FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Virtual Choreographies of Learning: AI-Driven Discovery in K-12 Online American Schools

Eduardo Salé Areias

Master in Informatics and Computing Engineering

Supervisor: Prof. Fernando Cassola Marques

Second Supervisor: Prof. Dennis Beck

July 27, 2026

Virtual Choreographies of Learning: AI-Driven Discovery in K-12 Online American Schools

Eduardo Salé Areias

Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

President: Professor António Soares
Examiner: Professor Leonel Morgado
Supervisor: Professor Fernando Cassola

July 27, 2026

Resumo

Os modelos de ensino totalmente online para o ensino básico e secundário (K–12), como os disponibilizados pela Connections Academy, servem mais de 100 000 alunos nos Estados Unidos, muitos dos quais transitam do ensino tradicional devido a dificuldades académicas ou pessoais prévias. Embora os relatórios nacionais mostrem frequentemente resultados mais baixos em exames padronizados para os alunos online, estes ambientes acolhem uma proporção significativamente maior de alunos em risco, o que dificulta comparações justas. Apesar da grande escala do ensino online, existe pouca investigação detalhada sobre a forma como os alunos efetivamente navegam nestas plataformas, como evoluem os seus padrões de estudo e como essas rotinas se relacionam com o seu progresso de aprendizagem.

Esta dissertação aborda estas questões adaptando o enquadramento das coreografias de aprendizagem para modelar padrões de comportamento semanais baseados em proporções de ações de aprendizagem online. Devido à indisponibilidade dos registos institucionais proprietários, o projeto adotou uma abordagem de engenharia de dados, desenvolvendo um simulador para gerar um conjunto de dados sintético de larga escala, concebido para refletir as regras pedagógicas e os ritmos letivos de um ambiente K–12 totalmente online. Foi então construído um *pipeline* analítico completo: métodos não supervisionados, sobretudo o algoritmo k-means, foram usados para identificar perfis comportamentais distintos, enquanto um modelo supervisionado de Random Forest foi treinado para prever o sucesso dos alunos e extrair sinais comportamentais precoces.

A análise revelou quatro coreografias de aprendizagem distintas, indicando que, neste ambiente simulado, a qualidade e a intenção do envolvimento estão associadas ao sucesso académico mais do que o simples volume de cliques. O modelo preditivo identificou que sessões de estudo frequentes e distribuídas, bem como a participação síncrona, estão positivamente associadas à aprovação, ao passo que sinalizou a "navegação improdutiva" — como a consulta excessiva do painel e a verificação repetida de notas sem consumo de conteúdo — como um sinal de risco precoce. Em última análise, este trabalho contribui com um *pipeline* analítico validado e agnóstico à origem dos dados, totalmente preparado para processar dados institucionais reais e apoiar os educadores com informação acionável para intervenções pedagógicas atempadas.

Palavras-chave: Ensino Online K–12 • Coreografias de Aprendizagem • Mineração de Dados Educacionais • Dados Sintéticos • Machine Learning

Abstract

Fully online K–12 education models, such as those provided by Connections Academy, serve over 100,000 students across the U.S., many of whom transition from traditional schools due to prior academic or personal challenges. While national reports often show lower standardized test scores for online learners, these environments enroll a significantly higher proportion of at-risk students, complicating fair comparisons. Despite the massive scale of online learning, there is limited granular research on how students actually navigate these platforms, how their study patterns evolve, and how these routines relate to their learning progress.

This dissertation addresses these questions by adapting the learning choreographies framework to model weekly behavior patterns based on proportions of online learning actions. Due to the unavailability of proprietary institutional logs, the project adopted a data engineering approach, developing a simulator to generate a large-scale synthetic dataset designed to reflect the pedagogical rules and academic rhythms of a fully online K–12 environment. An end-to-end analytical pipeline was then constructed: unsupervised methods, primarily k-means clustering, were used to identify distinct behavioral profiles, while a supervised Random Forest model was trained to predict student success and extract early behavioral signals.

The analysis surfaced four distinct learning choreographies, indicating that within this simulated environment, the quality and intent of engagement are associated with academic success more than the sheer volume of clicks. The predictive model found that frequent, distributed study sessions and synchronous participation are positively associated with passing, while it flagged "unproductive navigation"—such as excessive dashboard refreshing and repeated grade-checking without content consumption—as an early risk signal. Ultimately, this work contributes a validated, data-agnostic analytical pipeline, fully prepared to process real institutional data and to support educators with actionable insights for early pedagogical interventions.

Keywords: K–12 Online Learning • Learning Choreographies • Educational Data Mining • Synthetic Data • Machine Learning

The United Nations Sustainable Development Goals

The United Nations **SDG!**s (**SDG!**s) provide a global framework for achieving a better and more sustainable future for all. It includes 17 goals (`https://sdgs.un.org/goals`) to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

This dissertation contributes to SDG 4 by developing and validating an analytical pipeline, using a large-scale synthetic environment, to support pedagogical interventions and improve student outcomes in K-12 online learning.

The specific Sustainable Development Goal addressed in this work is:

SDG 4 Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.

SDG	Target	Contribution	Performance Indicators and Metrics
4	4.1	Developing a robust analytical pipeline, validated on synthetic data, to identify at-risk behaviors in K-12 online schools, providing a tool for early pedagogical interventions.	Identification of behavioral risk signals and early prediction of student success.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors, Professor Fernando Cassola Marques and Professor Dennis Beck. Thank you for your continuous guidance, patience, and invaluable feedback throughout this journey. Your expertise and encouragement were fundamental, especially during the challenging pivot to a synthetic data methodology and the development of the analytical pipeline.

I would also like to thank the faculty at the Faculdade de Engenharia da Universidade do Porto (FEUP) for providing a strong academic foundation over the course of my Master's degree.

To my girlfriend, my family, and my friends, thank you for your unwavering support, love, and motivation during the countless hours of coding, writing, and researching. I could not have reached this milestone without each of you by my side.

Eduardo Salé Areias

Declaração de Honra

Declaro que a presente dissertação é de minha autoria e não foi previamente apresentada e avaliada nesta ou em outra instituição. As referências a outros autores (afirmações, ideias, pensamentos), assim como a utilização de trabalhos publicados respeitam escrupulosamente as regras da atribuição, e encontram-se devidamente indicadas no texto e nas referências bibliográficas, de acordo corn as normas de referenciação. Tenho consciência de que a prática de plágio ou de autoplágio constituem ilícitos académicos, conforme estabelecido no Código de Ética da U.Porto.

Declaro, ainda, que utilizei ferramentas de Inteligência Artificial para a realização de parte(s) da presente dissertação, e que foram objeto de cuidada verificação para deteção de erros, imprecisões, atribuições infundadas ou outras formas de enviesamento de conteúdos e argumentos, bem como de determinação de autoria.

Eduardo Salé Areias

Eduardo Areias

“The goal is to turn data into information, and information into insight.”

Carly Fiorina

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

ANOVA Analysis of Variance

ARI Adjusted Rand Index

AUC Area Under the ROC Curve

EDM Educational Data Mining

EPM Educational Process Mining

FEUP Faculdade de Engenharia da Universidade do Porto

IEP Individualized Education Program

K-12 Kindergarten to 12th Grade

LA Learning Analytics

LMS Learning Management System

ML Machine Learning

NEPC National Education Policy Center

NLP Natural Language Processing

NMI Normalized Mutual Information

ROC Receiver Operating Characteristic

ROC-AUC Receiver Operating Characteristic – Area Under the Curve

RQ Research Question

SHAP SHapley Additive exPlanations

t-SNE t-Distributed Stochastic Neighbor Embedding

UMAP Uniform Manifold Approximation and Projection

URL Uniform Resource Locator

Chapter 1

Thesis Manifest

1.1 Context

Connections Academy, supported by Pearson Online and Blended Learning, operates fully online public schools serving K–12 students across 32 U.S. states and enrolling over 100,000 learners. Independent national reviews repeatedly find that students in full-time virtual schools, on average, underperform peers in traditional brick-and-mortar public schools on standardized measures [11]. However, the literature suggests these schools often serve a distinct population of “at-risk” students who have opted out of traditional schooling due to medical, social, or academic challenges [3]. These complex demographics make it crucial to understand what students actually *do* in virtual classrooms beyond simple attendance and how their behavioral routines relate to learning progress.

1.2 Motivation

Despite the scale of full-time online schooling, there is limited empirical detail about students' fine-grained learning activities: how they navigate content, participate in live sessions, and interact with teachers and peers and how these behaviors connect to outcomes [9]. Addressing this evidence gap can inform practice in large online school networks and help target support to the students who need it most, particularly those whose engagement patterns signal early risk of failure.

1.3 Problem

This dissertation investigates *virtual choreographies*—operationalized in this work as weekly behavioral patterns based on action proportions—as a lens to characterize K–12 students' behavior at scale and to examine how such patterns are associated with academic success. Concretely, the aim is to identify, describe, and relate behavior patterns observed in activity logs to indicators of performance in fully online schools, utilizing the platform-independent framework proposed by Cassola et al. [7].

1.4 Research Questions

RQ1 Identification. Which *virtual choreographies* (operationalized as weekly proportions of online learning actions) can be reliably discovered in synthetic activity data modeling a fully online K–12 environment using unsupervised clustering, and how stable are they over time?

RQ2 Robustness across contexts. To what extent do the discovered choreographies replicate across grade levels and geographic states (i.e., are they consistent behavioral habits rather than state- or grade-specific artifacts)?

RQ3 Association with outcomes. How are choreography *attributes*: such as temporal regularity, balance between synchronous and asynchronous participation, and frequency of feedback—associated with indicators of student success (e.g., progress, grades), when controlling for student background?

