various free-text topic descriptions provided by instructors to the corresponding standardized topics. Since instructors could enter different terms to refer to the same content, this mapping ensured consistent references across semesters.

Data Cleaning: During the data cleaning phase, we addressed missing values in class topic records. In cases where instructors did not specify the topic taught, we filled in these gaps by inferring the most likely topic based on data from previous semesters—considering the common sequence in which topics are usually presented. This approach ensured that missing entries were filled in a way that aligned with the typical instructional flow of each module.

The database was structured to store and manage student-related information, and is shown in Figure 22. The database schema consists of multiple tables, where the columns that were not used are shown in black. Additionally, the tables received from the institution are represented in blue, while the table developed to complement the analyses is shown in orange.

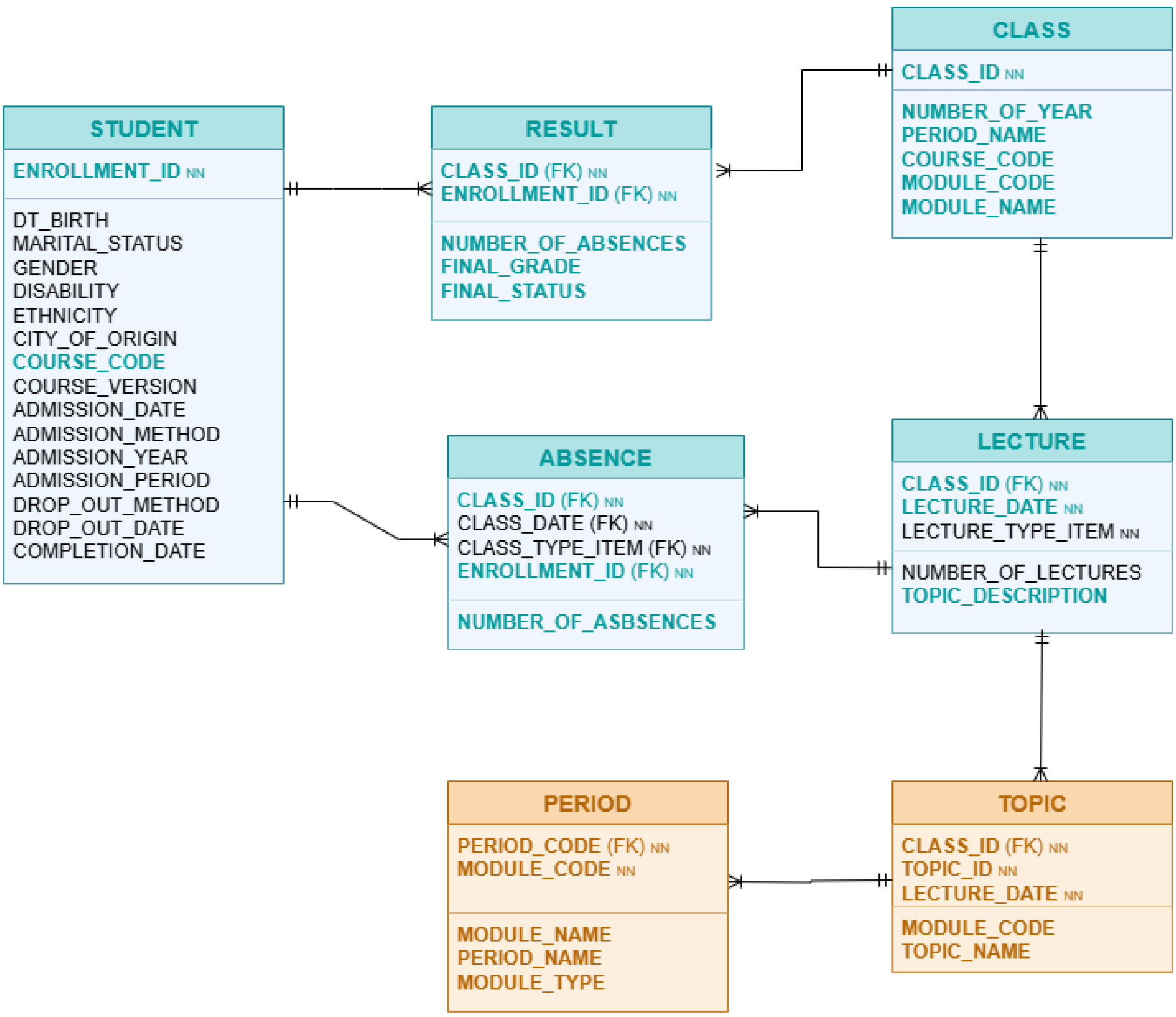


Figure 22 – Database diagram.

4.2 Evolution Experiments

This section presents the application of ***LearnVis*** through a set of evolution experiments that illustrate distinct scenarios from our data repository. Initially, we explore the general layout visualizations to investigate patterns of student performance and behavior throughout the modules of the course. Then, we refine the analysis by examining the relationships between attendance, academic performance, and the topics covered in class. We also discuss how the system represents students with multiple attempts, as well as failure and dropout cases in modules, and how this information is integrated into the different visual components. During the meetings described in Section 3, the experts reported the 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.

4.2.1 GSI002 - Introduction to Computer Programming

This module is offered in the first course period, 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 23 presents the layout produced from this module, which shows some homogeneous groups containing only students who succeeded or who failed, as well as some heterogeneous groups containing students in both situations. The layout confirms the high number of failure, in this case 61.7%. Figure 24a 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 on this module. One can also notice, by looking at the Module Attempt View (Figure 24b), that many students make several attempts at this module and that there is a significant number of student dropouts.

By analyzing a group of students who failed (Figure 25), it becomes clear that all of them obtained very low grades, highlighted in the Grades Distribution View (Figure 26a). 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 (2048), and the majority did not attend the topics covered in this module (Figure 26b). These students failed on their first attempt, and despite multiple attempts, they eventually dropped out (Figure 26c).

There are other students with similarly low grades — that is, grades insufficient for approval, with the majority of them having scored zero (Figure 27), which were placed in

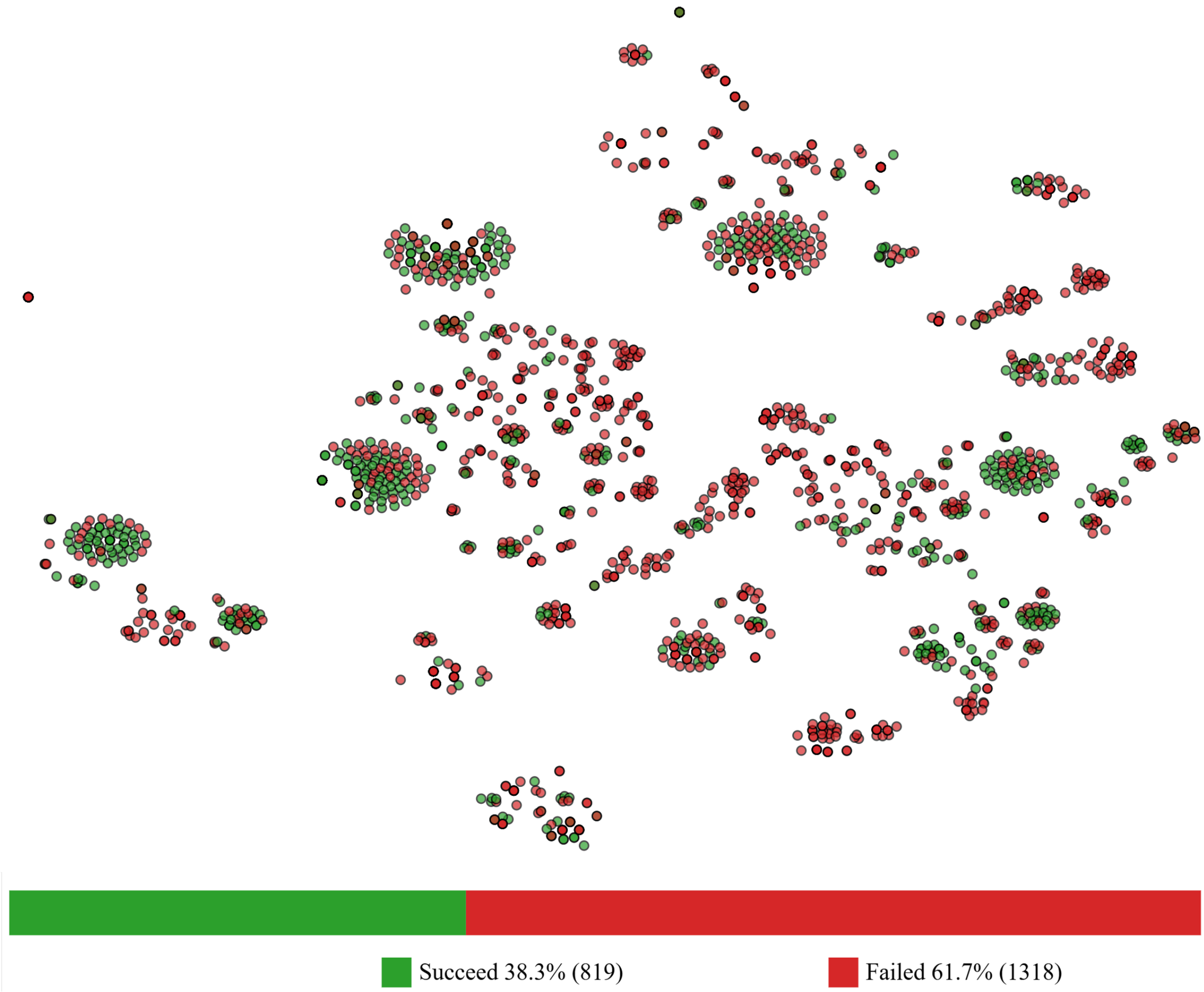
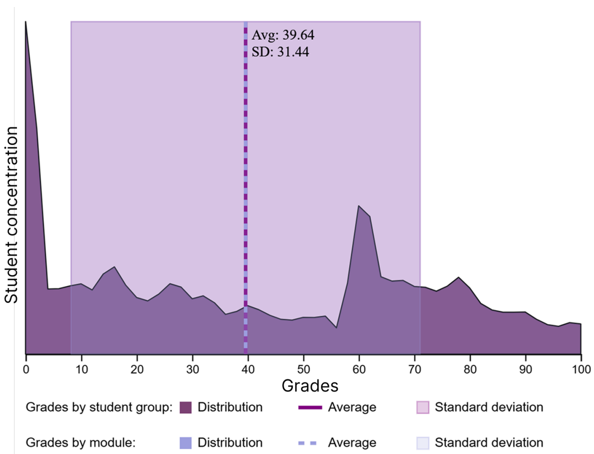
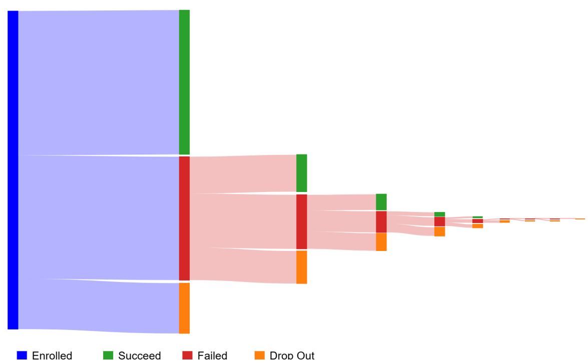


Figure 23 – Overview of student performance in the GSI002 module - Student Behavior View.



(a)



(b)

Figure 24 – Overview of student performance in the GSI002 module. **a)** Grades Distribution View, and **b)** Module Attempt View.

