Personalized Learning Model for Neurodivergent Students

Carly Carroll

The University of Rhode Island

May 15, 2025

Abstract

This study explores the development of personalized learning pathways for neurodivergent students using machine learning techniques. By analyzing behavioral engagement data such as click counts, study habits, and demographic variables, students were classified into engagement profiles and assigned tailored curriculum sequences designed to optimize learning outcomes.

The results indicate that adaptive pathways based on interaction patterns can enhance personalization in educational settings. While the students modeled in the study were not confirmed to be neurodivergent, the profiles and curriculum designs were guided by literature on neurodiversity and cognitive variation. There was a strict emphasis on the importance of inclusive model design and ethical considerations when working with diverse learners. Future work should validate these findings with explicitly neurodivergent populations and explore real-time learning adaptations.

Introduction

Background and Motivation

Personalized learning has recently emerged as a promising approach to modern education, leveraging technology to adapt instructional methods to the unique needs of individual learners. As education systems evolve, there is a growing recognition of the diverse way in which students process information, retain knowledge, and stay engaged in learning environments. Among the most impacted by the traditional educational methods are neurodivergent students, those who exhibit atypical neurological development. These students often face significant challenges in standard classroom settings, where uniform teaching methods can hinder their ability to thrive in academic settings.

Despite increased awareness of neurodiversity, the development of personalized learning models specifically designed for neurodivergent students remains limited. Current adaptive learning technologies primarily address general student engagement and performance without adequately accounting for the unique cognitive patterns characteristic of neurodivergent students. Addressing this gap requires a tailored approach that integrates machine learning with psychological principles to enhance educational outcomes.

Research Gap and Novelty

While adaptive learning has been widely explored in educational technology, most existing systems lack the refinement required to support neurodivergent learners. Studies have shown that attention patterns, executive functioning, and cognitive flexibility differ significantly between neurotypical and neurodivergent students, yet mainstream personalized learning tools often fail to accommodate these differences. There is a clear gap in the research regarding the development of machine learning models that cater specifically to the educational needs of neurodivergent individuals.

This paper addresses the gap by proposing a novel approach to personalized learning for neurodivergent students. The proposed model utilizes machine learning techniques, including reinforcement learning for adaptive feedback loops and cognitive modeling for attention and engagement patterns. Although the dataset does not include confirmed diagnosis, the model is designed with cognitive traits and behavioral patterns frequently associated with neurodivergent learners in mind. By incorporating insights from psychology and neurodiversity research, the model aims to optimize curriculum adjustment in a way that aligns with individual cognitive profiles.

Research Goals and Objectives

The primary goal of this paper is to develop and evaluate personalized learning models designed specifically for neurodivergent students. To achieve this, the study will focus on the following objectives:

1. Developing adaptive feedback systems using reinforcement learning to tailor responses based on individual responses.
2. Designing curriculum optimization techniques that incorporate insights from cognitive psychology and learning theory.
3. Modeling attention patterns and cognitive engagement to enhance focus and reduce cognitive overload.
4. Conducting empirical evaluations to assess the effectiveness of the proposed models in various learning environments.

Structure of the Paper

This paper begins by reviewing the current state of personalized learning models and the specific needs of neurodivergent students. It then presents the methodology used to develop the proposed learning models, followed by an analysis of their effectiveness based on empirical evaluations. The paper concludes with a discussion of the findings, implications for educational practice, and recommendations for future research.

Literature Review

Personalized learning for neurodivergent students is a complex and multifaceted area of research, drawing from fields as diverse as educational psychology, cognitive science, artificial intelligence, and educational technology. This section provides a comprehensive review of key studies that form the foundation of personalized learning systems for neurodivergent learners, focusing on neurodiversity in education, personalized learning systems, machine learning in education, cognitive and educational psychology, and the challenges and ethical considerations associated with these approaches. Despite increasing interest in inclusive education, few models have been empirically validated using data specific to neurodivergent populations. Many existing approaches rely on generalized behavioral proxies, limiting their ability to fully capture neurodivergent cognitive profiles. Understanding these interconnected disciplines is crucial for creating personalized strategies that address diverse neurodivergent needs.

Introduction to the Literature Review

The purpose of this literature review is to provide a critical analysis of the existing research on personalized learning models for neurodivergent students. This section covers foundational theories, current approaches, and emerging trends, while also identifying gaps in the literature that highlight opportunities for future research. The review is structured around five main themes: neurodiversity in education, personalized learning systems, machine learning in education, cognitive and educational psychology, and challenges and ethical considerations.