RQ4 Predictive utility. Do choreography-based representations improve early prediction of student success over standard activity metrics, and how early in the semester can such signals be detected?

RQ5 Actionability and Visualization. Which components within the choreographies (specific actions, transitions, or rhythms) most contribute to outcomes, and how can these be visualized to support teacher decision-making?

Given the unavailability of real institutional data, this work develops and validates the proposed approach on a synthetic dataset, as detailed in Chapter 4.

1.5 Hypotheses

Definition (Virtual Choreography). A *virtual choreography* is the structured set of actions carried out among agents (e.g., student, teacher, platform tools) that unfolds over time and within a specific learning space, abstracting specific platform clicks into semantic behaviors [7]. In this study, this theoretical concept is practically mapped through the aggregation of weekly action frequencies.

H1 — Regularity. Choreographies exhibiting stable temporal rhythms (e.g., consistent study times and revisits across the week) are positively associated with academic success.

H2 — Balanced participation. Choreographies that balance synchronous participation (live sessions) and asynchronous work (content study, assignment submission) relate to better outcomes than predominantly one-sided patterns.

H_Robustness — Context Stability (Addressing RQ2). We hypothesize that core behavioral habits will remain stable and consistently distributed across different geographic states and K–12 grade levels, reflecting underlying learning profiles rather than localized phenomena.

H3 — Feedback-centric interaction. Choreographies with frequent teacher-student feedback exchanges (asking questions, reading feedback, revising submissions) are positively associated with success.

H4 — Purposeful navigation. Choreographies characterized by efficient, task-aligned navigation (timely access to resources, limited detours) are associated with better progress than fragmented, sporadic engagement.

H5 — Risk signals (inverse). Choreographies marked by long inactivity gaps, last-minute bursts, or minimal interaction with teachers/peers correlate with weaker outcomes.

Chapter 2

Literature Review

2.1 Overview

This chapter describes the methodology used to identify and select the scientific literature relevant to this dissertation. It explains the type of literature review adopted, the search strategy, the inclusion and exclusion criteria, and the process followed to screen and select the final set of papers that compose the document corpus. In line with the thesis manifest, the review focuses on research about K–12 online learning and full-time virtual schools in the United States, with particular emphasis on their effectiveness, challenges and implications for students, teachers and school systems. This corpus provides the conceptual and empirical background needed to frame the later analysis of virtual choreographies of learning in Connections Academy schools.

2.2 Type of review

Given the objectives of this dissertation, a structured literature review was conducted, following principles of a systematic literature review while allowing some flexibility that is closer to a narrative synthesis. The goal of the review is to identify, select and document existing research on K–12 online learning and virtual schools, so as to support the definition of the dissertation research questions and to provide a solid basis for the later State of the Art chapter and for the empirical analysis of student activity data. The corpus prioritizes contemporary systematic reviews and empirical studies that reflect the current state of K–12 online learning technologies. Instead of early foundational work, the review relies on recent comprehensive analyses, such as the systematic reviews by Martin et al. [10] and Johnson et al. [9]. It also incorporates critical studies on student outcomes, engagement and school-level implementation of K–12 online programmes, specifically those addressing at-risk populations as explored by Beck [3, 4] and Toppin and Toppin [13], as well as earlier warnings about the maturity of the field [2].

2.3 Search strategy

The literature search was carried out in 2025, in the months leading up to this deliverable. The primary search engine used was Google Scholar. While discipline-specific databases like ERIC or Scopus are valuable, Google Scholar was selected as the primary source for its broader indexation of interdisciplinary research, capturing relevant studies at the intersection of Computer Science (Educational Data Mining) and Education that might otherwise be fragmented across distinct databases.

In addition, two discovery tools, Litmaps and ResearchRabbit, were used to expand the initial set of articles through citation-based exploration and visualisation of related work. The search targeted peer-reviewed journal articles, conference papers and book chapters written in English. Given the rapid evolution of educational technology and the impact of the COVID-19 pandemic, the search prioritized studies published between **2015 and 2025**. This timeframe ensures that the analyzed studies reflect the current technological infrastructure (modern LMS and learning analytics capabilities) and the post-pandemic reality of K–12 online education.

The main search string used in Google Scholar was iteratively refined, but was based on combinations of the following keywords:

- **Core topic terms:** “K–12 online learning”, “virtual schools”, “virtual schooling”, “online teaching”, “online school”, “distance education”.
- **Outcome / focus terms:** “effectiveness”, “student achievement”, “engagement”, “school choice”, “best practices”, “teacher training”, “at-risk students”.
- **Context terms:** “K–12”, “secondary education”, “United States”, “high school”.

An example of a typical query used in Google Scholar is:

> “K–12 online learning” OR “virtual schools” OR “virtual schooling” AND (effectiveness OR achievement OR engagement OR “distance education”)

The initial Google Scholar searches generated a broad set of potentially relevant results. These references were exported and then imported into Litmaps and ResearchRabbit. From this initial seed set, both tools were used to identify additional papers that frequently cite or are cited by the seed articles, visualise clusters of research, and surface key works. Through this process, an initial pool of approximately 150 records was assembled before applying the inclusion and exclusion criteria described below.

2.4 Inclusion and exclusion criteria

To ensure the relevance and quality of the selected literature, explicit inclusion and exclusion criteria were defined and applied during the screening process.

2.4.1 Inclusion criteria

A study was included in the corpus if it met all of the following criteria:

- **IC1** – The study focuses on K–12 online learning, virtual schools or virtual schooling (fully online or primarily online programmes).
- **IC2** – The study is peer-reviewed (journal article, conference paper or book chapter).
- **IC3** – The study is written in English.
- **IC4** – The study was published between **2015 and 2025**.
- **IC5** – The study addresses at least one of the following aspects: effectiveness or outcomes of virtual schooling, teaching practices, student engagement, or system-level issues such as school choice and policy.

2.4.2 Exclusion criteria

Studies were excluded if any of the following applied:

- **EC1** – The main focus is higher education, adult education or corporate training.
- **EC2** – The publication type is a thesis, dissertation, blog post, or other non-peer-reviewed document (with the exception of high-impact institutional reports, such as the NEPC annual report).
- **EC3** – The study is not available in full text.
- **EC4** – The study deals with general educational technology or blended learning without a clear focus on fully or primarily online K–12 programmes.
- **EC5** – The paper is purely theoretical or opinion-based without sufficient connection to K–12 online or virtual schooling.

2.5 Screening and selection process

The selection process followed three main stages: (i) title and abstract screening, (ii) full-text assessment, and (iii) citation-based expansion and refinement.

2.5.1 Stage 1: Title and abstract screening

After removing duplicates, approximately 150 unique records remained. Titles and abstracts were screened against the criteria. Studies clearly focusing on higher education or non-online contexts were removed, leaving approximately 60 papers.

2.5.2 Stage 2: Full-text assessment

The full text of the remaining studies was examined. Papers that mentioned online learning only superficially were excluded. This stage reduced the set to a core group of studies providing substantial evidence, including recent systematic reviews [9, 10] and research on at-risk students [4, 5]. Around 20 papers remained.

2.5.3 Stage 3: Citation-based refinement

The core set was used for backward and forward snowballing using Litmaps. After applying the same criteria to these additional articles, the final corpus consisted of **12 highly relevant papers**.

2.6 Summary of Selected Studies

To provide an overview of the state-of-the-art, Table 2.1 summarizes the key papers selected for the final corpus.

Author (Year)	Type	Focus	Key Insight
Martin et al. (2020)	Review	Effectiveness	Effectiveness depends on design/support, not just the medium.
Johnson et al. (2022)	Review	Teaching Practices	Importance of teacher facilitation in online settings.
Beck (2023, 2024)	Empirical	At-Risk Students	At-risk students perform better with substantial "advocate" support.
Curtis & Werth (2015)	Empirical	Engagement	Transactional distance leads to isolation; needs interaction.
Molnar et al. (2023)	Report	Policy	Virtual schools often lag in graduation rates compared to traditional.
Toppin & Toppin (2016)	Analysis	Demographics	Virtual schools attract students with prior academic/social issues.

Table 2.1: Summary of the core literature corpus.

2.7 Effectiveness of K–12 Online Learning

A central theme in the selected literature is the comparative effectiveness of full-time virtual schools versus traditional brick-and-mortar settings. Large-scale reports consistently highlight performance gaps [11], noting lower graduation rates. However, recent reviews [9, 10] caution against binary comparisons, suggesting effectiveness is dependent on instructional design and support. Johnson et al. [9] note that while average performance may be lower, online learning provides essential opportunities for credit recovery and advanced coursework.

2.8 Supporting At-Risk Populations

An important finding is that the demographic profile of virtual schools differs significantly from traditional ones. Toppin and Toppin [13] and Beck [3] observe that virtual schools frequently attract "at-risk" students: those with prior bullying, medical issues, or academic failure. Success for these learners relies heavily on support beyond the screen. Beck's research [4, 5] highlights the pivotal role of the "on-site facilitator" (typically a parent). Beck and Levine [5] found that parental engagement was a substantial predictor of success than many course-level variables during the pandemic.

2.9 Engagement and Interaction Patterns

Student engagement is cited as the primary predictor of retention. Curtis and Werth [8] identify strategies to foster engagement, noting that "transactional distance" leads to isolation. However, a methodological gap exists: most studies rely on surveys. There is limited research utilizing fine-grained log data to understand *temporal patterns* of engagement. This gap underscores the need for the data-driven approach proposed in this dissertation.

2.10 Limitations and threats to validity

The review is subject to limitations. First, Google Scholar does not index all databases, potentially missing some studies. Second, the search was limited to English publications. Third, the focus on U.S. virtual schools excludes international contexts. Finally, subjective judgement was involved in screening. These limitations will be considered in the State of the Art analysis.