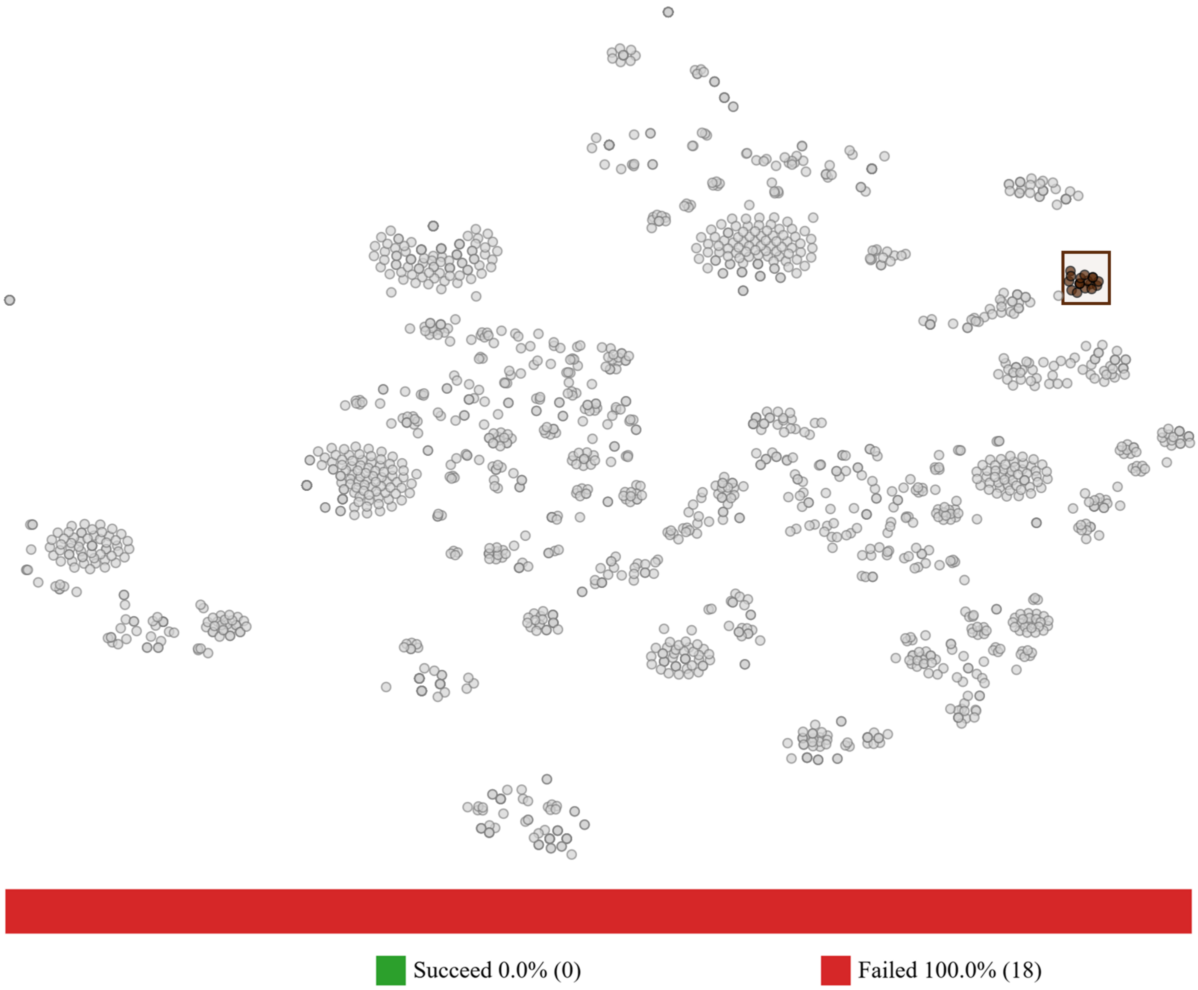


Figure 25 – Selection of a group composed only by students who failed - Student Frequency View

different groups. The difference between these students and the initially selected group is that, in these groups, the 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) (the topic name shown when hovering over its number). 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 28). 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 important 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

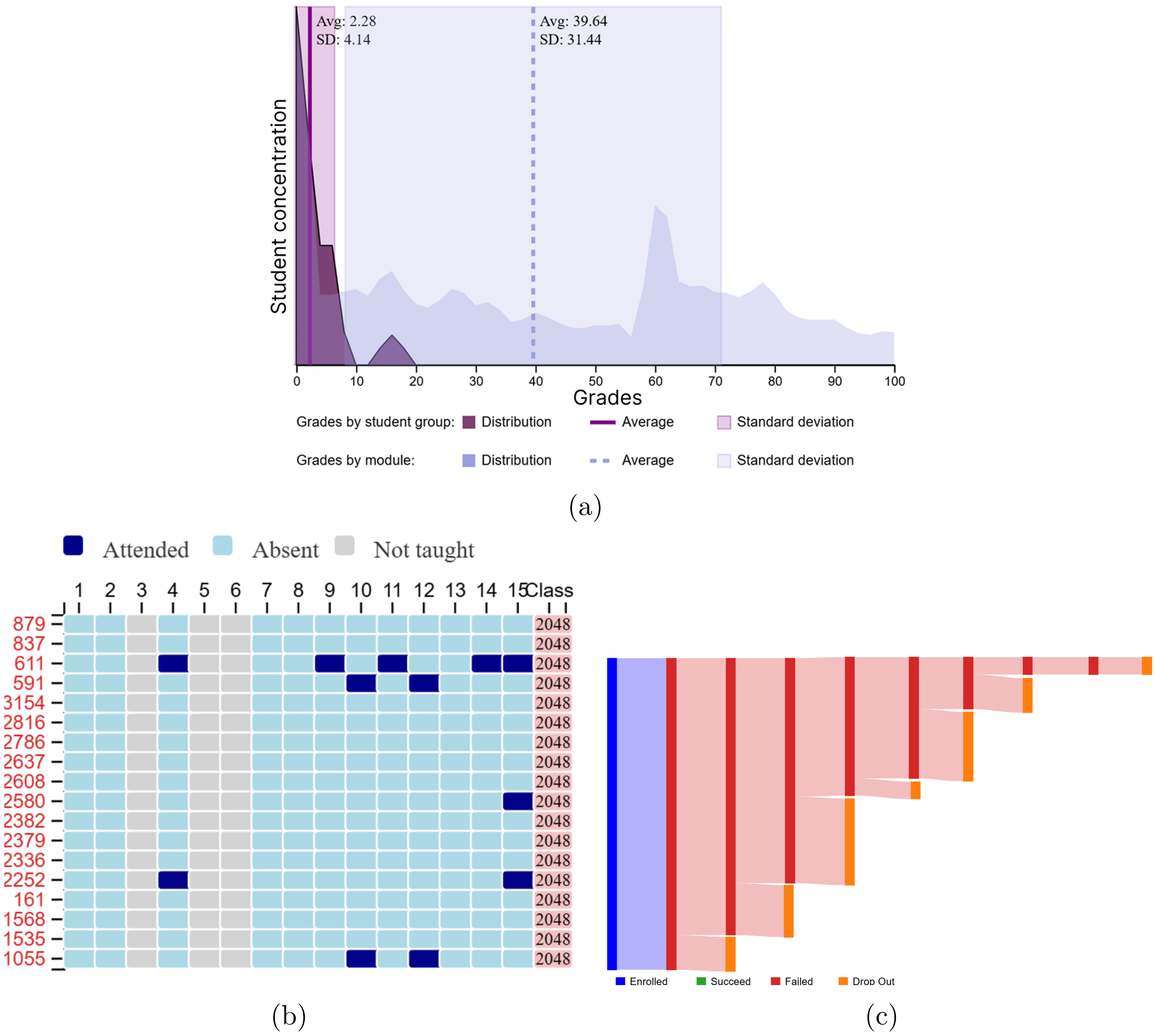


Figure 26 – Selection of a group composed only by students who failed. **a)** Grades Distribution View, **b)** Student Frequency View, and **c)** Module Attempt View.

these patterns may thus contribute significantly to decision making.

Analyzing another selected group who were present in all the topics taught (Figure 31), it can be noticed that 57.9% of students succeeded (Figure 29). The Module Attempt View shows that most of the students who succeeded did so on their first attempt (Figure 32). 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 30), indicating a higher average grade compared to the overall one.

Four students from this group succeeded in subsequent attempts, as shown in Figure 33a. One notices that all of them attended all topics in each attempt (Figures 33b to 33e), 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 29), the approval rate was below 60.0%. It is interesting to observe that the students

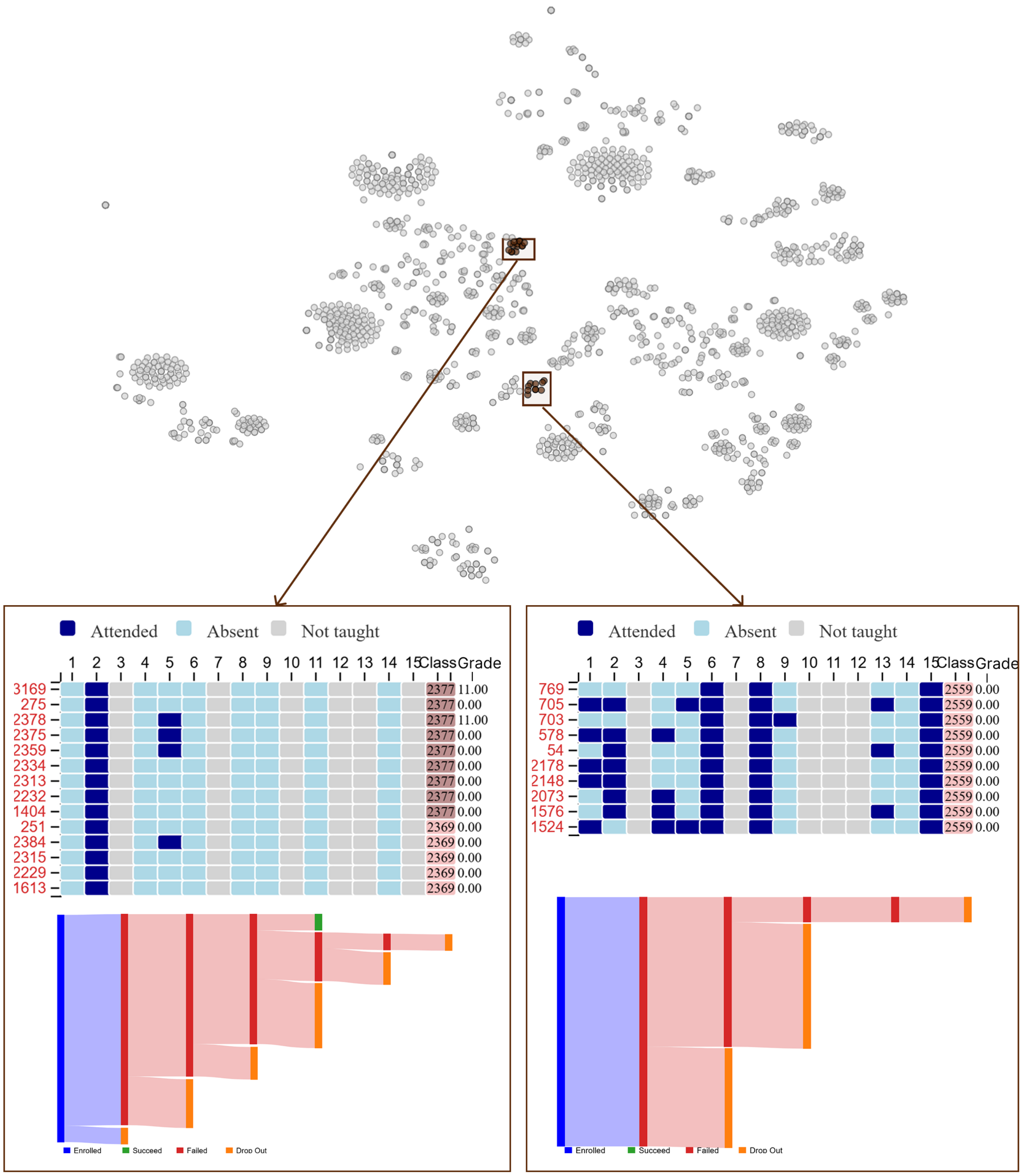


Figure 27 – Group of students with low grades.