Neurodiversity in Education

Neurodiversity refers to the recognition and acceptance of diverse neurological conditions, such as autism, ADHD, and dyslexia, as natural variations in human cognition rather than deficits. Armstrong (2010) advocates for a strengths-based approach, emphasizing that neurodiverse students can excel when their unique abilities are nurtured. This perspective aligns with personalized learning by promoting tailored educational experiences that build on individual strengths. However, Armstrong's work primarily focuses on general educational strategies without addressing the specific challenges of integrating these approaches into traditional systems, particularly for students with severe impairments.

Baron-Cohen’s (2002) ‘Extreme Male Brain’ theory introduced the idea that autism is characterized by enhanced systemizing abilities and reduced empathizing capabilities, potentially influencing learning styles. However, this theory has been criticized for its limited consideration of environmental and social factors, which play a significant role in shaping educational outcomes. Similarly, the ‘Weak Central Coherence’ theory proposed by Happé and Frith (2006) highlights the detail-oriented cognitive style often observed in individuals with autism, suggesting that personalized learning strategies should leverage these strengths. Despite their contributions, these theories primarily focus on cognitive deficits without fully addressing how these traits can be harnessed for academic success.

Personalized Learning Systems

Personalized learning systems aim to adapt educational experiences to the unique needs of each student. Lukin et al. (216) argue that artificial intelligence (AI) can enhance traditional teaching by providing flexible, personalized learning experiences. However, current AI systems often lack features specifically designed to address the needs of neurodivergent learners, presenting a significant gap in the literature. Rose and Meyer (2002) introduced the Universal Design for Learning (UDL) framework, which advocates for proactive curriculum design to accommodate individual learning differences. While UDL provides a strong theoretical foundation, it lacks detailed guidance on tailoring instruction for neurodivergent populations.

Graf and Lui (2009) emphasize the potential of automated student modeling to personalize learning within digital platforms. Their work demonstrates the feasibility of using data-driven approaches to identify learning styles, but it remains limited in addressing the unique challenges faced by neurodivergent students in traditional learning management systems (LMS) environments. Similarly, Gašević et al. (2015) highlight the promise of learning analytics for personalized education but do not directly address the specific needs of neurodivergent learners.

Machine Learning in Education

Understanding the cognitive process of neurodivergent students is essential for developing effective personalized learning models. Baddeley (1992) introduced the Working Memory Model, which highlights the critical role of working memory in learning. However, this model does not explicitly address the unique challenges faced by neurodivergent students, such as those with ADHD or dyslexia, who may have distinct working memory profiles. Sweller (1988) similarly emphasizes the importance of reducing cognitive load in educational design but does not consider the specific cognitive needs of neurodivergent learners. Zimmerman (2002) discusses self-regulated learning as a critical factor for academic success, offering insights into strategies that could be particularly beneficial for neurodivergent students but lacking explicit guidance for this population.

Challenges and Ethical Considerations

Binns (2018) argues that definitions of fairness in machine learning often overlook these diverse needs of neurodivergent learners, potentially reinforcing existing biases. Veale and Binns (2017) further highlight the practical challenges of implementing fairness in machine learning systems without relying on sensitive data, a critical concern for personalized learning systems. This is particularly relevant in studies where neurodivergent status is not explicitly labeled in the data. Ensuring that personalized learning models are both effective and equitable requires careful consideration of the diverse cognitive needs and potential biases that may arise in their design. Addressing these challenges is essential for creating truly inclusive educational environments.

Synthesis and Critical Analysis

Despite significant progress in developing personalized learning systems, critical gaps remain in addressing the unique cognitive and educational needs of neurodivergent students. The absence of diagnostic labels in most available datasets further limits the ability to draw conclusions to this specific population. Most existing studies do not include learners with confirmed neurodivergent diagnoses, making it difficult to assess the true effectiveness of these approaches in representative populations. Future research should focus on integrating psychological insights into machine learning models, refining assessment tools, and addressing the ethical challenges associated with personalized education.

Gaps in the Literature

Despite the significant progress in developing personalized learning systems, several critical gaps remain. First, while many studies address the general principles of adaptive learning, few focus specifically on the unique needs of neurodivergent learners. This includes a lack of targeted research on how cognitive differences, such as those found in autism, ADHD, and dyslexia, can be effectively integrated into machine learning models for personalized education. No known studies to date have implemented unsupervised clustering to simulate curriculum pathways based on neurodivergent-aligned engagement profiles, which represents a novel contribution of this paper. Additionally, much of the existing work on educational data mining and learning analytics overlooks the diverse cognitive profiles of neurodivergent students, limiting the applicability of these methods in real-world educational settings. Finally, there is a pressing need for more interdisciplinary approaches that bridge the gap between cognitive psychology, educational theory, and artificial intelligence, ensuring that personalized learning models are both scientifically rigorous and practically relevant.