Chapter 3

State of the Art

3.1 Overview

While the previous chapter focused on the educational context of K–12 online learning and the needs of at-risk populations, this chapter reviews the technological landscape. Specifically, it addresses the field of Educational Data Mining (EDM) and the computational methods available for discovering patterns in student behavior. The chapter defines the scope of EDM, contrasts traditional Educational Process Mining with the proposed *Virtual Choreographies* framework, and reviews the unsupervised machine learning algorithms selected for this study.

3.2 Educational Data Mining

Educational Data Mining (EDM) is defined as the area of scientific inquiry centered on the development of methods for making discoveries within the unique kinds of data that come from educational settings [12]. In their comprehensive survey, Romero and Ventura [12] distinguish EDM from Learning Analytics (LA):

- **Learning Analytics (LA):** Often focuses on human-led decision making (e.g., visual dashboards for teachers) and relies on statistics.
- **Educational Data Mining (EDM):** Emphasizes automated discovery and the development of algorithms to find hidden patterns without human intervention.

In the context of K–12 online schools, EDM is typically applied to three main tasks [12]:

1. **Prediction:** Using historical data (grades, login frequency) to forecast student outcomes or dropout risk.
2. **Structure Discovery:** Finding underlying structures in data, such as grouping students with similar learning strategies.
3. **Relationship Mining:** Identifying relationships between variables.

This dissertation focuses primarily on *Structure Discovery*, using unsupervised learning to find patterns not immediately obvious to teachers.

3.3 Modeling Student Behavior

To analyze student engagement beyond simple metrics, researchers traditionally use techniques to model the strict *sequence* of student actions, though aggregated behavioral patterns can offer an alternative perspective.

3.3.1 Educational Process Mining (EPM)

A dominant approach is Educational Process Mining (EPM). As detailed by Bogarín et al. [6], EPM applies process mining techniques to educational data, treating learning as a series of events (e.g., `Quiz Start` → `Quiz End`). While effective for structured tasks, Bogarín et al. [6] note that EPM faces challenges in flexible environments. In full-time virtual schools, EPM often results in "spaghetti models": complex, tangled diagrams that are difficult to interpret.

This lack of readability makes such models less suitable for teachers who need actionable insights (addressing RQ5), necessitating a higher-level abstraction like Virtual Choreographies.

3.3.2 The Virtual Choreographies Framework

To address the limitations of rigid process maps, this dissertation adopts the concept of **Virtual Choreographies**. Cassola et al. [7] define a *Virtual Choreography* as a platform-independent representation of actions, interactions, and events that unfold over time. Unlike raw clickstreams, a Virtual Choreography maps technical events into **semantic actions** (e.g., mapping `URL:/math/quiz/1` to `Assessment`). This abstraction allows for:

- **Platform Independence:** The analysis focuses on the behavior, not the system.
- **Routine Discovery:** While theoretically grounded in temporal sequences, in this work, this abstraction is practically operationalized by aggregating actions into weekly proportions. This enables the identification of clusters of students who share similar behavioral routines (e.g., "Late Night Crammers" vs. "Steady Workers") regardless of the specific course.

3.4 Unsupervised Learning Techniques

The discovery of these choreographies relies on unsupervised machine learning algorithms.

3.4.1 Clustering Algorithms

Clustering groups objects so that objects in the same group are more similar to each other than to those in other groups. In EDM, the most widely used algorithm is **k-means**, due to its computational efficiency and interpretability [12]. However, k-means requires the number of clusters (k) to be specified, necessitating the use of validation metrics such as the Elbow Method or Silhouette Score.

3.4.2 Dimensionality Reduction

Student behavioral data is often high-dimensional. To visualize clusters effectively, dimensionality reduction is required. **t-SNE** and **UMAP** are state-of-the-art non-linear techniques used to project high-dimensional data into 2D space. These visualizations allow researchers to "see" the separation between different student groups and visually assess whether the identified choreographies represent distinct behavioral patterns. For this work, UMAP was preferred over t-SNE. While both are non-linear techniques, t-SNE excels at preserving local structure — grouping similar points together — but does not preserve global distances well. UMAP, by contrast, retains the overall structure of the data, ensuring that profiles which are conceptually opposite, such as the Steady Worker and the Low Engagement choreographies, appear distinct in the 2D map. UMAP also scales to thousands of student-weeks in a fraction of the time, which made it a practical advantage for this study.

3.5 Summary

Current research highlights the need for better support systems for at-risk students. Technologically, while EPM offers tools to analyze sequences, it can be overly rigid for online schooling. By adapting the *Virtual Choreographies* framework [7] and combining it with standard EDM clustering techniques [12], this thesis aims to discover interpretable patterns of student behavior that can inform the educational strategies discussed in the literature review.

Chapter 4

Methodology

4.1 Overview

This chapter describes the methodological process used to analyze, characterize, and predict the academic success of students in fully online schooling. The pipeline comprises these stages: data extraction and preparation, pattern discovery through k-means clustering and UMAP visualization, association of the discovered patterns with academic outcomes, and prediction of student success using a Random Forest model.

4.2 Synthetic Data as a Methodological Choice

The data used in this dissertation is synthetic. The real datasets were expected first from Pearson and subsequently from Stride Inc., but neither could be delivered in time. The synthetic dataset had originally been created as a working substrate to allow development to proceed while the real data was pending; once it became clear the real data would not arrive, continuing with the synthetic data was the only viable path.

This choice carries a significant methodological advantage. Because the dataset was generated by the author, the true behavioral archetype of each student is known by construction. This ground truth, which is never exposed to the clustering algorithm, makes it possible to measure whether the discovered clusters actually recover the underlying behaviors, using metrics such as the Adjusted Rand Index (ARI). With real, anonymized data, these true labels would be unknown, and such a quantitative validation of the clustering would not be possible.

A second advantage is openness. As the data does not belong to real students or institutions and was created entirely by the author, it can be used and published without privacy or licensing constraints, supporting full reproducibility of the pipeline.

This approach has an explicit limitation. Synthetic results cannot be validated in the same way as real data, and they do not constitute empirical claims about real Connections Academy students. What they validate is the methodology, the pipeline and analytical methods developed to carry out the study.

The generator's parameters were designed through informed intuition rather than calibrated against measured empirical distributions. The five archetypes were specified so that each would be behaviorally plausible and distinct from the others, drawing on the patterns of engagement and disengagement discussed in the literature reviewed earlier [4, 8]. They are not intended to reproduce exact empirical frequencies from real online schools, but to provide coherent, recognizable behavioral profiles against which the discovery pipeline can be developed and validated.

4.3 Synthetic Data Generation

The synthetic generator builds an academic year of activity for thousands of students, each assigned to one of five behavioral archetypes that describe how they study. The Steady Worker is a consistent, engaged student, focused on content study, assignments, and assessments. The Balanced Engager is the most social and interactive profile, driven by live sessions and discussion forums. The Night Crammer studies intensively, but concentrates activity late at night or in the early morning. The Minimal Browser mostly navigates pages, checks the dashboard, and looks up grades, with little actual studying. Finally, the Disengaged Ghost is a near-absent student, showing minimal activity of any kind.

Students are sampled from these archetypes with embedded correlations. At-risk students are more likely to be assigned to the lower-engagement archetypes, reflecting the weaker outcomes associated with these profiles. The dataset reproduces a realistic institutional structure: 32 schools across 32 U.S. states, spanning 13 grade levels from Kindergarten to Grade 12 (K–12).

Activity is generated as event-level logs, one student at a time. Each event records an action type, drawn according to the student's archetype, along with the time of day and, where applicable, the session duration. A fixed random seed (42) is used throughout the generation process to ensure full reproducibility.

4.4 Feature Engineering

While the theoretical concept of a learning choreography implies a sequence of actions over time, this methodology practically operationalizes it through the aggregation of weekly behavior patterns and proportions. The unit of analysis is the student-week: one row per student per calendar week. This choice provides enough data points for the clustering algorithm to find patterns, while letting each student be represented as a sequence of weeks across the year. Two alternatives were considered, the student-day and the student-year, but the former would generate an excessive volume of sparse data, and the latter would not retain enough observations per student. The student-week was the most balanced option.

Weekly features follow a two-block design. The WHAT block captures the proportion of each semantic action type performed during the week (for example, how much content study, assessment, or forum activity). The WHEN block independently captures the proportion of activity falling in each of four time-of-day bins: Morning, Afternoon, Evening, and Night.

The two blocks are kept separate rather than crossed. In an early attempt, action and time-of-day were combined into a single set of features, which caused the time dimension to dominate k-means: the resulting clusters separated students almost entirely by period of day rather than by behavior. Decoupling the two blocks resolved this and let the clustering focus on what students do, with time of day as a secondary signal.

Login and Logout actions are deliberately excluded from the WHAT block. They are universal, every student performs them, and therefore carry no discriminative value; they merely mark session boundaries. Finally, aggregate metadata (number of sessions, active days, total actions) is computed but kept outside the clustering feature space, so that students are grouped by the shape of their behavior rather than by sheer activity volume.

Pattern Discovery (Clustering) At this stage, the goal of clustering is to discover the shape of behavior, for which k-means was chosen as a clustering algorithm widely used in educational data mining [12]. Because k-means groups data by computing mathematical distances between points, raw action counts would let differences in scale dominate: the algorithm would simply separate students who click a lot from students who click little, failing to capture the actual behavioral patterns. To prevent this, each weekly vector is L1-normalized into proportions. After normalization, two students with the same behavioral mix, for instance, both spending half their activity on content study and half on forums, are treated as similar even if one performed many more actions than the other. Clustering therefore focuses on the composition of behavior rather than its volume.