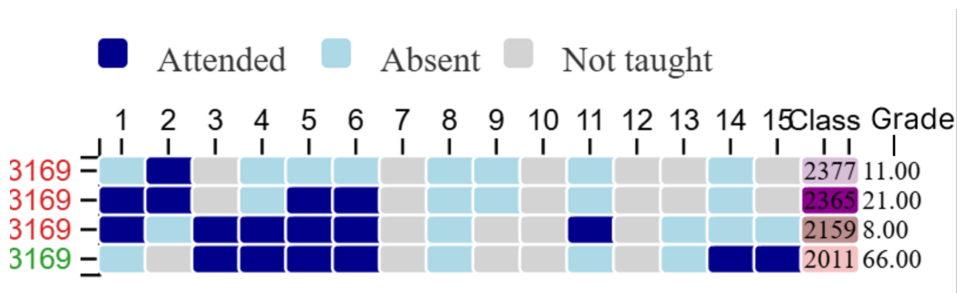


Figure 28 – Student 3169's behavior across module attempts.

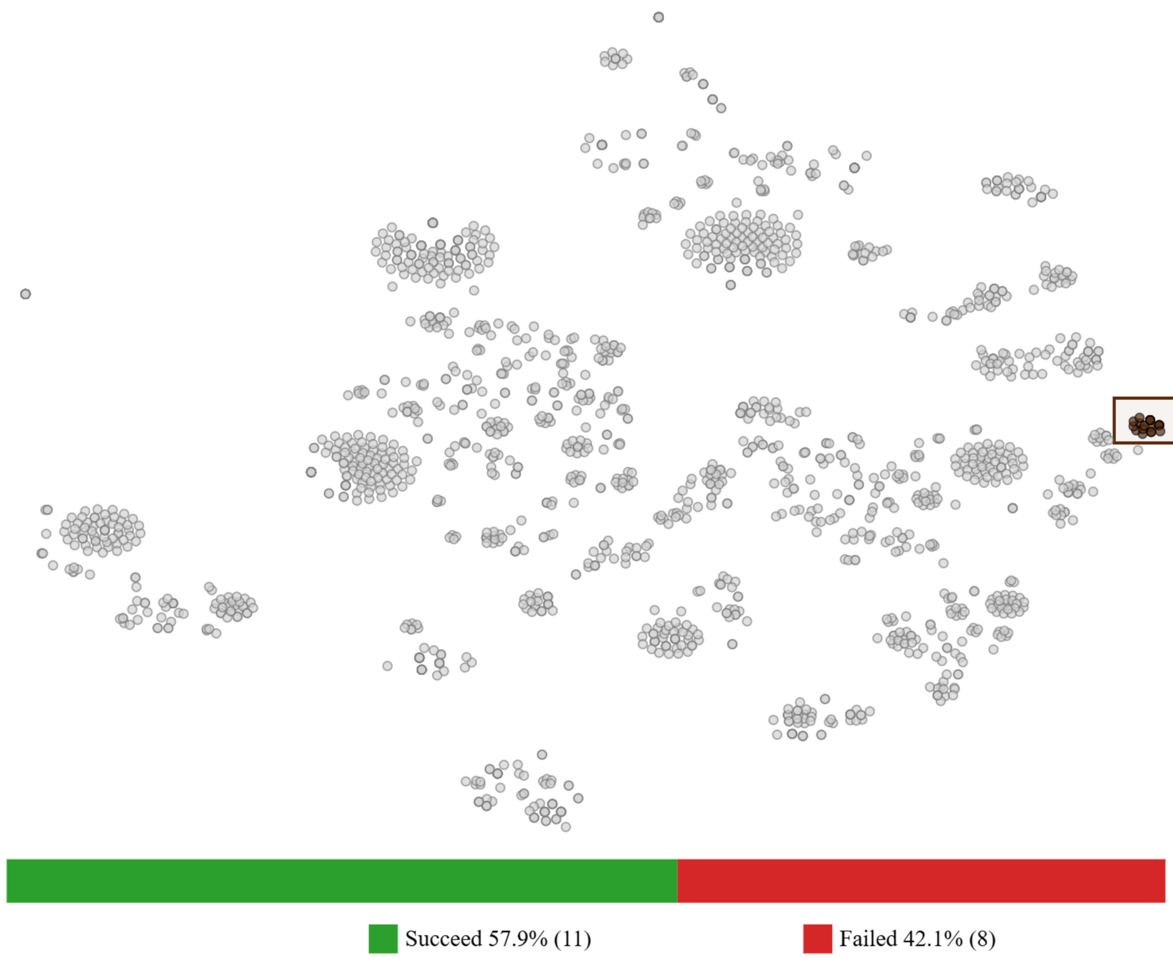


Figure 29 – Selection of a group containing students that succeeded and students that failed - Student Behavior View

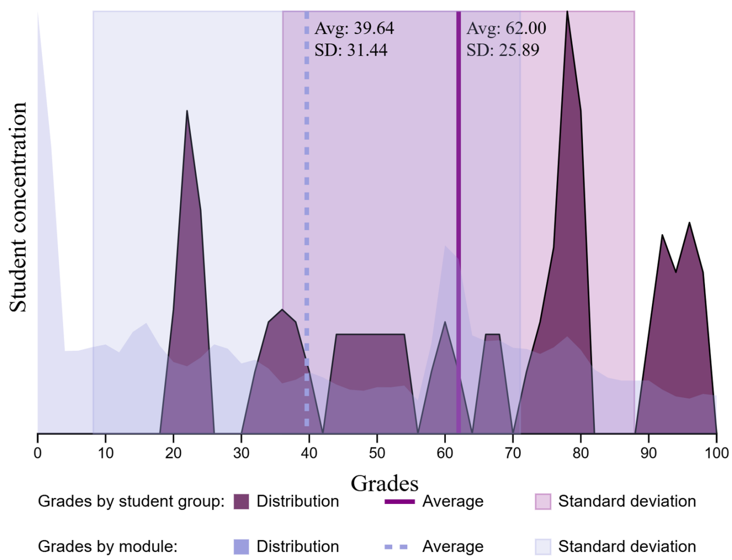


Figure 30 – Selection of a group containing students that succeeded and students that failed - Grades Distribution View

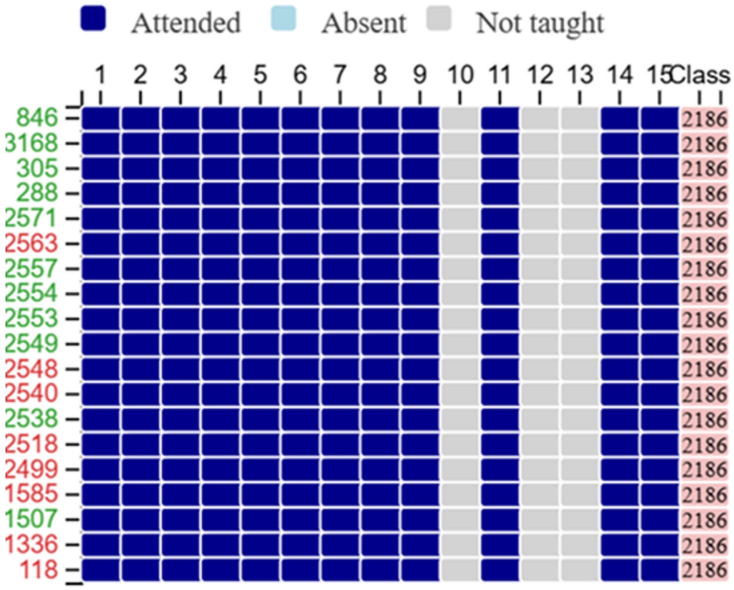


Figure 31 – Selection of a group containing students that succeeded and students that failed - Student Frequency View.

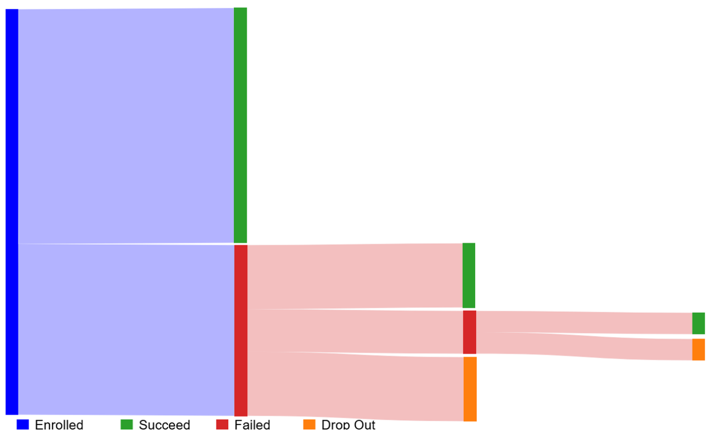


Figure 32 – Selection of a group containing students that succeeded and students that failed - Module Attempt View.

migrated to groups with even higher approval rates in subsequent attempts (Figure 33a). 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.

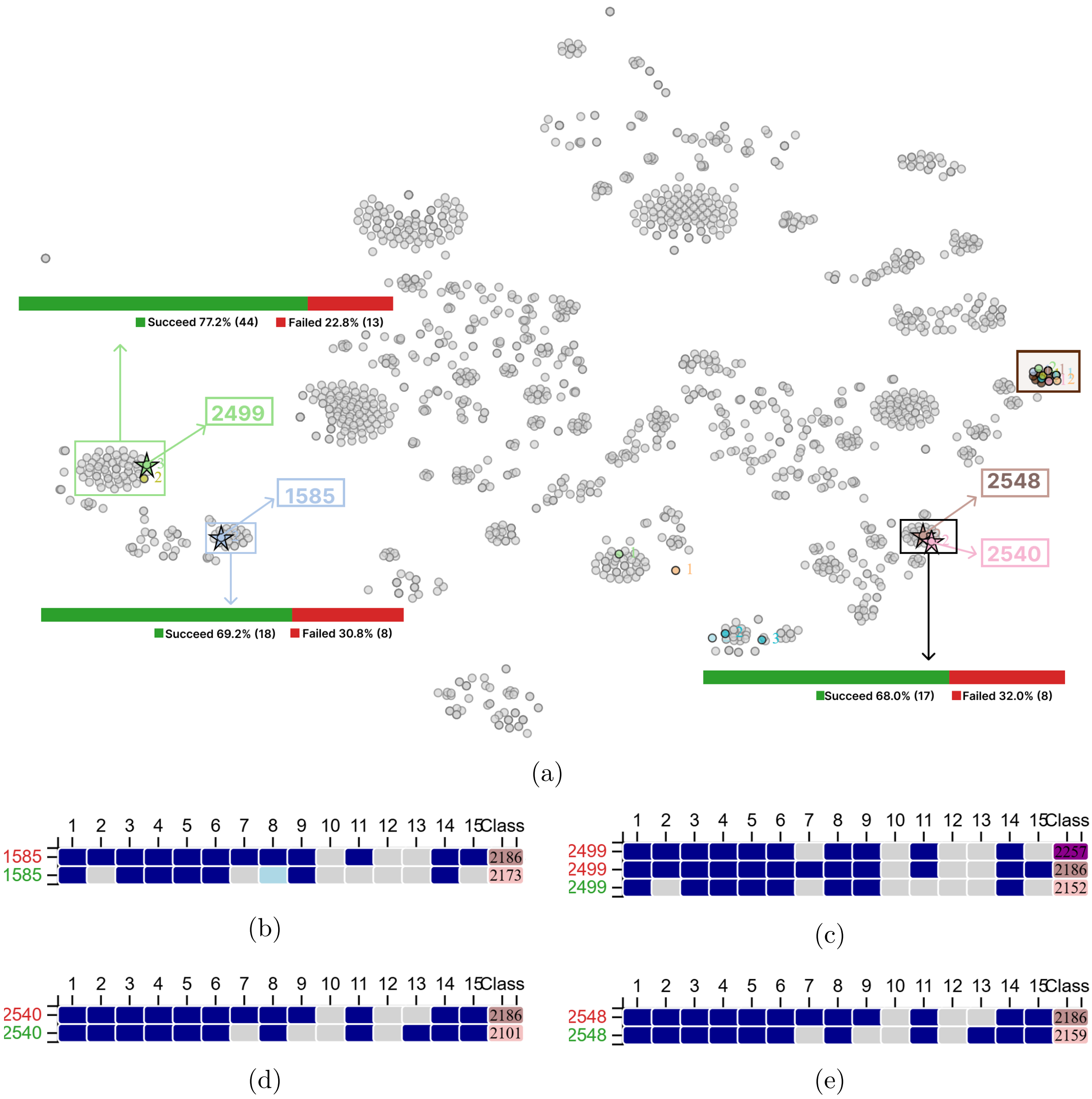


Figure 33 – Analysis of students who succeeded after multiple attempts. **a)** Student Behavior View selection, **b)** Student Frequency View for student 1585, **c)** Student Frequency View for student 2499, **d)** Student Frequency View for student 2540, and **e)** Student Frequency View for student 2548.