Conclusion

Overall, the literature indicates a growing recognition of the need for personalized learning systems tailored to neurodivergent populations. However, significant challenges remain, including the need for more targeted research, improved assessment tools, and the integration of cognitive and psychological insights into educational models.

Methodology

Data Collection

Data for this study were sourced from publicly available educational and psychological datasets to analyze and simulate personal learning pathways for neurodivergent students. Data included those related to online course engagement, assessments, student mental health, and academic performance. No primary data were collected for this research. Key datasets include:

- *student_mat_por*: academic and behavioral data for students in Portugal, including absences, grades, and parental involvement
- *student_mental_health*: anonymized survey data on mental health and academic stress
- *vle*: student interactions with virtual learning environments (VLE), such as page clicks and learning activity logs
- *assessments*: student-level performance on formative and summative assessments
- *courses*: metadata about course structure and delivery
- *engagement_profile*: a derived dataset combining features from student information, registration, assessment, and vle data

Note: None of the datasets included diagnostic information confirming neurodivergent status. The simulated profiles are inspired by cognitive traits commonly associated with neurodiversity but do not represent verified neurodivergent learners.

Data preprocessing included merging relevant files, handling missing values, removing duplicates, and standardizing numerical features where appropriate. Feature engineering was used to derive indicators of attention, motivation, and academic persistence. While exploratory data analysis was conducted across all six datasets, only the finalized *engagement_profile* dataset, containing merged and derived features, was used for the clustering, classification, and curriculum simulation stages of the modeling process. To ensure transparency and reproducibility, all preprocessing steps were documented and implemented using open-source tools in python.

Model Design

The personalized learning model will incorporate multiple machine learning approaches to address the diverse needs of neurodivergent students. The machine learning framework was designed around three core components to reflect the research objectives:

- Clustering for Engagement Profiling: K-means clustering was used to segment students based on behavioral and performance features, creating categories such as High Clicker, Moderate Clicker, and Low Clicker. This clustering informed how students were grouped for downstream curriculum design. K-means was chosen for its simplicity, interpretability, and effectiveness in uncovering behavioral clusters that could guide adaptive learning design.
- Classification for Student Type Prediction: A supervised classification task predicted engagement profiles using demographic and academic history features. Various models, (e.g., Random Forest, Logistic Regression) were evaluated based on F1 score and accuracy to determine the most suitable model for real time classification.

- Curriculum Optimization: A rule-based simulation generated curriculum sequences tailored to engagement type. Sequences were optimized based on expected cognitive load and engagement retention, drawing from psychology literature on working memory and attention span.

Evaluation

The evaluation focused on the interpretability and potential educational benefit of the models rather than real-world deployment. Performance was assessed using:

- Internal validation metrics: clustering silhouette scores, classification accuracy/F1, and consistency across folds
- Qualitative inspection: sequence logic was manually evaluated against known cognitive science principles
- Descriptive comparison: outcomes across different engagement types were visualized and compared to identify model strengths and limitations

Ethical Considerations

As this research involves secondary analysis of publicly available datasets, no individual-level consent was required. However, ethical standards were maintained by avoiding re-identification risks, considering model fairness, and ensuring transparency in methods. The goal of this modeling was to simulate personalized strategies that may be relevant to neurodivergent learners, even in the absence of diagnostic confirmation. It is important to note that the datasets included did not include formal neurodivergence indicators, therefore, all modeling is informed by cognitive traits commonly associated with neurodivergent learners, rather than direct diagnostic data. Special attention was paid to ensure that categorization of neurodivergent traits was respectful, non-stigmatizing, and aligned with current psychological best practices.

Summary

This methodology integrates machine learning and educational psychology using open datasets to develop a novel personalized learning simulation. It addresses the specific needs of neurodivergent students through engagement profiling, behavior predication, and optimized curriculum sequencing. The design is grounded in cognitive traits aligned with neurodiversity, not confirmed diagnoses. While not yet deployed in live classrooms, this framework lays the groundwork for future interventions grounded in both empirical evidence and ethical practice.