Standardization is then applied so that each feature contributes comparably to the distance computation, preventing frequent actions from overwhelming rarer but discriminative ones. This step, however, required care. Some action types are very rare, almost no student performs them, so their columns have a near-zero standard deviation. Standardizing divides by the standard deviation, so these columns would trigger division-by-zero errors that corrupt the model. The `safe_standardize` function avoids this by detecting near-constant columns (standard deviation below a small threshold) and neutralizing them to zero rather than dividing by them, so they simply carry no weight in the clustering.

The number of choreographies, k, was selected by evaluating solutions across a range of $k = 2$ to 10, computing both the inertia (Elbow method) and the Silhouette score. These automatic metrics did not, on their own, give a definitive answer. The Silhouette score favored sharp mathematical partitions, but the UMAP projection showed that behaviors merged continuously rather than forming isolated islands; relying on the metric alone would have split genuine profiles in half. Both $k = 4$ and $k = 5$ were therefore examined directly. Validation against the ground-truth archetypes showed that the two solutions recovered the original behaviors with essentially the same accuracy (ARI ≈ 0.53). The deciding factor was interpretability: at $k = 5$, two of the clusters were both low-engagement profiles, separated only by time of day, which made them redundant. The final choice of $k = 4$ retains the same fidelity to the underlying structure while producing distinct, non-redundant choreographies.

The final clustering was produced by fitting k-means with $k = 4$, ten initializations (`n_init` = 10), and a fixed random seed (42) to guarantee reproducible results. For visual diagnosis, the resulting clusters were projected into two dimensions using Uniform Manifold Approximation and Projection (UMAP), with n_neighbors $= 15$ and min_dist $= 0.1$. It is important to stress that UMAP was used strictly for human inspection: it allowed visual verification of whether the discovered groups separated distinctly, but had no influence whatsoever on the cluster assignments, which were determined solely by k-means. Removing the projection would leave the clusters unchanged.

4.5 Evaluation Design

RQ1 (Identification) was evaluated by comparing the algorithm's cluster assignments against the generator's ground truth. A contingency table (cluster $\times$ archetype) was built, and the purity of each cluster was computed, identifying the dominant student profile within each discovered group. To assess this correspondence quantitatively, two agreement metrics were applied: the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). These metrics quantify the extent to which the algorithm recovered the behavioral structure originally embedded in the synthetic dataset.

RQ2 (Robustness) was assessed along two dimensions. Temporal stability was measured over the full academic year using week-over-week ARI and the percentage of students retained in the same cluster from one week to the next. A transition matrix was built to map how students moved between choreographies; the concentration of values along the main diagonal indicates the degree to which students maintained consistent learning routines rather than shifting abruptly between profiles. Cross-context stability examined whether the discovered patterns were universal rather than demographic artefacts. The percentage distribution of each choreography was evaluated across grade levels and states, checking that the standard deviation across contexts remained small. Finally, to assess whether the shape of behavior changed with schooling level, the cosine similarity of each profile's behavioral signature was computed across grade bands (K–5 versus 6–12), measuring how closely the underlying structure of these profiles is preserved.

RQ3 (Association with Outcomes) was assessed by first identifying each student's dominant choreography, the profile in which they spent most of their weeks across the year. With students grouped by dominant profile, a one-way ANOVA was applied to their final grades to test whether the performance differences between profiles were statistically significant, and the eta-squared statistic was used to measure the magnitude of the effect. Finally, to make the comparison fairer, these results were stratified by the student's initial at-risk status. This made it possible to examine to what extent final grades reflected the learning routines themselves rather than difficulties the student already carried beforehand.

RQ4 (Predictive Utility) was evaluated by training a Random Forest classifier to predict whether a student would pass, using features drawn only from the first 4 and 8 weeks of the year. This strict temporal cutoff prevents data leakage: it ensures the model cannot "peek" at end-of-year behavior to anticipate the outcome. Two models were compared using the ROC-AUC metric: a baseline using only activity-volume features, and an experimental model adding the student's dominant choreography. ROC-AUC was chosen over accuracy because the classes are imbalanced (most students pass), which would make accuracy misleading.

RQ5 (Actionability) was addressed by analyzing the trained prediction model to identify which actions carried the most weight in the final outcome, measuring feature importance in two ways: Gini importance and permutation importance. Two approaches were used because Gini importance has a well-known bias: it tends to overvalue features with many distinct values, such as navigation clicks, which can be misleading. Permutation importance is more reliable, as it measures the actual drop in model performance when a feature is shuffled. Once the most important behaviors were identified, the direction of effect was examined, checking whether a high proportion of a given action was associated with passing or failing, which yields the practical, actionable insights about which behaviors to encourage. Mean absolute SHAP values (via a TreeExplainer) were also considered as a means of corroborating these rankings.

4.6 Reproducibility and Implementation Details

The pipeline can be reproduced in full, from start to finish. The guiding rule was to fix the random seed (42) at every stage of the project, from the generation of the synthetic students to the train/test split in the Random Forest. This forces the computation to make exactly the same decisions on any machine, guaranteeing identical results across runs. The analysis is organized as a sequence of Python scripts that must be executed in a defined order: synthetic data generation, feature engineering, clustering, validation, stability analysis, outcome association, early prediction, and interpretation, documented in the repository. The entire work was made available in a public repository [1], relying only on standard Python libraries (pandas, numpy, scikit-learn, umap-learn, matplotlib, seaborn, scipy).

Chapter 5

Results

5.1 Overview

This chapter presents the empirical results of the study, structured sequentially around the five research questions: the identification of behavioral profiles (RQ1), their temporal and cross-context robustness (RQ2), their association with final grades (RQ3), their predictive utility (RQ4), and finally the interpretability of the signals of success and risk (RQ5). The aim of this section is to report the data and metrics objectively, leaving the deeper interpretation of these findings and their connection to the methodological context for the Discussion chapter.

5.2 RQ1: Identification of Choreographies

The first step was to decide how many groups to create. As illustrated in Figure 5.1, the automatic metrics did not point to a single definitive value. Since the UMAP projection showed that behaviors blend somewhat rather than forming isolated islands, both four and five groups were tested. Both reached essentially the same accuracy, so the simpler, more direct option was chosen: four groups to avoid redundant profiles.

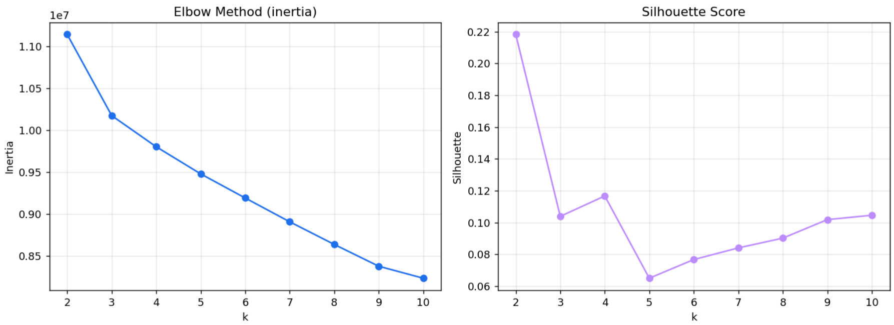

Figure 5.1: K-selection results: silhouette scores and inertia for $k = 2..10$. The silhouette peak at $k = 4$ supports the choice of four clusters.

As shown in Figure 5.2, clustering recovered the underlying structure with an Adjusted Rand Index of 0.532 and a Normalized Mutual Information of 0.593. These values are not perfect, but they indicate a good recovery: the model captured the main behavioral structure, while reflecting a natural overlap between students. This is visible in the purity of each group. The algorithm isolated the Balanced Engager with high purity (95.8%) and separated the Steady Worker (63.2%) and the Minimal Browser (56.8%) well. The Night Crammer, however, reached only 42.6% purity: studying late at night or early in the morning is such a dominant trait that the algorithm grouped together students who behaved differently during the rest of the day, as long as they shared this nocturnal timing.

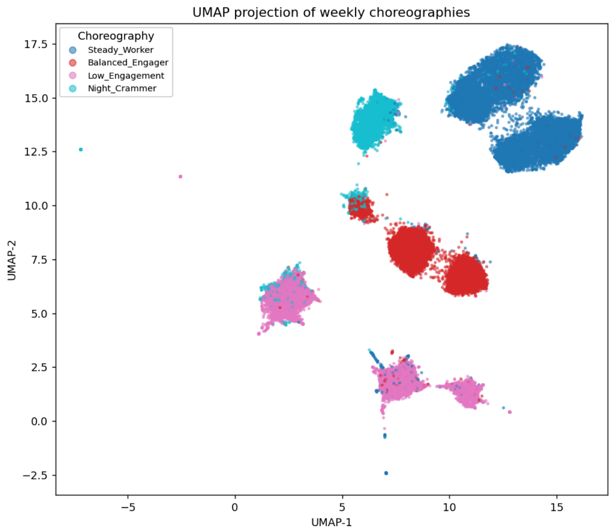

Figure 5.2: UMAP projection of the four discovered choreographies.

5.3 RQ2: Robustness (Temporal and Cross-Context)

The temporal stability analysis showed that students maintain a high degree of consistency in their learning patterns throughout the year (see Figure 5.3). This translated into a week-over-week Adjusted Rand Index (ARI) of around 0.80 and a same-profile retention rate of roughly 91%. The only marked fluctuation occurred during the winter break, where the ARI dropped to about 0.50, reflecting the disruption of routines typical of a holiday period. However, it is important to acknowledge that this exceptionally high baseline stability is partially a structural artifact of the synthetic data generation process, wherein students are simulated from a fixed underlying archetype.