4.2.2 GSI009 - Information Systems Career

This module is offered in the second course period and has a history of more students succeeding than failing. Figure 34 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 35). Many students received a grade of 90.0. Most of the students succeeded on their first attempt (Figure 36), and some students, despite receiving approval grades, failed due to insufficient attendance, as marked in A in Figure 37. An interesting case appears in marker B (Figure 37),

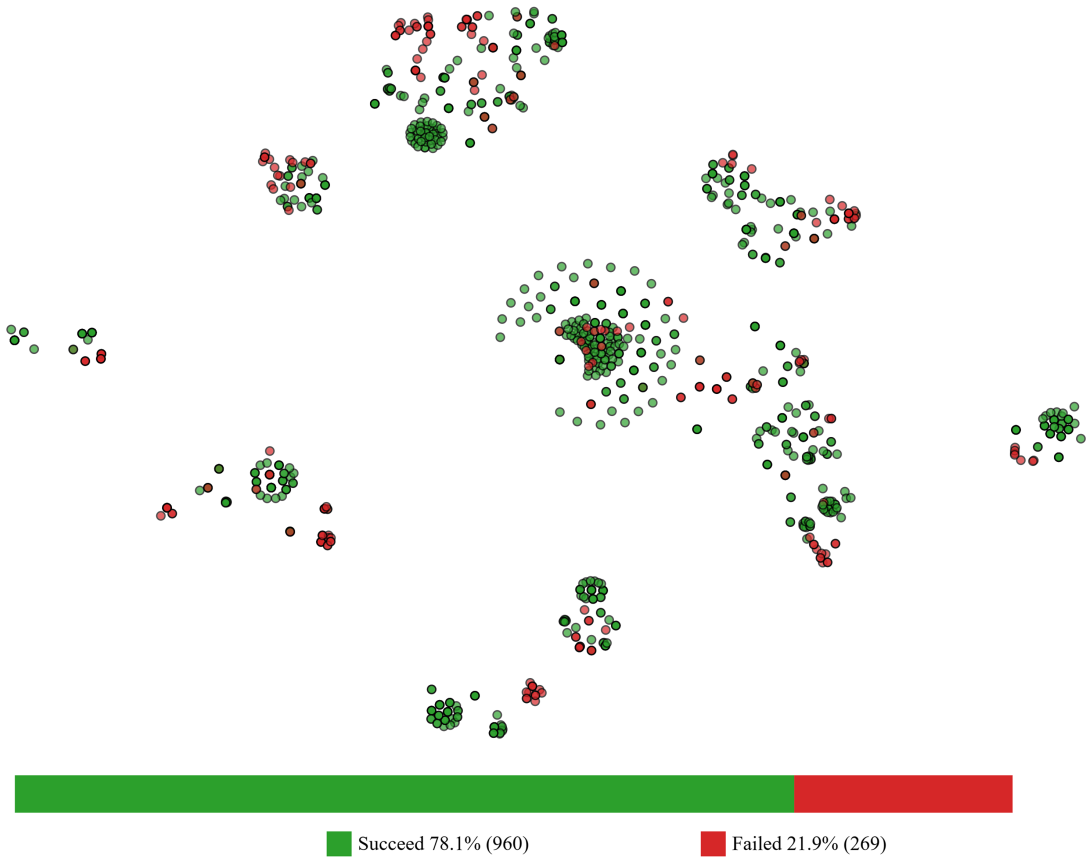


Figure 34 – Overview of student performance in the GSI009 module - Student Behavior View.

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 38) shows an average grade of approximately 5.7 (Figure 39). This group had a significant number of absences in almost all topics (Figure 40), and most of the students dropped out the module (Figure 41), except for one student (ID 2449). This student failed the first attempt receiving a grade of 6.0 and missed the topics *Ethics in Computer Science* (2), *Intellectual Property* (4), *Information Technology Professional Profile* (7), and *Information Technology Career* (10) (Figure 42). In the second attempt the student received a grade of 90.0, and was absent only from the topic *Ethics in Computer Science* (2) (Figure 42). The layout suggests that students who attended these topics may have been more likely to succeed,

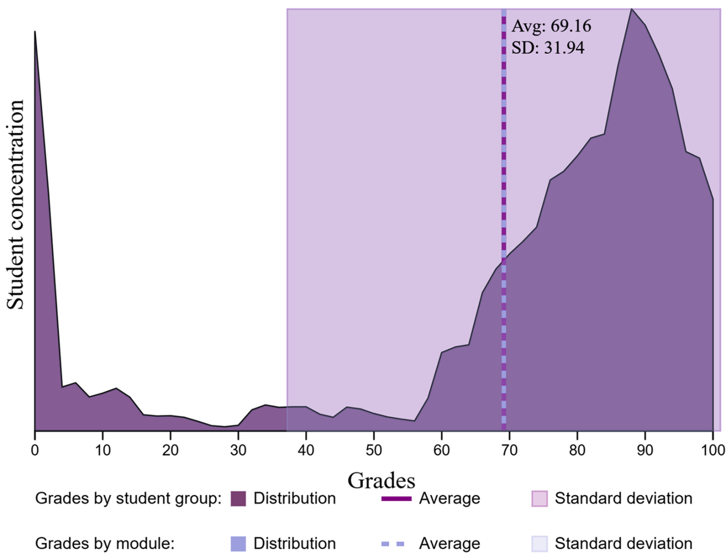


Figure 35 – Overview of student performance in the GSI009 module - Grades Distribution View.

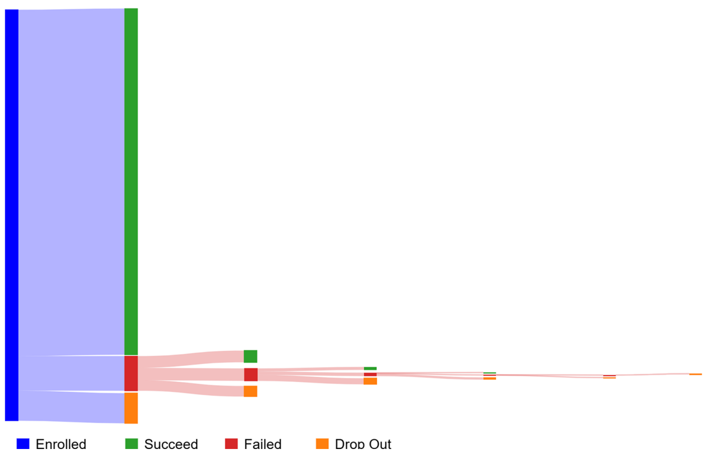


Figure 36 – Overview of student performance in the GSI009 module - Module Attempt View.

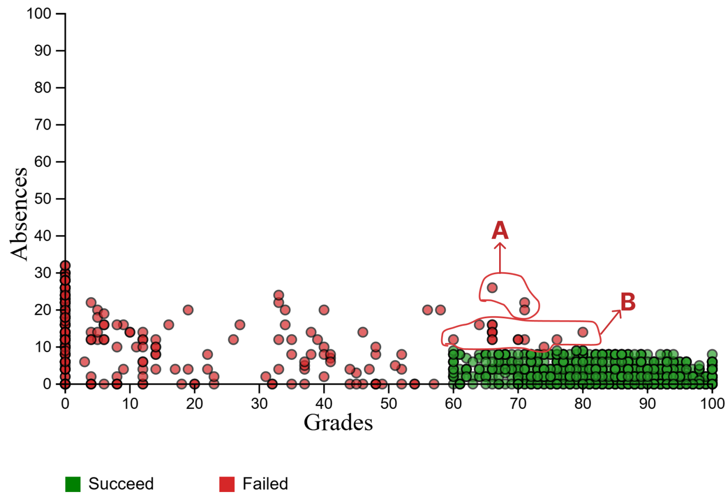


Figure 37 – Overview of student performance in the GSI009 module - Absences vs. Grades View.

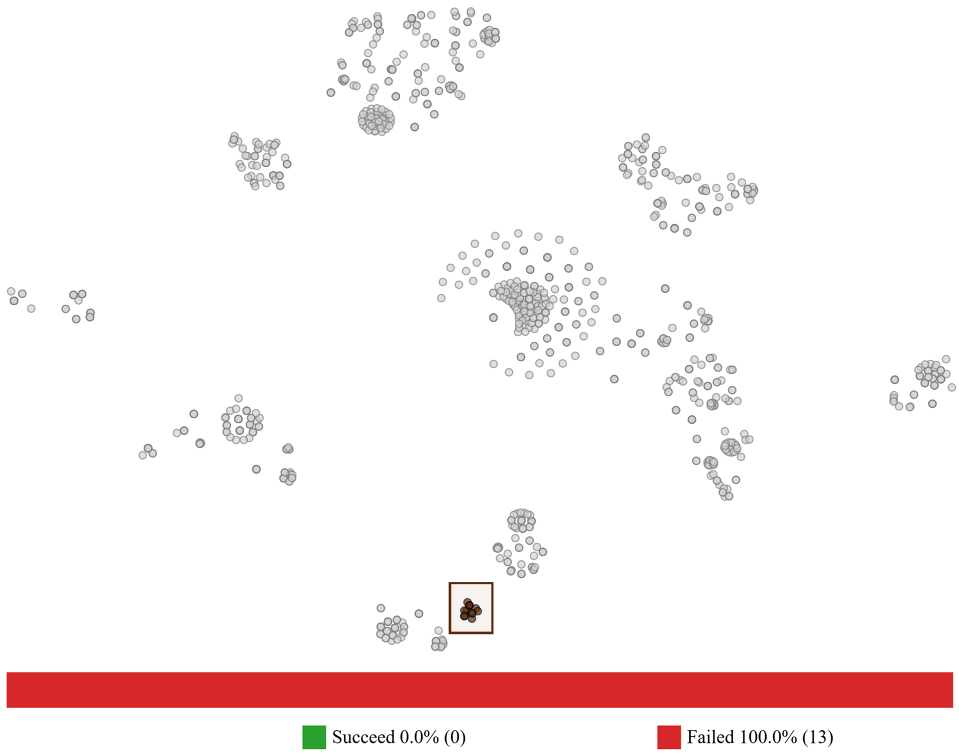


Figure 38 – Selection composed only by students who failed - Student Behavior View.

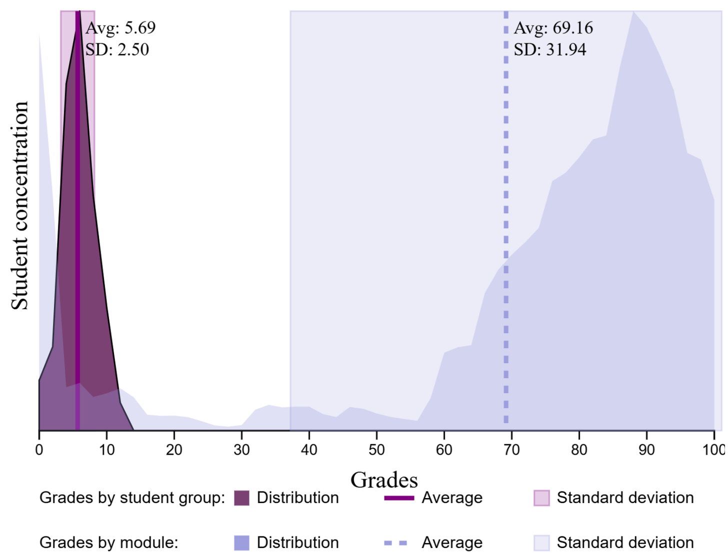


Figure 39 – Selection composed only by students who failed - Grades Distribution View.

potentially guiding further analysis by the course coordinator to identify which topics are most critical to academic success.