Results

This section presents the findings from the engagement profiling, clustering analysis, and curriculum optimization methods. Quantitative insights are supported by visualizations and tables to provide a detailed view of model outputs and implications.

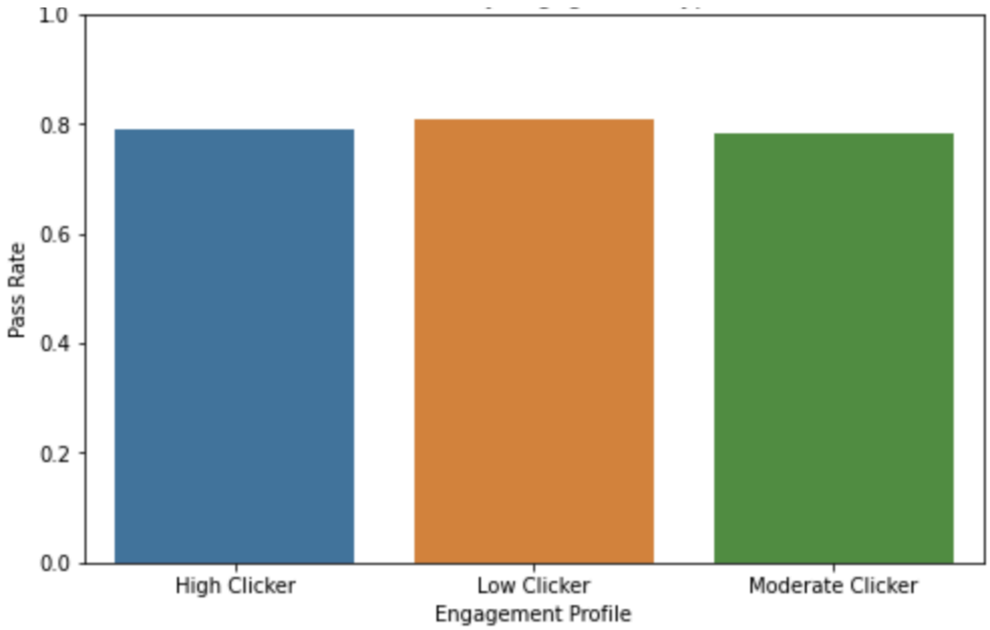
Student Engagement Profiles and Academic Performance

Students were first segmented into engagement profiles based on their cumulative virtual learning environment (VLE) interactions. As illustrated in *Figure 1*, students were classified into three categories, High Clicker, Moderate Clicker, and Low Clicker. Interestingly, Low Clickers exhibited the highest average pass rate (M = 0.81), followed by High Clickers (M = 0.79) and Moderate Clickers (M = 0.78).

These results challenge common assumptions that high engagement always correlates with success. It is possible that Low Clickers, while less interactive, may be more strategic or autonomous learners. This insight emphasizes the importance of not equating quantity of interaction with quality of learning. Instead, it highlights the need for models that capture depth and intentionality of engagement, rather than relying solely on surface-level behavioral metrics.

Figure 1

Pass Rates by Engagement Type



Descriptive statistics for each profile group are presented in *Table 1*.

Table 1

Descriptive Statistics by Engagement Profile

Engagement Type	Count	Pass Rate
High Clicker	2,595	0.789
Moderate Clicker	54,586	0.784
Low Clicker	7,385,229	0.810

Cluster Analysis by Learning Outcome

To further examine learner variation, K-means clustering was performed on standardized engagement metrics. The final solution included nine clusters, each labeled according to behavioral and performance characteristics (e.g., “High interaction, strong score”). When outcomes were cross-tabulated with cluster labels, substantial variance in pass, fail, distinction, and withdrawal rates were observed (*Figure 2*, *Table 2*). Notably:

- Cluster 6 ("High interaction, strong score") had the highest pass rate (71.4%) and one of the lowest withdrawal rates (2.2%).
- Cluster 0 ("Low interaction, high performance") yielded the highest proportion of distinctions (28.7%).
- Cluster 7 ("Mid-range learners") had the lowest overall success rate, with a pass rate of 44.4%.