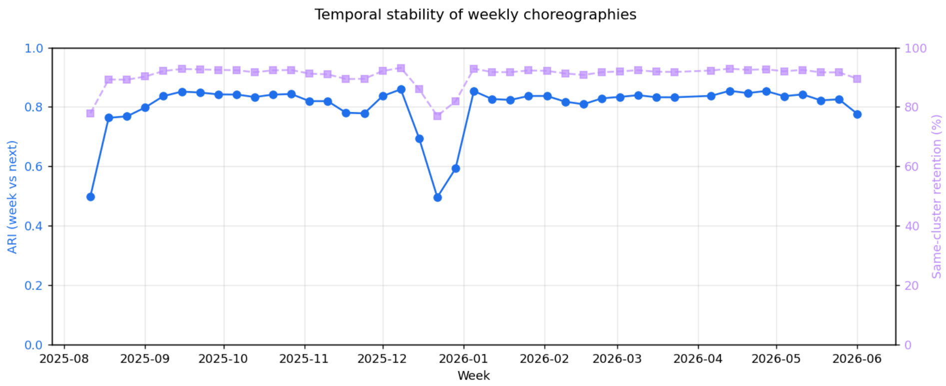

Figure 5.3: Temporal stability of weekly choreographies, displaying week-over-week ARI (blue solid line) and same-cluster retention rates (purple dashed line).

The transition matrix (Figure 5.4) confirmed the "stickiness" (or inertia) along the main diagonal of each group, but with differences between them. The Steady Worker was the most stable, retaining 95.6% of its members from one week to the next. The Balanced Engager showed solid retention at 91.9%, and the Low Engagement profile kept 87.9%. The Night Crammer proved the most volatile, retaining only 79.2% of its students, with many shifting to other behavioral patterns in subsequent weeks. Notably, when students did leave the Night Crammer profile, they moved most often into the Low Engagement profile (around 9% of transitions), suggesting a possible trajectory of disengagement over time.

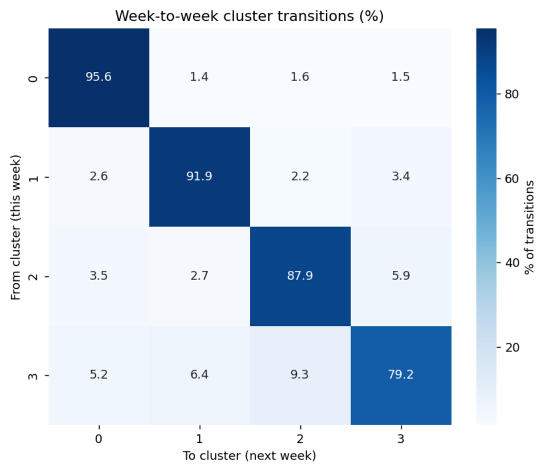

Figure 5.4: Week-to-week cluster transition matrix (%), highlighting the inertia along the main diagonal for each distinct behavioral profile.

The cross-context stability analysis demonstrated that the discovered learning patterns are remarkably consistent and independent of geography or student age. The presence of the four profiles was nearly identical across all regions and grade levels, as illustrated in Figures 5.5 and 5.6. The variation (spread) between contexts stayed consistently below 3.5 percentage points, indicating that the profiles appear uniformly rather than reflecting a specific year's curriculum or a particular state's policies.

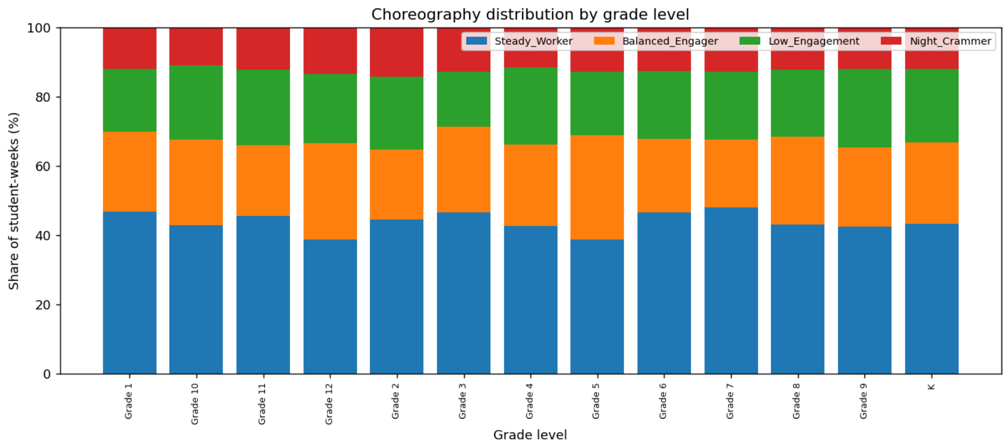

Figure 5.5: Choreography distribution by grade level, showing a highly consistent share of student-weeks across all K–12 grades.

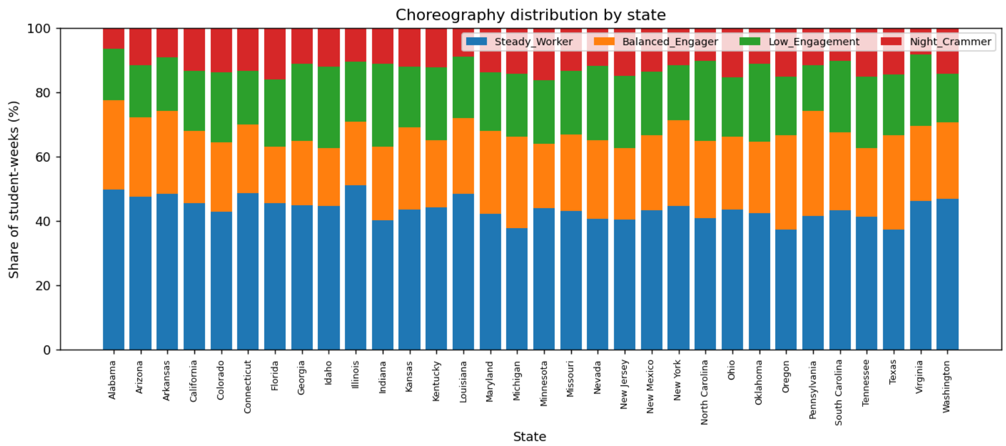

Figure 5.6: Choreography distribution by state, demonstrating uniform profile presence across different geographical regions.

Furthermore, when cosine similarity was used to compare whether the "shape" of platform behavior differed between younger students (elementary school, K–5) and older ones (secondary school, 6–12), the result was approximately 0.9999. This confirms mathematically that the behavioral signature of each profile remains essentially the same across schooling levels: the actions that define a Night Crammer in 3rd grade correspond to the exact same pattern of activity and timing in a 10th-grade student.

5.4 RQ3: Association with Outcomes

The analysis revealed a hierarchy in both average grades and pass rates across the four profiles (see Figure 5.7). At the top, with virtually identical and positive results, were the Balanced Engager and the Steady Worker. The Balanced Engager led slightly, with an average grade of 74 and a pass rate of 82%, followed closely by the Steady Worker, with an average of 73 and 81% of its students passing. In an intermediate position, the Night Crammer recorded an average grade of 69 and a lower pass rate of 70%. At the bottom was the Low Engagement profile: students with this dominant choreography reached an average grade of only 62 and a pass rate of 52%, meaning barely more than half of this group passed.

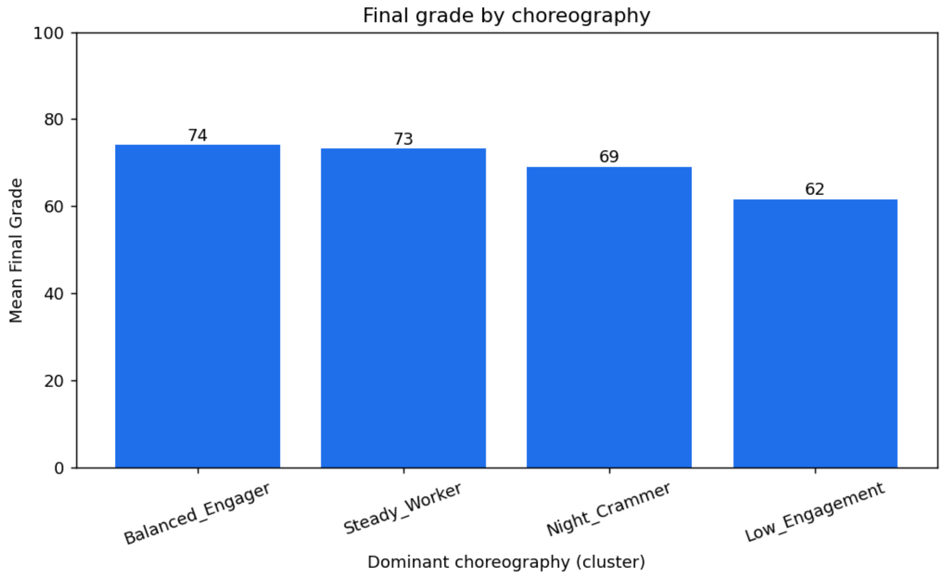

Figure 5.7: Mean final grade by dominant choreography, illustrating the performance hierarchy among the four behavioral profiles.