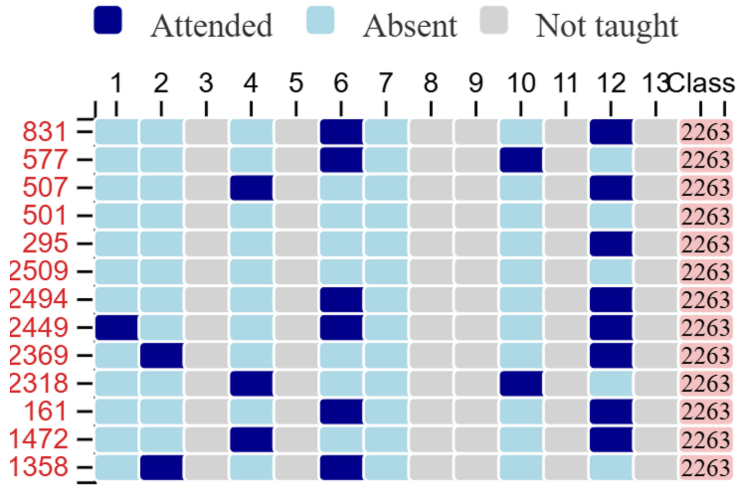


Figure 40 – Selection composed only by students who failed - Student Frequency View.

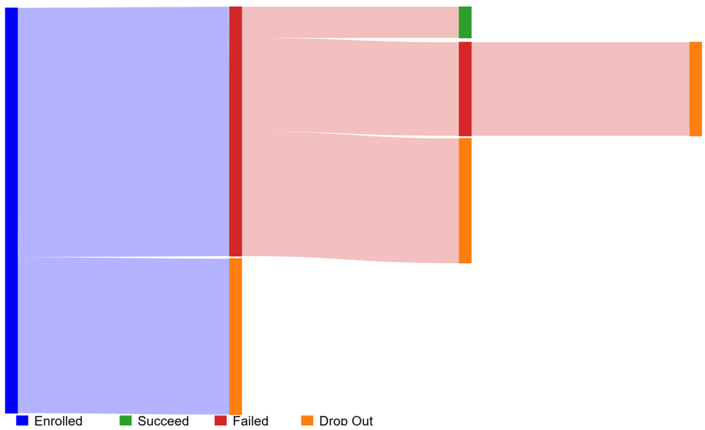


Figure 41 – Selection composed only by students who failed - Module Attempt View.

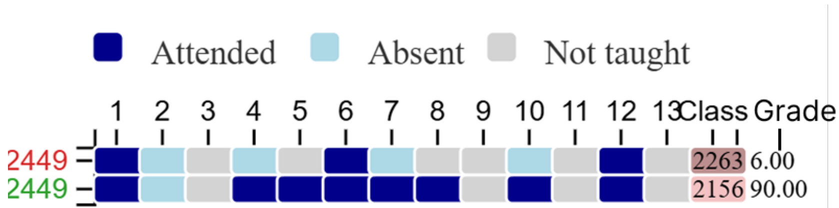


Figure 42 – Analysis of Student ID 2449 - Student Frequency View

Selecting a group of students who succeeded (Figure 43), the average grade for this group is 89.4, which is higher than the average grade for the module in general, which is approximately 69.2 (Figure 45). All of them attended all topics (Figure 44). Most succeeded on the first attempt (Figure 46), except for one student (ID 843), who succeeded on the second attempt (Figure 47). 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 48).

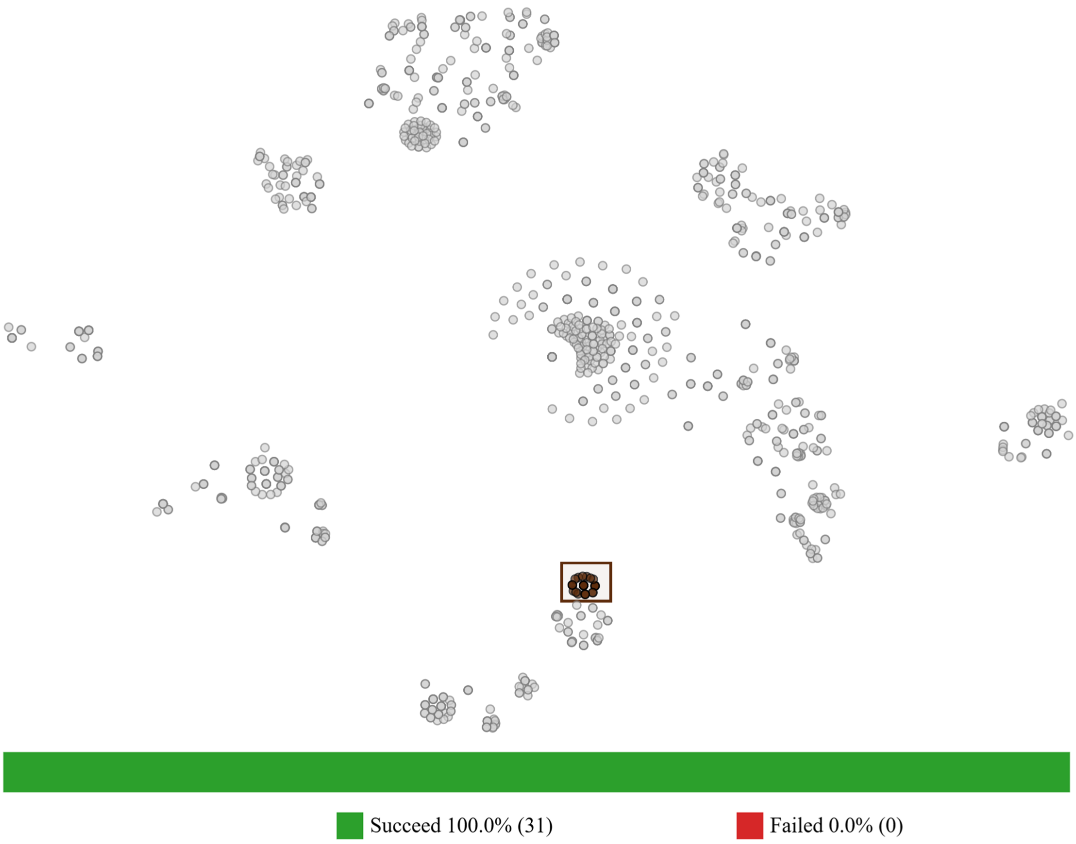


Figure 43 – Selection of a group containing only students that succeeded - Student Behavior View.

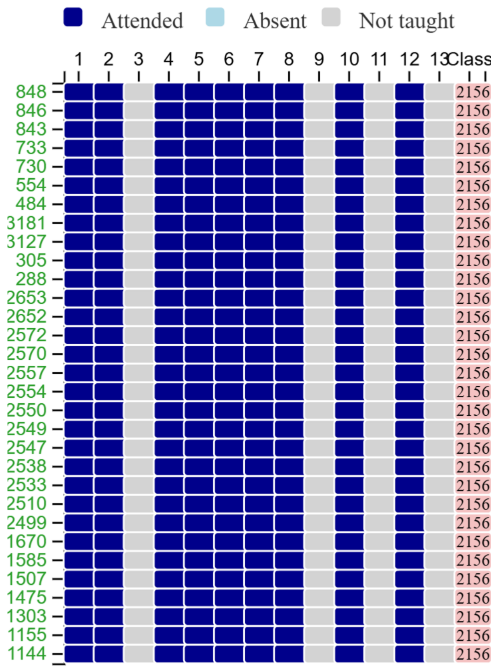


Figure 44 – Selection of a group containing only students that succeeded - Student Frequency View.

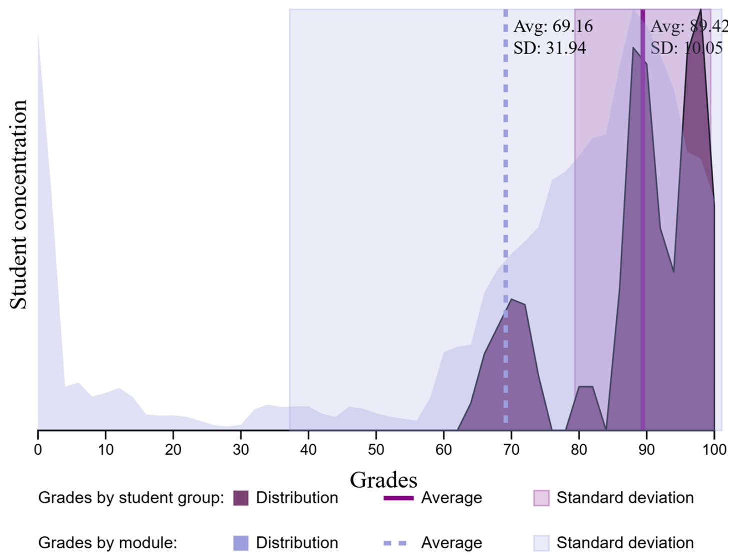


Figure 45 – Selection of a group containing only students that succeeded - Grades Distribution View.

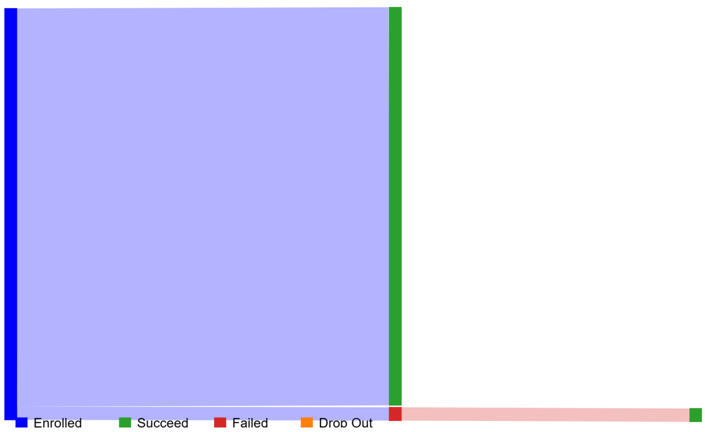


Figure 46 – Selection of a group containing only students that succeeded - Module Attempt View.

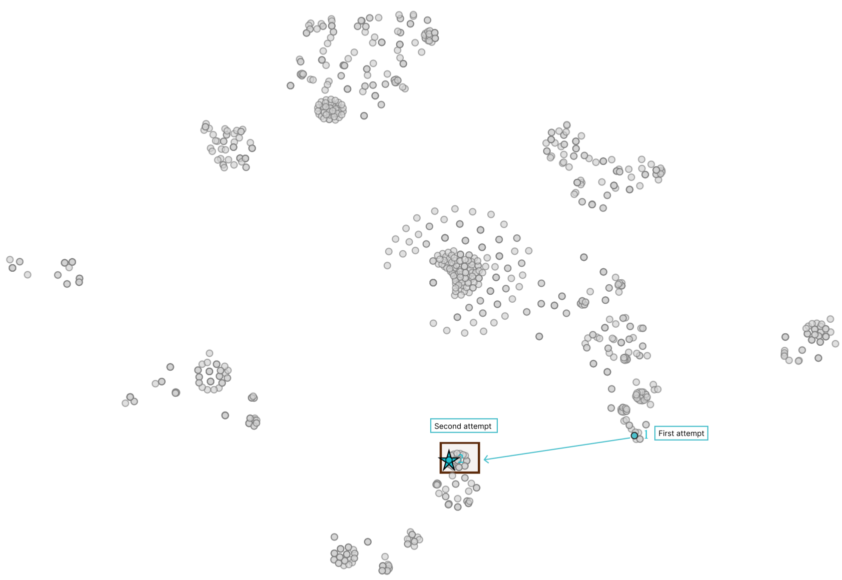


Figure 47 – Analysis of Student ID 843 - Student Behavior View.