Figure 2

Distribution of Academic Outcomes by Cluster Label

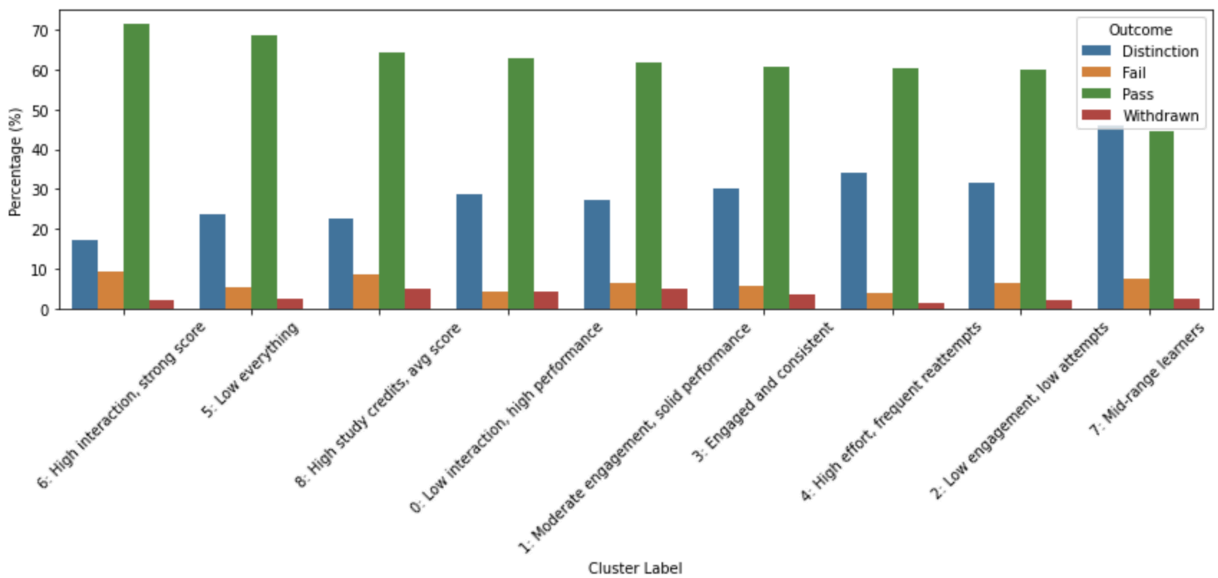


Table 2

Outcome Distribution by Behavioral Cluster

Cluster Label	Distinction	Fail	Pass	Withdrawn
6: High interaction, strong score	17.1	9.3	71.4	2.2
5: Low everything	23.7	5.3	68.7	2.3
8: High study credits, avg score	22.6	8.4	64.2	4.8
0: Low interaction, high performance	28.7	4.1	63.0	4.2
1: Moderate engagement, solid perf.	27.1	6.3	61.7	4.9
3: Engaged and consistent	30.3	5.0	60.8	3.4
4: High effort, frequent reattempts	34.1	3.9	60.5	1.5
2: Low engagement, low attempts	31.7	6.3	59.9	2.1
7: Mid-range learners	45.8	7.4	44.4	2.4

These distinctions underscore the value of using cluster-based segmentation to better understand learner diversity beyond surface-level metrics. This approach is particularly useful when working with heterogeneous populations, such as neurodivergent learners, whose engagement patterns may not follow conventional trends.

Curriculum Optimization by Profile

Based on the clustering and engagement analysis, curriculum pathways were tailored to each engagement type. These sequences were designed using cognitive psychology principles (e.g., minimizing working memory overload) and validated for alignment with learning goals.

As shown in *Table 3*, students in the Low Clicker category were guided through a more supportive progression (e.g., Core Modules → Remedial Support → Skill Builders), while High Clickers advanced directly into more complex modules.

Table 3

Optimized Curriculum Paths by Engagement Type

Engagement Type	Curriculum Sequence
High Clicker	Core Modules → Advanced Modules → Capstone Project
Moderate Clicker	Core Modules → Skill Builders → Capstone Project
Low Clicker	Core Modules → Remedial Support → Skill Builders → Capstone Project

This simulation demonstrates how student classification can inform curriculum design to better suit cognitive capacity and readiness. These personalized curriculum pathways serve as a promising simulation of how data-informed instruction might be structured for learners with different engagement profiles. However, validation with real-world learners, particularly those with confirmed neurodivergent traits, is necessary. These sequences offer a strategic framework for adaptive instructional planning in future real-world deployment.