To assess the performance differences observed between the profiles, the ANOVA test offered two distinct perspectives, and it is essential to separate statistical significance from the actual magnitude of the effect. In terms of significance, the test returned a very high F-statistic (F = 907) with a p-value below 0.001, indicating that the variation in average grades between the four groups is statistically significant rather than due to chance. On its own, however, significance can be misleading: in large samples the test is so sensitive that almost any minor fluctuation produces a very low p-value, which could lead to overvaluing irrelevant differences. The effect size addresses this. The eta-squared value ($\eta^2 \approx 0.35$) measures the practical weight of the result and shows that the difference is genuinely large, meaning the behavioral profile explains roughly 35% of the total variance in students' final grades. Taken together, these metrics indicate not only that grade differences exist between the groups, but that they are substantial and positively associated with how students engage with the platform.

Introducing the at-risk history into the analysis added an essential layer of honesty to the results, revealing two important points (as illustrated in Figure 5.8). First, where a direct comparison was possible, at-risk status showed a negative association of its own, independent of how the student studied. Within the same choreography, at-risk students consistently scored about 10 to 12 points lower than their non-flagged peers: in the Steady Worker group, the average fell from 73.9

to 63.3, and in the Night Crammer group, from 73.4 to 61.3. Second, this analysis runs into a serious limitation in the distribution of the data. The Balanced Engager profile contained almost no at-risk students, while the Low Engagement profile was composed almost entirely of them. This lack of overlapping students across groups means that choreography and initial risk status are strongly confounded by construction. As a result, the two factors cannot be fully disentangled: it is not possible to determine with full rigor whether a Low Engagement student's failure is associated with their low-engagement routine itself, or whether it reflects the academic difficulties they already carried beforehand.

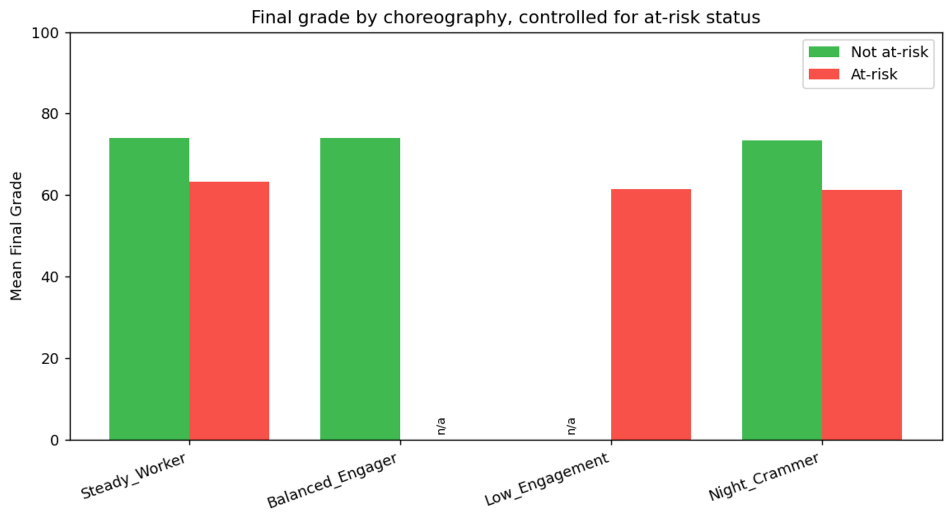

Figure 5.8: Mean final grade by choreography, controlled for at-risk status. Note the missing data (n/a) indicating the confounding between the Balanced Engager and Low Engagement profiles with baseline risk.

5.5 RQ4: Predictive Utility

Looking at the results objectively, the baseline model sits consistently above the experimental model, as shown in Figure 5.9. The numbers reflect this: at 4 weeks, the baseline reached an ROC-AUC of 0.825, while the experimental model (volume + choreography) fell slightly below at 0.824. At 8 weeks, the same pattern held, with the baseline at 0.838 and the experimental model at 0.833. The direct answer, therefore, is that the choreography did not improve the model's predictive performance. Adding the behavioral profile produced a negative AUC gain—in practice, the choreography added no predictive power over a simple count of interaction volume.

However, the RQ4 result carries a dual message, and the second part is highly positive. Early prediction of student success works and is reliable: in a setting with imbalanced classes, predicting the outcome as early as week 4 with an AUC around 0.83 is a solid result. What this experiment ultimately indicates is a principle of parsimony. The early-warning signal works well from the start of the semester, but for strictly predictive purposes, knowing *how much* a student works is sufficient, without needing to add complex information about *how* they work.

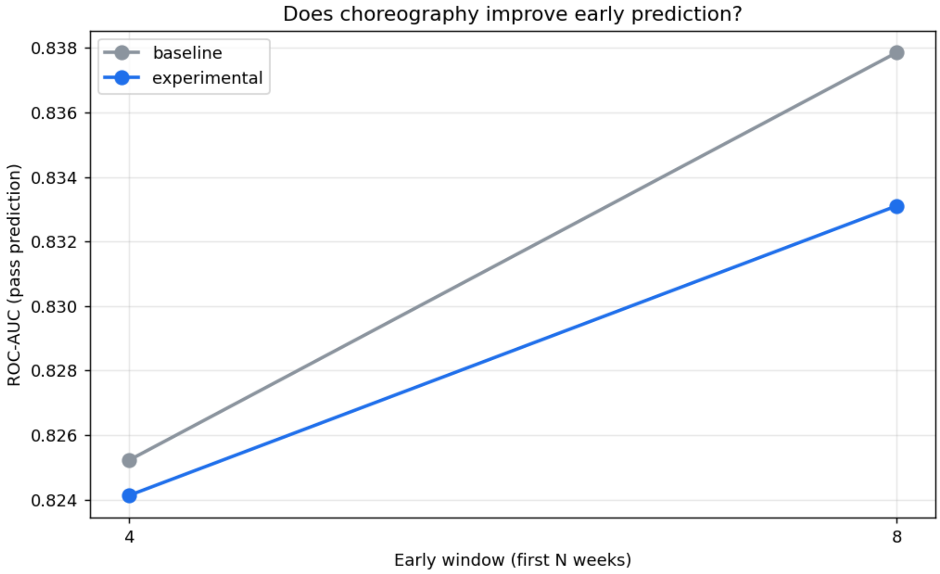

Figure 5.9: ROC-AUC comparison between the baseline model (volume only) and the experimental model (volume + choreography) at 4 and 8 weeks.

5.6 RQ5: Actionability and Interpretation

The Random Forest model (which achieved an ROC-AUC of 0.86) separates behaviors into two groups: indicators of success and signals of risk, as illustrated in Figure 5.10. On the success side, active participation and engagement volume emerge as the primary indicators of passing. The most prominent signal overall is the number of sessions: a high session count is associated with an increase in the probability of success from 71% to 98%. Starting and submitting assignments rank at the top of importance, corresponding to pass rates of 94–95% (compared to a 74% baseline), and joining synchronous sessions is associated with an increase from 80% to 93%. Consuming content, watching videos and studying course materials also proved positively associated with passing.

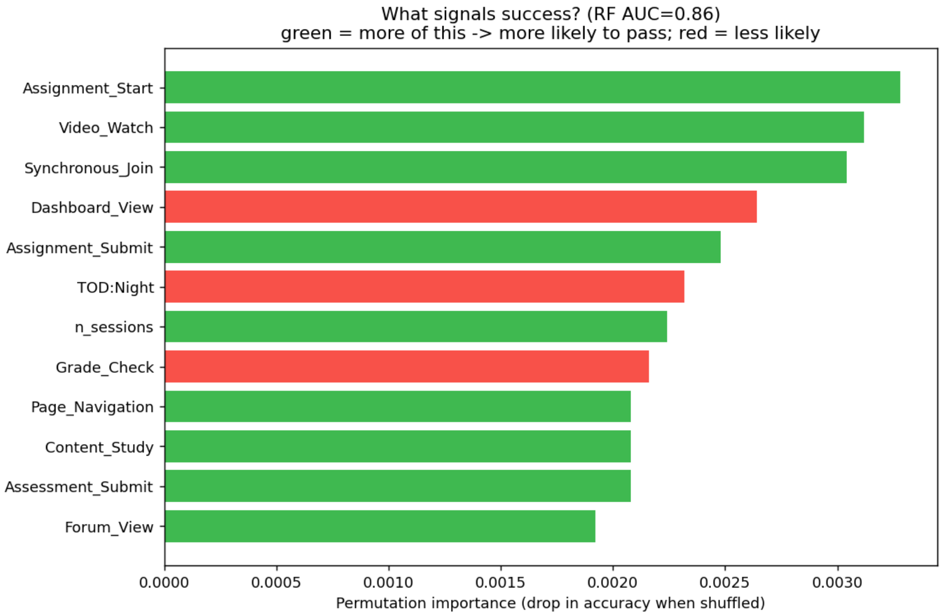

Figure 5.10: Permutation feature importance (drop in accuracy when shuffled). Green bars indicate behaviors that signal a higher likelihood of passing, while red bars flag risk signals.

On the risk side, the model surfaces the most counter-intuitive finding of the analysis. Studying late at night is associated with a higher risk of failure. More strikingly, the model flags "unproductive navigation": excessive visits to the dashboard and repeated grade-checking as risk indicators; students whose activity is dominated by this kind of clicking exhibit a lower pass rate (73% compared to 95%). The key lesson is that merely being "online" is not the same as producing work. A student who appears active on a teacher's dashboard, but whose activity is dominated by refreshing the home page and repeatedly checking grades, may in fact be caught in a cycle of unproductive monitoring and at high risk of failing, signaling a need for intervention.

It should be noted that the model identifies statistical associations rather than causal relationships. Furthermore, the distinctness of these behavioral signals partially reflects the underlying rules of the synthetic data generation process, a limitation that will be further addressed in the discussion.