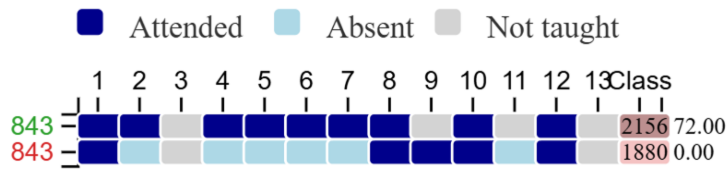


Figure 48 – Analysis of Student ID 843 - Student Frequency View.

The analysis of this module suggests that students do not face significant difficulties in succeeding. The layout is in accordance with the students' feelings about this module, in terms of approval challenge. However, the layout also highlighted the importance of the attendance for ensuring approval.

4.2.3 GSI019 - Web Programming Module

This module is offered in the fourth course period and has a history of more students succeeding than failing. The layout in Figure 49 shows that 58.9% of the students succeeded, and that a significant portion of the students obtained a grade between 60.0 and 90.0 (Figure 50), but one also notices a considerable number of students which obtained a grade of 0.0. Figure 51 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 51.4% of the students succeeded on their first attempt, while 17.2% dropped out and 14.1% failed (Figure 52). Some of the students who failed required up to four attempts before succeeding.

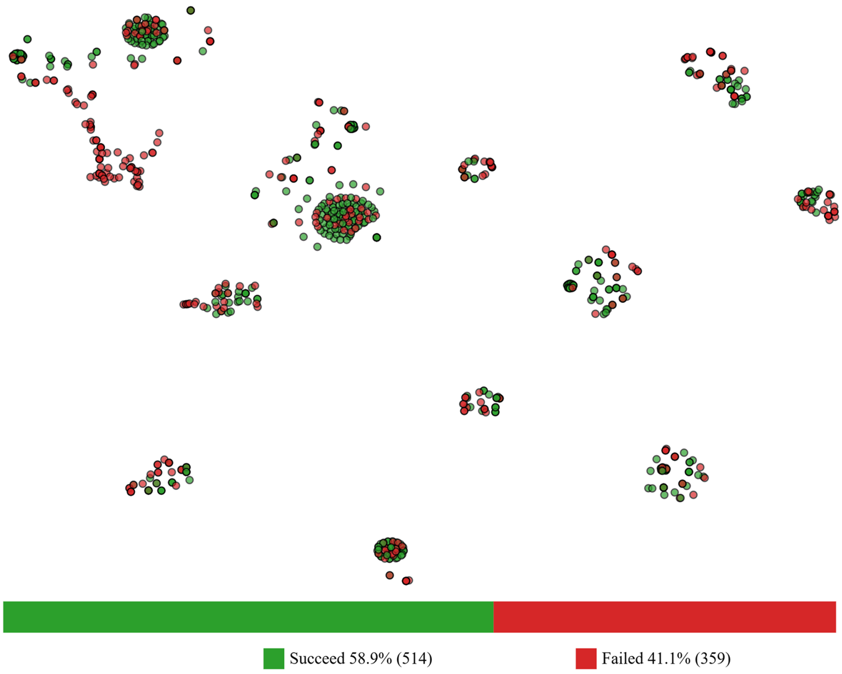


Figure 49 – Overview of student performance in the GSI019 module - Student Behavior View.

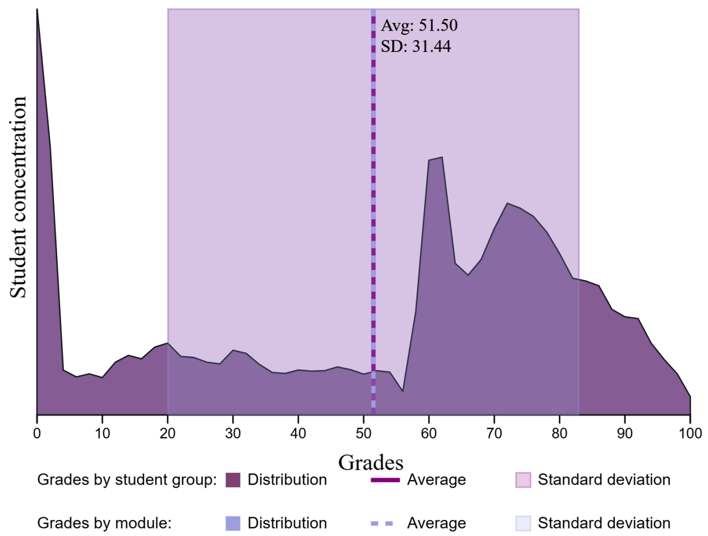


Figure 50 – Overview of student performance in the GSI019 module - Grades Distribution View.

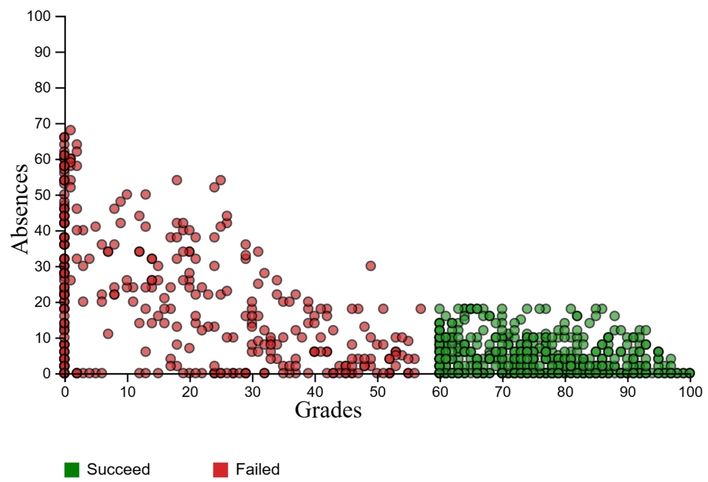


Figure 51 – Overview of student performance in the GSI019 module - Absences vs. Grades View.

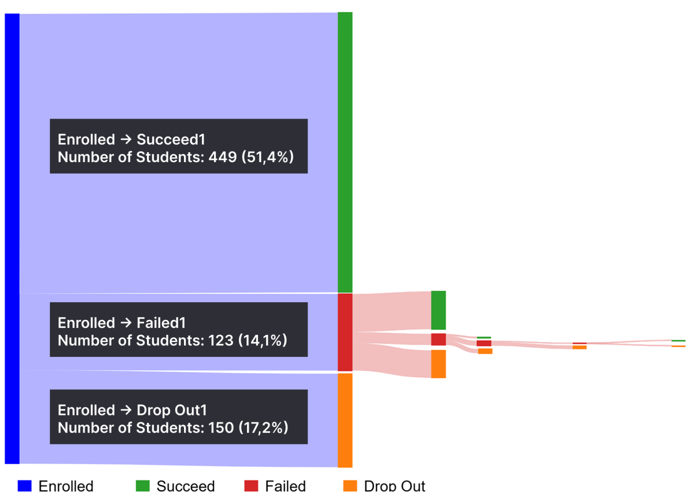


Figure 52 – Overview of student performance in the GSI019 module - Module Attempt View.

When selecting the group highlighted in Figure 53, it is possible to notice that the overall performance in the module was positive and the approval rate was 70.5%. Figure 54 shows a higher concentration of grades between 60 and 90, confirming their approval. All students from this group attended all topics (Figure 55), and most of them succeeded in their first attempt (Figure 56).

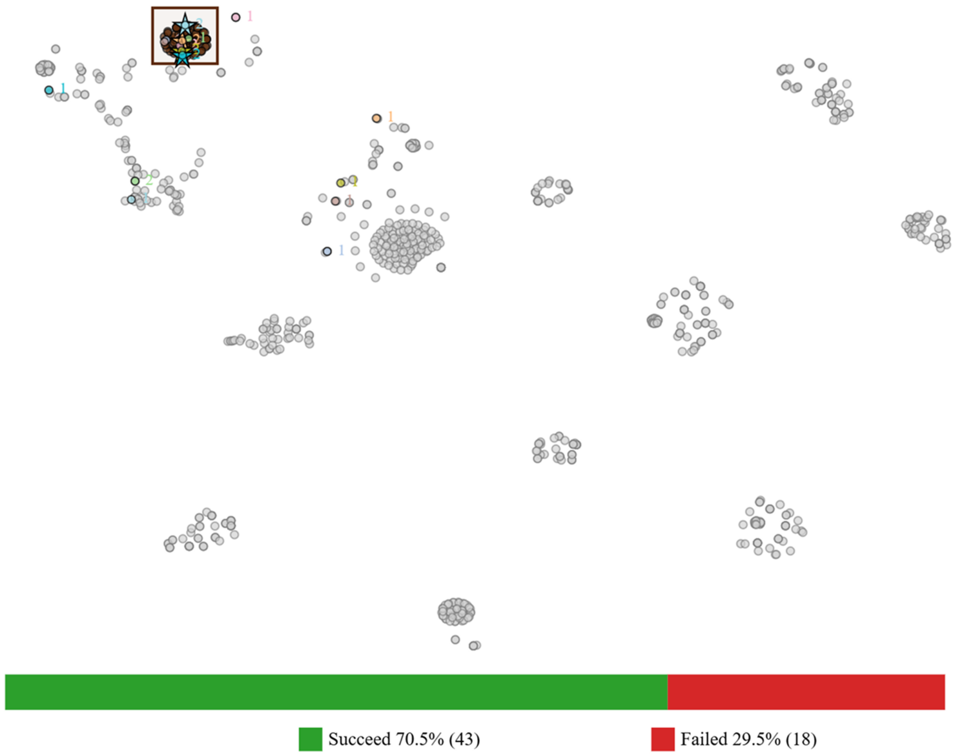


Figure 53 – Selection of a group of students in the GSI019 module - Student Behavior View.

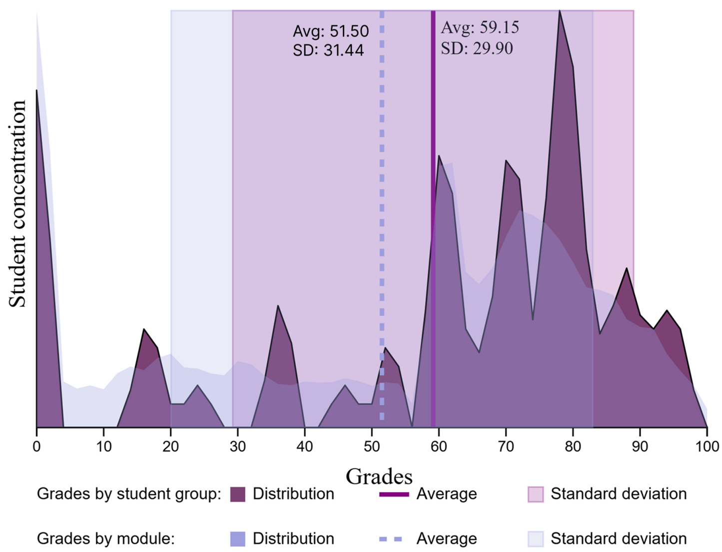


Figure 54 – Selection of a group of students in the GSI019 module - Grades Distribution View.

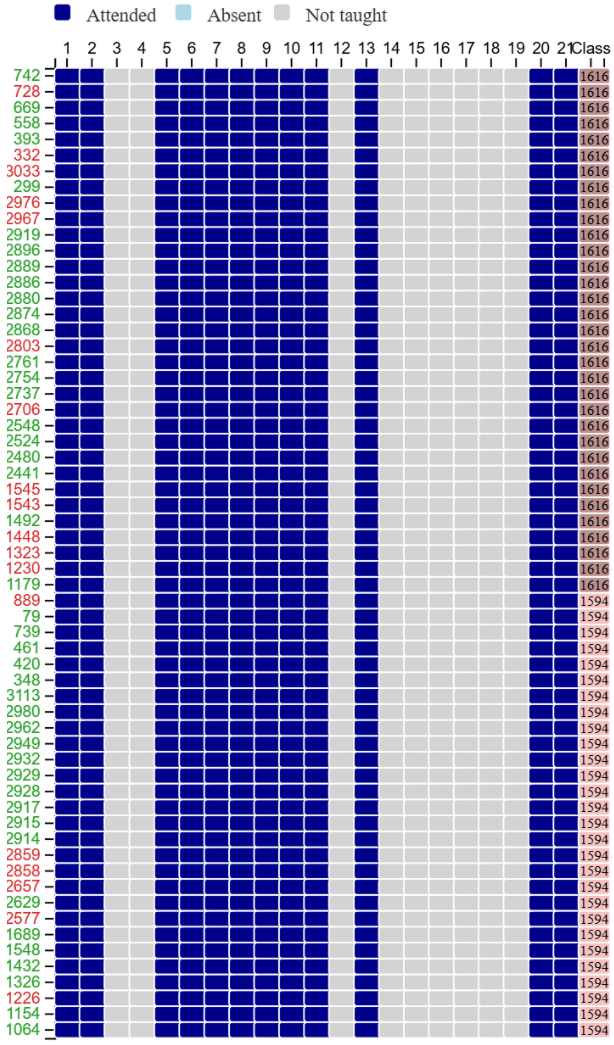


Figure 55 – Selection of a group of students in the GSI019 module - Student Frequency View.

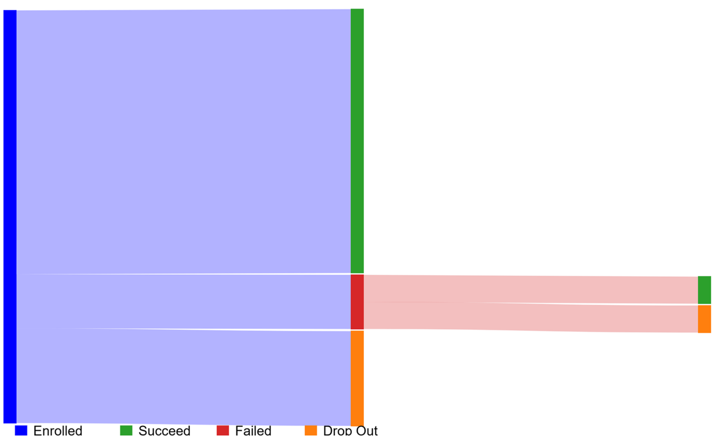


Figure 56 – Selection of a group of students in the GSI019 module - Module Attempt View.

However, there is a group of students who either dropped out or failed, with no more than two attempts at the module (Figure 56). Four of these students succeeded in their second attempt. As shown in Figure 57, they all missed some topics during their first attempt. In contrast, during their second attempt, they regularly attended all topics, which may have contributed to their eventual success. This change in attendance behavior highlights the importance of consistent participation in achieving academic approval.

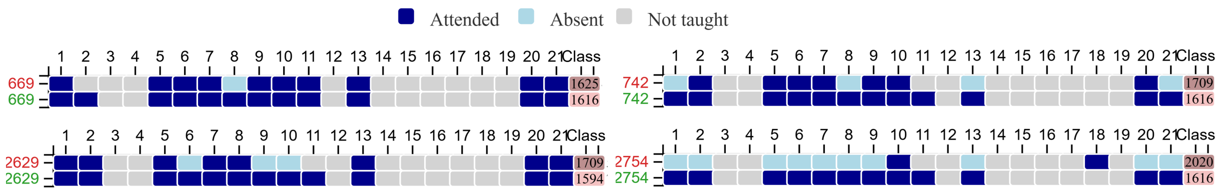


Figure 57 – Selection of a group of students in the GSI019 module.

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 later succeeded 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.

Chapter 5

Conclusion

This study introduced ***LearnVis***, a visualization system designed to support the analysis of student performance across semesters in a course program. The system aims to facilitate the identification of relevant academic patterns by providing users with an interactive visual interface for exploring educational data. We demonstrated its application using real data from students enrolled in a Information Systems course at a Brazilian university, with the goal of understanding student behavior and highlighting critical aspects such as performance by module, the influence of attendance, patterns of repeated attempts, as well as the correlation between absences in specific topics and the final outcomes in each module.

Throughout this work, we showed that the layouts developed provided effective and interactive analyzes, offering managers and researchers a new way to visualize and understand academic performance. With the *Student Behavior View* layout, it was possible to identify student profiles with similar behavioral patterns, which facilitated group analysis and provided a greater understanding of the difficulties encountered. In combination with the other layouts, we investigated student attendance and assessed the direct impact of presence in classes on final performance. This approach suggests, for example, that being absent from specific topics may be directly related to failures and that the timing of absence is as relevant as the overall number of missed classes.

Additionally, the system provides complementary visualizations that enrich the analysis. One highlighted feature is the ability to select groups of students and identify which ones are repeating a particular module, differentiated by colors and accompanied by a glyph indicating approval in later attempts. This visualization makes it easier to track students over time, allowing users to understand if and how performance changes after a failure. Such insights would be difficult to obtain through tabulated data or traditional charts, which do not allow integrated and dynamic analysis across multiple variables.

Among the most relevant results obtained with ***LearnVis***, we highlight the identification of recurring behavioral patterns, such as groups of students who repeatedly fail certain modules, or who show dropout behavior after consecutive failures. We also demonstrated that certain topics taught in class have a greater impact on the final average than others, suggesting that targeted interventions on specific content could significantly improve overall student performance. These findings are especially valuable to instructors, as they provide concrete evidence on where to focus pedagogical efforts and student support strategies.

The system developed can thus serve as a strategic tool for institutional policy-making. By enabling quick and visual identification of critical situations—such as modules with high failure rates, low engagement, or irregular attendance—***LearnVis*** supports the creation of precise and evidence-based actions to improve academic outcomes, increase student retention.

The proposed visual strategy and the system itself were applied in this research to data provided by the Computing and Informatics courses of the Federal University of Uberlândia. However, we believe that this approach is generalizable and can be easily adapted for use with datasets from other institutions or programs, as long as they share similar structures. The flexibility of the tool allows it to be applied in different educational contexts, including institutions seeking to evaluate the impact of curriculum reforms, pedagogical changes, or institutional policies for student monitoring.

5.1 Limitations

We developed *LearnVis* based on iterative discussions with educational experts, guided by the requirements and analytical tasks described in Section 3. 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. 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.

5.2 Future Work

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.

Bibliography

ALI, S.; HAFEEZ, Y.; HUMAYUN, M.; JAMAIL, N. S. M.; AQIB, M.; NAWAZ, A. Enabling recommendation system architecture in virtualized environment for e-learning. **Egyptian Informatics Journal**, Elsevier, v. 23, n. 1, p. 33–45, 2022. DOI: 10.1016/j.eij.2021.05.003.

ARCINAS, M. M.; MEENAKSHI, M.; BAHALKAR, P. S.; BHATURKAR, D.; LALAR, S.; RANE, K. P.; GARG, S.; OMAROV, B.; RAGHUVANSHI, A. An efficient course recommendation system for higher education students using machine learning techniques. **Bulletin of Electrical Engineering and Informatics**, v. 14, n. 2, p. 1468–1475, 2025. DOI: 10.11591/eei.v14i2.771.

BADAL, Y. T.; SUNGKUR, R. K. Predictive modelling and analytics of students' grades using machine learning algorithms. **Education and Information Technologies**, Springer, v. 28, n. 3, p. 3027–3057, 2023. DOI: 10.1007/s10639-022-11299-8.

BELLAJ, M.; DAHMANE, A. B.; BOUDRA, S.; SEFIAN, M. L. Educational data mining: Employing machine learning techniques and hyperparameter optimization to improve students' academic performance. **International Journal of Online & Biomedical Engineering**, v. 20, n. 3, 2024. DOI: 10.3991/ijoe.v20i03.46287.

BORG, I.; GROENEN, P. J. **Modern multidimensional scaling: Theory and applications**. [S.l.]: Springer Science & Business Media, 2005. DOI: 10.1007/0-387-28981-X.

BOSTOCK, M.; OGIEVETSKY, V.; HEER, J. **D^3 data-driven documents**. [S.l.]: IEEE, 2011. 2301–2309 p. DOI: 10.1109/TVCG.2011.185.

BRANDT, J. Z.; TEJEDO-ROMERO, F.; ARAUJO, J. F. F. E. Fatores influenciadores do desempenho acadêmico na graduação em administração pública. **Educação e Pesquisa**, SciELO Brasil, v. 46, p. e202500, 2020. DOI: 10.1590/s1678-4634202046202500.

CARD, S. K.; MACKINLAY, J.; SHNEIDERMAN, B. **Readings in information visualization: Using vision to think**. San Francisco, CA: Morgan Kaufmann, 1999. 686 p. ISBN 978-1-55860-533-6.

CHEN, Z.; WANG, J.; XIA, M.; SHIGYO, K.; LIU, D.; ZHANG, R.; QU, H. Stugptviz: A visual analytics approach to understand student-chatgpt interactions.

IEEE Transactions on Visualization and Computer Graphics, v. 31, n. 1, p. 908–918, 2025. DOI: 10.1109/TVCG.2024.3456363.

CLEVELAND, W. S. **Visualizing Data**. Summit, NJ: Hobart Press, 1993. 360 p. ISBN 0-9634884-0-6.

COX, T. F.; COX, M. A. A. Multidimensional scaling. In: CHEN, C. houh; HÄRDLE, W.; UNWIN, A. (Ed.). **Handbook of Data Visualization**. Berlin, Heidelberg: Springer, 2008. p. 315–347. DOI: 10.1007/978-3-540-33037-0$_{1}$4.

DENG, H.; WANG, X.; GUO, Z.; DECKER, A.; DUAN, X.; WANG, C.; AMBROSE, G. A.; ABBOTT, K. Performancevis: Visual analytics of student performance data from an introductory chemistry course. **Visual Informatics**, Elsevier, v. 3, n. 4, p. 166–176, 2019. DOI: 10.1016/j.visinf.2019.10.004.

ETEMADPOUR, R.; MOTTA, R.; PAIVA, J. G. de S.; MINGHIM, R.; OLIVEIRA, M. C. F. D.; LINSEN, L. Perception-based evaluation of projection methods for multidimensional data visualization. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 21, n. 1, p. 81–94, 2014. DOI: 10.1109/TVCG.2014.2330617.

ETEMADPOUR, R.; ZHU, Y.; ZHAO, Q.; HU, Y.; CHEN, B.; SHARIER, M. A.; ZHENG, S.; PAIVA, J. G. S. Role of absence in academic success: an analysis using visualization tools. **Smart Learning Environments**, SpringerOpen, v. 7, n. 1, p. 1–25, 2020. DOI: 10.1186/s40561-019-0112-3.

FIOR, C. A.; POLYDORO, S. A. J.; PELISSONI, A. M. S.; DANTAS, M. A.; MARTINS, M. J.; ALMEIDA, L. d. S. Impacto da autoeficácia e do rendimento acadêmico no abandono de estudantes do ensino superior. **Psicologia Escolar e Educacional**, SciELO Brasil, v. 26, p. e235218, 2022. DOI: 10.1590/2175-35392022235218t.

GARCIA-ZANABRIA, G.; GUTIERREZ-PACHAS, D. A.; CAMARA-CHAVEZ, G.; POCO, J.; GOMEZ-NIETO, E. SDA-Vis: A visualization system for student dropout analysis based on counterfactual exploration. **Applied Sciences**, v. 12, n. 12, 2022. DOI: 10.3390/app12125785.

GHAZVINI, A.; SHAREF, N. M.; SIDI, F. B. Prediction of course grades in computer science higher education program via a combination of loss functions in lstm model. **IEEE Access**, v. 12, p. 30220–30241, 2024. DOI: 10.1109/ACCESS.2024.3351186.

GOMES, L. S. I.; SCHLICKMANN, R. Política de apoio pedagógico no ensino superior: um estudo de caso. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, SciELO Brasil, v. 29, p. e024003, 2024. DOI: 10.1590/1982-57652024v29id2756283.

GUTIERREZ-PACHAS, D. A.; GARCIA-ZANABRIA, G.; CUADROS-VARGAS, E.; CAMARA-CHAVEZ, G.; GOMEZ-NIETO, E. 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**, v. 13, n. 2, p. 1–19, 2023. DOI: 10.3390/educsci13020154.

HUSSAIN, S.; KHAN, M. Q. Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. **Annals of data science**, Springer, v. 10, n. 3, p. 637–655, 2023. DOI: 10.1007/s40745-021-00341-0.