Chapter 6

Discussion

6.1 Interpretation of Findings

The central message emerging from this discussion is that learning choreographies are consistently associated with academic outcomes, indicating that the quality and intent of a student's engagement distinguish success more than the mere volume of clicks. The behavioral patterns mapped here showed that risk signals, such as unproductive navigation, are analytically identifiable at an early stage. The need for rigor, however, requires acknowledging that these dynamics were observed in a synthetic data environment designed to simulate the routines of thousands of students. The main contribution of this dissertation is therefore the construction of a robust methodological and predictive pipeline, now fully prepared to receive the institution's real data and to inform pedagogical interventions in the field.

6.2 Revisiting the Hypotheses (H1–H5)

6.2.1 H1 — Regularity

The results support H1. The premise that stable, regular temporal routines are associated with academic success is borne out by the data on two fronts. In terms of the behavioral profiles (RQ3), the Steady Worker, which by design represents a student with a consistent, disciplined weekly cadence, achieved top results, with an average final grade of 73 and an 81% pass rate, making it one of the most effective routines against failure.

In terms of the predictive model (RQ5), the number of sessions stood out as the most prominent signal of success: activity spread across many sessions, which inherently reflects a regular and sustained routine rather than effort concentrated in a single day, raised the probability of passing to 98%. Conversely, late-night study, typically associated with irregular last-minute effort, was flagged as a risk signal.

Taken together, the results are consistent with H1, suggesting that consistency is more advantageous than sporadic effort, though it should be recognized that this association was, in part, built into the design of the synthetic data.

6.2.2 H2 — Balanced Participation

The results support H2. The premise that a balance between synchronous participation and asynchronous work translates into better outcomes is borne out by the data in two ways.

First, the Balanced Engager embodies this hypothesis: characterized precisely by combining independent study and assignment submission with attendance in live sessions and social interaction, it was the best-performing group in RQ3, with the highest average grade and pass rate, outperforming more one-dimensional patterns.

Second, in RQ5 the `Synchronous_Join` variable emerged as one of the primary predictive signals of success, alongside purely asynchronous actions such as `Assignment_Submit` and `Content_Study`, suggesting that both components contribute to positive outcomes [8].

The data therefore support this hypothesis without establishing causality. As with H1, it should be acknowledged that the consistency with which the Balanced Engager mirrors this theory partly reflects the synthetic generation process, which translated these theoretical premises into the dataset.

6.2.3 H3 — Feedback-centric Interaction

Looking honestly at the data and at the decisions made along the way, H3 cannot be considered adequately tested. The interaction metadata — which held the richest record of message exchanges and direct teacher–student feedback — was deliberately left out of the clustering, which was built exclusively on activity and navigation logs to define the choreographies.

Consistently, in the RQ5 analysis the `Feedback_Review` action does not appear among the variables with the greatest predictive power: the Random Forest gave priority to starting and submitting assignments, synchronous attendance, raw session volume, and the risk behaviors, leaving feedback consumption in a secondary statistical role. The data therefore do not provide evidence to confirm H3.

Rather than a weakness, this delimits a boundary of the present work and points directly to a promising direction for future research: incorporating the interaction metadata to test the role of feedback explicitly.

6.2.4 H4 — Purposeful Navigation

H4 finds substantial support in the results, but it calls for a methodological nuance. The core premise that purposeful navigation is associated with better outcomes is supported by the data. In RQ5, the most surprising finding was precisely that "unproductive navigation" acts as a prominent risk signal, dropping the pass rate from 95% to 73%. Consistently, the profiles dominated by this scattered behavior, such as Low Engagement, obtained the lowest grades and pass rates in RQ3.

There is, however, an important nuance. The model did not measure navigation efficiency in its strict, mechanical sense: the speed or directness of reaching a resource. What it captured was the qualitative difference between productive navigation, oriented toward coursework, and empty or anxious navigation.

The results thus support the premise of H4 while refining its focus: the real differentiating factor is not the efficiency of the path, but the productive intent of the navigation. Being caught in cycles of empty monitoring is detrimental to a student's progress.

6.2.5 H5 — Risk Signals

H5 is the hypothesis that finds the most consistent support across the analysis. The data support the premise that inactivity, late or irregular effort, and a lack of productive interaction are associated with the poorest outcomes, and this support appears at several stages of the results.

In RQ3, the Low Engagement profile, the embodiment of inactivity and minimal interaction, was by far the lowest-performing group, with an average grade of only 62 and a concerning pass rate of 52%. In RQ5, the predictive model captured exactly the behaviors described in the hypothesis: late-night study was flagged as a risk signal, supporting the danger of last-minute effort, while empty, anxious navigation dropped the probability of passing from 95% to 73%, showing that being connected without producing is nearly as associated with failure as total inactivity.

Finally, in RQ2, the transition matrix revealed the Night Crammer as the most unstable profile: when these students cannot sustain their nocturnal sessions, they tend to slide into the Low Engagement pattern, falling into the highest-risk group.

The data support H5, although the sharpness with which extreme disengagement and irregular study manifest themselves also derives from the architecture of the data generation.

6.3 Relation to Prior Work

In the literature, Beck [3, 4] has argued that the success of at-risk students in online education does not depend solely on what they do in front of the screen, but is deeply tied to external factors, in particular the presence of an "advocate". Our results are consistent with this view. By showing that, within the same study choreography, whether the disciplined routine of a Steady Worker or the effort of a Night Crammer, students flagged as at-risk consistently obtain lower grades, it becomes apparent that the behavior visible on the platform does not explain everything. This persistent disadvantage suggests that the student's background and the absence of an external support network continue to exert a negative association on performance, regardless of their dedication within the virtual environment. In this sense, our findings reinforce the idea that analyzing system metrics alone is insufficient for vulnerable students, and they support the importance of the family and social support structures emphasized by the author.

The results also show a parallel with the work of Curtis and Werth [8] on the concept of "transactional distance" and the risks of isolation in online education. According to these authors, the lack of interaction in virtual environments creates a psychological and communicational barrier, transactional distance, that severely harms student engagement and success. Our data reflect this perspective. The effectiveness of the Balanced Engager, the best-performing profile in the cohort (grade 74, 82% pass rate), illustrates the student who refuses isolation: by combining autonomous (asynchronous) study with social (synchronous) presence, this student reduces transactional distance and benefits from a richer learning environment. Consistently, in RQ5 participation in live sessions (`Synchronous_Join`) emerged as one of the main predictors of passing, rising from 80% to 93%. In this sense, our findings are consistent with the thesis of Curtis and Werth, suggesting that mitigating isolation through synchronous moments is associated not merely with a better "experience" but with a tangible increase in the likelihood of academic success. It should be noted, however, that this association was itself built into the synthetic data, so the results illustrate rather than independently confirm their argument.

Our results also establish a structural link with the conclusions of the systematic reviews by Martin et al. [10] and Johnson et al. [9]. The central thesis of these authors is that the effectiveness of online education is not dictated by the technological medium itself, but by how instructional design, facilitation, and student engagement are structured. Our data are consistent with this perspective. Observing that students immersed in the same learning environment show such disparate academic trajectories and success rates, contrasting the 81% pass rate of the Steady Worker with only 52% for the Low Engagement profile, suggests that the technology acts as a neutral stage. What truly shapes the outcome is not the platform, but the manner and quality of the student's engagement within it. In this way, our analysis reinforces the idea underlying these reviews: the focus of educators and researchers should not be on questioning the "medium", but on understanding and supporting the patterns of participation that take place inside it.

6.4 Limitations

The main technical limitations of this work center on five points.

The synthetic nature of the data. This is the central limitation. Although the dataset was generated from realistic logics and distributions, the data reflect the mathematical rules imposed on them. As a result, the sharpness of the profiles does not capture the full organic noise of human behavior, and the findings cannot be directly generalized to real students without subsequent validation.

Inflated robustness and stability (RQ2). As a consequence of the previous point, the high temporal consistency and the uniformity of the choreographies across different states and grade levels largely reflect the stability of the generator itself. The natural drift that local sociocultural factors would introduce in real life could not be observed.

Statistical confounding between choreography and initial risk (RQ3). The "empty cells" problem constrained this analysis. Because the Balanced Engager profile was generated with almost no at-risk students, and the Low Engagement profile was composed almost entirely of them, the two factors became mathematically entangled by construction. This prevented us from rigorously isolating what is attributable to the student's routine and what is the weight of their prior history.

Exclusion of the interaction metadata (H3). The hypothesis on the importance of feedback (H3) was not adequately tested. Because the direct interactions were excluded from the clustering in order to focus on the activity and navigation logs, feedback consumption did not carry enough weight in the analysis to confirm or reject its influence.

Coupling between volume and behavior (RQ4). The fact that the choreography added no predictive power over the baseline model illustrates a limitation in the separation of variables. In the way the routines were designed and generated, how a student studies is strongly coupled with how much they study, so the machine learning model could not extract useful new information.

6.5 Threats to Validity

External validity, which is the ability to transfer and generalize these findings to the real world, is most threatened by the synthetic nature of the dataset. Although the generation process respected plausible pedagogical logics and academic rhythms, the behavioral data are ultimately a product of predefined rules, probabilities, and mathematical distributions. They lack the organic noise, the human unpredictability, and the complex socioeconomic conditions that affect real students. It cannot therefore be assured that the profiles identified here, or their success metrics, would manifest identically in a real school population.

Mitigation of this threat is, however, structurally anticipated by the design of the project. The entire analytical pipeline, from the transformation of raw logs into student-week matrices, through clustering, to the predictive machine learning models was developed to be agnostic to the origin of the data. The natural mitigation will therefore be to feed this pipeline with the institution's real data once it becomes available, replicating the methodology to test its validity in the field.

Causality cannot be claimed; only associations can be reported. There are two concrete reasons for this. The first is confounding. As explored in RQ3, there is an inseparable overlap between the behavioral profile and the at-risk factor. With too few students to compare across all scenarios, the effect of each variable cannot be isolated, and it is impossible to state whether a student's failure was directly caused by a weak clicking routine or determined by their prior history of academic risk and vulnerability. The second, and more fundamental, reason is the circularity inherent to synthetic data. Because the dataset was generated by us, the relationships between behaviors and final grades were, to some extent, parameterized by the generator's rules. Claiming a pure cause-and-effect relationship in this context would amount to circular reasoning.

They are, inevitably, an approximation. Although the student-week matrix and the categorization of thousands of clicks into high-level semantic actions attempt to operationalize Cassola's theoretical concept of a learning choreography, log data capture only the student's digital trace on the platform. They do not measure genuine cognitive effort, the level of attention on the other side of the screen, or offline study: a student may be generating `Video_Watch` events while looking at their phone. Furthermore, the labels assigned to the groups are ultimately heuristic, qualitative interpretations based on the data distributions; another researcher could look at the same action proportions and assign slightly different names or pedagogical meanings. The main threat to construct validity, therefore, is assuming that click counts and timestamps represent the entirety of a student's learning routine, when they are in fact only its digital shadow.

The statistical inferences are broadly sound, and this is the category in which the study shows the greatest methodological robustness. To avoid drawing premature or illusory conclusions from the data, several deliberate decisions were made. First, a focus on effect size: in RQ3, using eta-squared rather than relying solely on ANOVA p-values avoided the classic large-sample trap, measuring the real magnitude of the effect rather than mere statistical significance. Second, fair metrics for imbalanced data: in RQ4 and RQ5, adopting ROC-AUC instead of plain accuracy was crucial — with a base pass rate of 87.6%, accuracy would have given a false sense of predictive power, whereas AUC evaluated the model by its genuine ability to distinguish the two classes. Third, unbiased feature importance: in RQ5, permutation importance was used instead of Gini impurity, neutralizing the algorithm's known bias toward inflating the importance of high-cardinality variables and ensuring that the extracted success and risk signals were genuine.

One honest threat nonetheless remains. The entire data generation, train/test split, and model initialization rested on a single random seed. While this ensures full reproducibility for anyone running the code, it also means the inferences depend on that specific initialization, and with other seeds the marginal numbers could fluctuate slightly. A more robust statistical analysis would require running multiple seeds to confirm that the conclusions hold beyond any variance introduced by initialization.

Chapter 7

Conclusion

7.1 Summary of Contributions

The main contributions of this dissertation fall along three axes: methodological, analytical, and practical. The value of the work lies not only in the results obtained, but in the architecture built to reach them.

Methodological contribution. The robust achievement of this thesis is the creation and validation of a complete, structured, and reproducible analytical pipeline. From the conversion of raw logs into weekly activity matrices, through clustering, and culminating in predictive modeling and interpretation, the entire system was designed to be strictly agnostic to the origin of the data. The main methodological legacy is therefore a fully operational analytical infrastructure, ready to process the school institution's real data.

Analytical contribution. Applying this pipeline to a large-scale synthetic environment made it possible to map four distinct learning choreographies and to show associations between behavioral patterns and academic outcomes. The thesis consolidates findings such as the identification of counter-intuitive risk signals and the demonstration that early prediction is feasible, while also reporting negative results with full transparency, namely that adding the choreography did not improve on the predictive power of a model based on activity volume alone.

Practical contribution. For an educational institution, this work translates into a tool for early warning and pedagogical management. The thesis shows that risk profiles can be identified within the first four weeks of the school year. By indicating that the quality of interaction and task-oriented navigation matter, it offers teachers a lens beyond mere "being online", enabling timely interventions to support students caught in cycles of unproductive monitoring or irregular effort.

7.2 Project Limitations

The external circumstances that shaped and constrained the course of this dissertation are distinct from the purely technical limitations and center on three logistical and operational factors.

The inaccessibility of institutional data. The principal limitation of this project was the failure to obtain real data from the institutions initially expected (such as Pearson or Stride). This logistical blocker forced a substantial restructuring of the entire dissertation. Rather than working on an existing school dataset, it became necessary to pivot toward data engineering, requiring the full development of a log simulator to make the research viable.

The compressed Master's timeline. A dissertation operates within a fixed and demanding time window. The substantial time invested in overcoming the absence of data, conceptualizing, coding, and validating the synthetic generator to ensure it reflected plausible school dynamics consumed a significant portion of the schedule. With data available from the outset, that time could have been channeled into exploring additional algorithms or into the exhaustive testing of hypotheses that ultimately remained secondary, such as the text analysis of interactions for H3.

The autonomous development of the infrastructure. In many Educational Data Mining contexts, or within large research teams, data extraction and architecture are already established. In this project, the entire ecosystem from the foundation of action generation to the final analytical and predictive pipeline had to be designed, built, and tested individually. While this enriched the technical robustness of the work, this engineering effort naturally limited the breadth of the final analysis.

7.3 Future Work

The avenues for future work follow naturally from the caveats and mitigations mapped throughout the dissertation. The next logical steps center on moving to a real-world setting, deepening the variables analyzed, and expanding the analytical toolkit.

Applying the infrastructure to real data. This is an immediate step. Since the entire pipeline was designed to be agnostic to the origin of the data, the central goal is to feed this architecture with the partner institution's operational logs as soon as they become available. This would close the project's loop, making it possible to test whether the choreographies and risk signals extracted from the synthetic environment manifest in a real school population.

Testing H3 through the interaction metadata. To address the limitation acknowledged in the Discussion, future work should recover the direct communication data and incorporate the teacher–student message flow into the clustering, possibly drawing on Natural Language Processing (NLP) techniques to assess the content of feedback. This would finally allow the role of feedback and transactional distance in relation to success rates to be examined directly.

Strengthening statistical robustness. To address the threat to conclusion validity introduced by the use of a single seed, future iterations of the predictive model should integrate mechanisms such as Monte Carlo cross-validation or the execution of multiple seeds, ensuring that the performance metrics (AUC) and feature-importance signals hold beyond any variance from algorithmic initialization.

Algorithmic exploration and sequential analysis. Finally, the analytical approach can be considerably expanded. Testing non-parametric clustering algorithms could capture student groupings with more complex contours than those identified by k-means. Moreover, a natural evolution of this work would be to move from proportional counts to process mining, analyzing the exact sequential order of clicks to understand not only what a student does in a session, but their step-by-step navigation trajectory.

References

[1] Eduardo Salé Areias. *Virtual Choreographies of Learning: AI-Driven Discovery in K-12 Online American Schools*. https://github.com/EduardoAreias1906/feup-thesis. GitHub repository. 2026.

[2] Michael K. Barbour. "Virtual Education: Not Yet Ready for Prime Time?" In: *The Test-Based Education Reforms: Lessons from a Failed Agenda*. Ed. by William J. Mathis and Tina Trujillo. Charlotte, NC: Information Age Publishing, 2016, pp. 407–429.

[3] Dennis Beck. "Be Prepared: Online School Experience and Student Achievement During the Pandemic". In: *Frontiers in Education* 8 (2023), p. 1161003. DOI: 10.3389/feduc.2023.1161003.

[4] Dennis Beck. "At-risk and online: Parent perceptions of at-risk learner's supports in a fully online school". In: *Journal of Research on Technology in Education* (2024). (Ahead-of-print). DOI: 10.1080/15391523.2023.2285497.

[5] Dennis Beck and J. Levine. "The role of the advocate in cyber schools during the COVID-19 pandemic". In: *Journal of Online Learning Research* 6.3 (2020), pp. 195–217. URL: https://www.learntechlib.org/primary/p/217172/.

[6] Alejandro Bogarín, Rebeca Cerezo, and Cristóbal Romero. "A survey on educational process mining". In: *WIREs Data Mining and Knowledge Discovery* 8.1 (2018), e1230. DOI: 10.1002/widm.1230.

[7] Fernando Cassola, Leonel Morgado, et al. "Using Virtual Choreographies to Identify Office Users' Behaviors to Target Behavior Change". In: *Energies* 15.12 (2022), p. 4354. DOI: 10.3390/en15124354.

[8] Heidi Curtis and Loredana Werth. "Fostering Student Success and Engagement in a K-12 Online School". In: *Journal of Online Learning Research* 1.2 (2015), pp. 163–190. URL: https://www.learntechlib.org/primary/p/151187/.

[9] Carla C. Johnson et al. "Online Teaching in K-12 Education in the United States: A Systematic Review". In: *Review of Educational Research* 93.3 (2022), pp. 353–411. DOI: 10.3102/00346543221105550.

[10] Florence Martin, T. Sun, and C. D. Westine. "A systematic review of research on K-12 online teaching and learning: Comparison of research from two decades 2000 to 2019". In: *Journal of Research on Technology in Education* 52.3 (2020), pp. 353–374. DOI: 10.1080/15391523.2019.1605825.

[11] Alex Molnar et al. *Virtual Schools in the U.S. 2023*. Boulder, CO: National Education Policy Center, 2023. URL: https://nepc.colorado.edu/publication/virtual-schools-annual-2023.

[12] Cristobal Romero and Sebastian Ventura. “Educational Data Mining and Learning Analytics: An Updated Survey”. In: *WIREs Data Mining and Knowledge Discovery* 10.3 (2020), e1355. DOI: 10.1002/widm.1355.

[13] Ian N. Toppin and Sheila M. Toppin. “Virtual schools: The changing landscape of K-12 education in the U.S.” In: *Education and Information Technologies* 21.6 (2016), pp. 1571–1581. DOI: 10.1007/s10639-015-9402-8.