JOHNSON, B.; SHNEIDERMAN, B. **Tree-maps: A space filling approach to the visualization of hierarchical information structures**. [S.l.], 1998.

JOIA, P.; COIMBRA, D.; CUMINATO, J. A.; PAULOVICH, F. V.; NONATO, L. G. Local affine multidimensional projection. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 17, n. 12, p. 2563–2571, 2011. DOI: 10.1109/TVCG.2011.220.

JOLLIFFE, I. T. Principal component analysis for special types of data. In: ______. **Principal component analysis**. 2. ed. [S.l.]: Springer, 2002. cap. 8, p. 338–372. ISBN 978-0-387-95442-4. DOI: 10.1007/0-387-22440-8$_1$3.

KEIM, D. A.; HAO, M. C.; DAYAL, U.; HSU, M. Pixel bar charts: a visualization technique for very large multi-attribute data sets. **Information Visualization**, SAGE Publications Sage UK: London, England, v. 1, n. 1, p. 20–34, 2002. DOI: 10.1057/palgrave.ivs.9500003.

KOCSIS, Á.; MOLNÁR, G. Factors influencing academic performance and dropout rates in higher education. **Oxford Review of Education**, Taylor & Francis, p. 1–19, 2024. DOI: 10.1080/03054985.2024.2316616.

LAKSHMI, S.; MAHESWARAN, C. P. Effective deep learning based grade prediction system using gated recurrent unit (gru) with feature optimization using analysis of variance (anova). **Automatika**, Taylor & Francis, v. 65, n. 2, p. 425–440, 2024. DOI: 10.1080/00051144.2023.2296790.

LEOPOLD, T.; BATTISTA, A. D.; JATIVA, X.; SHARMA, S.; LI, R.; GRAYLING, S. Future of jobs report 2025. In: **World Economic Forum. https://www.weforum.org/publications/the-future-of-jobs-report-2025/**. [S.l.: s.n.], 2025.

LIN, L.; ZHOU, D.; WANG, J.; WANG, Y. A systematic review of big data driven education evaluation. **Sage Open**, SAGE Publications Sage CA: Los Angeles, CA, v. 14, n. 2, p. 21582440241242180, 2024. DOI: 10.1177/21582440241242180.

LIU, J.; WAN, G.; LIU, W.; LI, C.; PENG, S.; XIE, Z. High-dimensional spatiotemporal visual analysis of the air quality in china. **Scientific Reports**, Nature Publishing Group UK London, v. 13, n. 1, p. 5462, 2023. DOI: 10.1038/s41598-023-31645-1.

LIU, S.; CUI, W.; WU, Y.; LIU, M. A survey on information visualization: recent advances and challenges. **The Visual Computer**, Springer, v. 30, p. 1373–1393, 2014. DOI: 10.1007/s00371-013-0892-3.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. 11, p. 2579–2605, 2008. Disponível em: <http://www.jmlr.org/papers/v9/vandermaaten08a.html>.

MANDINACH, E. B.; ABRAMS, L. M. Data literacy and learning analytics. In: LANG, C.; SIEMENS, G.; WISE, A. F.; GASEVIC, D.; MERCERON, A. (Ed.). **The handbook of learning analytics**. Canada: SoLAR, 2022. cap. 19, p. 196–204. DOI: 10.18608/hla22.019.

MANPOWERGROUP. **Escassez de talentos no Brasil e no mundo: quem detém o talento, detém o futuro**. 2022. Accessed on 14 May 2023. Disponível em: <https://blog.manpowergroup.com.br/escassez-de-talentos-no-brasil-e-no-mundo-quem-detem-o-talento-detem-o-futuro>.

MARTINS, M. V.; BAPTISTA, L.; MACHADO, J.; REALINHO, V. Multi-class phased prediction of academic performance and dropout in higher education. **Applied Sciences**, MDPI, v. 13, n. 8, p. 1–15, 2023. DOI: 10.3390/app13084702.

MARTINS, R. M.; BERGE, E.; MILRAD, M.; MASIELLO, I. Visual learning analytics of multidimensional student behavior in self-regulated learning. In: SCHEFFEL, M.; BROISIN, J.; PAMMER-SCHINDLER, V.; IOANNOU, A.; SCHNEIDER, J. (Ed.). **Transforming Learning with Meaningful Technologies**. [S.l.], 2019. p. 737–741. DOI: 10.1007/978-3-030-29736-7_78.

MAYR, G. **Die Gesetzmässkeit im Gesellschaftsleben statistische Studien**. [S.l.]: R. Oldenbourg, 1877. DOI: 10.1515/9783486724653.

MCKINNEY, W. et al. Data structures for statistical computing in python. **SciPy**, v. 445, n. 1, p. 51–56, 2010. DOI: 10.25080/Majora-92bf1922-00a.

MUNZNER, T. **Visualization analysis and design**. New York: CRC press, 2014. 428 p. DOI: 10.1201/b17511.

MUSHTAQ, I.; KHAN, S. N. Factors affecting students' academic performance. **Global Journal of Management and Business Research**, v. 12, n. 9, p. 17–22, 2012. Disponível em: <https://journalofbusiness.org/index.php/GJMBR/article/view/100221>.

NAGY, M.; MOLONTAY, R. Interpretable dropout prediction: towards xai-based personalized intervention. **International Journal of Artificial Intelligence in Education**, Springer, v. 34, n. 2, p. 274–300, 2024. DOI: 10.1007/s40593-023-00331-8.

OQAIDI, K.; AOUHASSI, S.; MANSOURI, K. Towards a students' dropout prediction model in higher education institutions using machine learning algorithms. **International Journal of Emerging Technologies in Learning (Online)**, v. 17, n. 18, p. 103–117, 2022. DOI: 10.3991/ijet.v17i18.25567.

PACHAS, D. A. G.; GARCIA-ZANABRIA, G.; CUADROS-VARGAS, A. J.; CAMARA-CHAVEZ, G.; POCO, J.; GOMEZ-NIETO, E. A comparative study of who and when prediction approaches for early identification of university students at dropout risk. In: IEEE. **2021 XLVII Latin American Computing Conference (CLEI)**. [S.l.], 2021. p. 1–10. DOI: 10.1109/CLEI53233.2021.9640119.

PALLATHADKA, H.; WENDA, A.; RAMIREZ-ASÍS, E.; ASÍS-LÓPEZ, M.; FLORES-ALBORNOZ, J.; PHASINAM, K. Classification and prediction of student performance data using various machine learning algorithms. **Materials Today: Proceedings**, Elsevier, v. 80, n. 3, p. 3782–3785, 2023. DOI: 10.1016/j.matpr.2021.07.382.

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on**

Visualization and Computer Graphics, IEEE, v. 14, n. 3, p. 564–575, 2008. DOI: 10.1109/TVCG.2007.70443.

PEI, B.; CHENG, Y.; AMBROSE, A.; DZIADULA, E.; XING, W.; LU, J. Learningviz: a dashboard for visualizing, analyzing and closing learning performance gaps—a case study approach. **Smart Learning Environments**, Springer, v. 11, n. 1, p. 1–25, 2024. DOI: 10.1186/s40561-024-00346-1.

PORTELA, A. **Qualidade da educação está associada a maiores taxas de crescimento, revela estudo**. 2023. Portal FGV. Accessed on 14 May 2023. Disponível em: <https://portal.fgv.br/noticias/qualidade-educacao-esta-associada-maiores-taxas-crescimento-revela-estudo>.

RABELO, A. M.; ZáRATE, L. E. A model for predicting dropout of higher education students. **Data Science and Management**, v. 8, n. 1, p. 72–85, 2025. ISSN 2666-7649. DOI: 10.1016/j.dsm.2024.07.001.

RAO, J.; WANG, B.; LI, C. A k-means algorithm for personalized recommendation for public art education in higher education. In: SPRINGER. **The World Conference on Intelligent and 3D Technologies**. [S.l.], 2024. p. 443–452. DOI: 10.1007/978-981-97-9124-8$_3$6.

RIEHMANN, P.; HANFLER, M.; FROEHLICH, B. Interactive sankey diagrams. In: IEEE. **IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.** [S.l.], 2005. p. 233–240. DOI: 10.1109/INFVIS.2005.1532152.

SEMESP, I. **Mapa do Ensino Superior no Brasil 2024**. Semesp, 2024. Acesso em: 28 fev. 2025. Disponível em: <https://www.semesp.org.br/mapa/>.

SHAHBAZI, Z.; BYUN, Y.-C. Agent-based recommendation in e-learning environment using knowledge discovery and machine learning approaches. **Mathematics**, MDPI, v. 10, n. 7, p. 1192, 2022. DOI: 10.3390/math10071192.

SILVA, P. T. d. F.; SAMPAIO, L. M. B. Políticas de permanência estudantil na educação superior: reflexões de uma revisão da literatura para o contexto brasileiro. **Revista de Administração Pública**, SciELO Brasil, v. 56, p. 603–631, 2022. DOI: 10.1590/0034-761220220034.

SILVERMAN, B. W. **Density estimation for statistics and data analysis**. [S.l.]: Routledge, 2018. DOI: 10.1201/9781315140919.

SUSNJAK, T.; RAMASWAMI, G. S.; MATHRANI, A. Learning analytics dashboard: a tool for providing actionable insights to learners. **International Journal of Educational Technology in Higher Education**, Springer, v. 19, n. 1, p. 12, 2022. DOI: 10.1186/s41239-021-00313-7.

TENENBAUM, J. B.; SILVA, V. d.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000. DOI: 10.1126/science.290.5500.2319.

TINTO, V. Through the eyes of students. **Journal of college student retention: Research, theory & practice**, Sage Publications Sage CA: Los Angeles, CA, v. 19, n. 3, p. 254–269, 2017. DOI: 10.1177/1521025115621917.

TSUNG, S.; WEI, H.; LI, H.; WANG, Y.; XIA, M.; QU, H. Blocklens: visual analytics of student coding behaviors in block-based programming environments. In: **Proceedings of the Ninth ACM Conference on Learning@ Scale**. [S.l.: s.n.], 2022. p. 299–303. DOI: 10.1145/3491140.3528298.

VENKATESAN, R. G.; KARMEGAM, D.; MAPPILLAIRAJU, B. Exploring statistical approaches for predicting student dropout in education: A systematic review and meta-analysis. **Journal of Computational Social Science**, Springer, v. 7, n. 1, p. 171–196, 2024. DOI: 10.1007/s42001-023-00231-w.

WILCOX, P.; WINN, S.; FYVIE-GAULD, M. 'it was nothing to do with the university, it was just the people': the role of social support in the first-year experience of higher education. **Studies in higher education**, Taylor & Francis, v. 30, n. 6, p. 707–722, 2005. DOI: 10.1080/03075070500340036.

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

ZAKI, M. J.; MEIRA, W. **Data mining and analysis: fundamental concepts and algorithms**. [S.l.]: Cambridge University Press, 2014. DOI: 10.1017/CBO9780511810114.

ZHANG, H.; DONG, J.; LV, C.; LIN, Y.; BAI, J. Visual analytics of potential dropout behavior patterns in online learning based on counterfactual explanation. **Journal of Visualization**, Springer, p. 1–19, 2022. DOI: 10.1007/s12650-022-00899-8.

Appendix

APPENDIX A

Project files

This repository contains the source code, datasets, and documentation related to the development of ***LearnVis***, the visual analytics system created as part of this master's thesis. The project includes the implementation of interactive visualizations, data processing scripts, and analytical tools designed to support the exploration and analysis of academic data.

A.1 Repository Structure

The repository is structured as follows:

- **data:** contains the processed query data tables used in each layout.
- **database_scripts:** includes the SQL scripts used to build each data table mentioned above, as well as the backup file (.bak) of the database, which contains all the original tables and the additional tables created for the development of the layouts.
- **javascript:** contains the interactive layouts developed for the system.
- **python:** contains the script used to apply t-SNE for generating the final data view used in the *Student Behavior View* layout.
- **css:** contains custom visual styles applied to the layouts.
- **index.html:** the main HTML file that serves as the structure for the system's interface.

Link the repostiory: <https://github.com/angelicago/LearnVis_System.git>.