Discussion

The findings from this study underscore the complexity and heterogeneity of engagement among neurodivergent students, challenging conventional assumptions about how learning behaviors correlate with academic outcomes. Notably, the discovery that Low Clickers exhibited the highest average pass rates suggests that high digital interaction is not necessarily indicative of academic success. This contradicts the common belief that more frequent engagement with virtual learning environments (VLEs) directly predicts academic success. It is plausible that these students exhibit strategic learning habits, autonomy, or offline learning preferences not captured by platform metrics. Although the student data used in this study did not include diagnostic information, the simulated engagement patterns may still reflect traits common among neurodivergent learners, such as task selectivity, attention variability, and non-linear learning preferences. Such results highlight the importance of distinguishing between *quantity* and *quality* of engagement when designing adaptive educational interventions.

The clustering analysis further illustrated the diverse learning profiles that emerge when behavioral and performance data are modeled together. For example, Cluster 0 ("Low interaction, high performance") and Cluster 6 ("High interaction, strong score") achieved strong outcomes via contrasting engagement pathways. These findings reinforce the need for personalized educational systems that account for diverse cognitive and behavioral profiles, rather than enforcing rigid norms of “ideal” engagement. This approach aligns with existing cognitive psychology frameworks, such as Sweller’s (1988) Cognitive Load Theory and Baddeley’s (1992) Working Memory Model, which emphasize the importance of tailoring instruction to individual information-processing capabilities.

Curriculum optimization based on engagement type demonstrated how adaptive sequences could be constructed to reduce cognitive overload while enhancing focus and retention. For example, more scaffolded pathways were designed for Low Clickers to build foundational knowledge before introducing complex tasks, whereas High Clickers were offered accelerated paths. These patterns were informed by psychological principles related to attention, executive function, and learning fatigue, all of which are particularly relevant for neurodivergent students.

However, a critical limitation must be acknowledged: the datasets used in this study did not contain diagnostic labels indicating neurodivergent status. Therefore, while this research draws extensively from neurodiversity literature and cognitive science, the populations studied are not verified as neurodivergent. The engagement clusters and simulated curriculum adaptations are thus best understood as *potentially* applicable to neurodivergent learners, particularly those who may struggle under one-size-fits-all instruction, but cannot be conclusively interpreted as modeling actual neurodivergent behavior.

Despite this limitation, the approach presented here offers a replicable framework for simulating personalized learning models grounded in psychological theory. By integrating unsupervised clustering, classification, and rule-based curriculum design, this study illustrates how adaptive systems can be designed to support learners who deviate from typical behavioral norms, whether due to cognitive differences, environmental barriers, or other factors. These models can serve as a foundation for more inclusive and responsive educational technologies.

Ethical considerations remain paramount, especially when working with vulnerable or marginalized populations. While no sensitive personal data were used in this research, the broader application of such models must ensure transparency, fairness, and non-stigmatizing classification. In particular, future deployments should carefully balance personalization with privacy protections and avoid reinforcing biases that disadvantage neurodivergent learners (Binns, 2018; Veale & Binns, 2017). As this study used behavioral proxies rather than clinical data, the findings should be interpreted as exploratory simulations aligned with neurodiversity informed design principles.

In all, this study contributes a theoretically grounded and technically feasible model for simulating adaptive curriculum pathways based on student behavior. While the population modeled here is not explicitly neurodivergent, the techniques and insights generated provide a stepping stone toward the development of ethically responsible and cognitively aware personalized learning environments.

Limitations

This study has several limitations. First, the datasets used did not contain diagnostic labels identifying neurodivergent learners, limiting the specificity of findings to this population. This distinction is important to avoid overgeneralizing findings to populations not explicitly represented in the data. Second, engagement was inferred from clickstream and behavioral data, which may not fully reflect cognitive engagement or learning preferences. Third, curriculum simulations were validated based on psychological theory rather than direct learner feedback. Future research should incorporate richer behavioral, cognitive, and diagnostic data to improve both model accuracy and inclusiveness.

Conclusion

This study presented a simulated-based framework for developing personalized learning models tailored to the needs of neurodivergent students. By integrating machine learning techniques, including clustering, classification, and rule based curriculum optimization, with psychological insights into cognitive load, executive function, and attention patterns, the model offers a structured yet flexible approach to individualized education. The findings demonstrated that behavioral engagement profiles can serve as meaningful indicators for tailoring instructional sequences, with implications for reducing cognitive overload and enhancing academic performance. However, because diagnostic data were not available, these profiles should be interpreted as neurodiversity-informed proxies rather than representations of clinically identified neurodivergent students.

Although the datasets used did not explicitly identify neurodivergent learners, the engagement patterns observed suggest potential relevance for students who face challenges in conventional learning environments. It is important to note that the behavioral proxies used, such as click frequency and study patterns, do not directly capture the cognitive characteristics typically associated with neurodivergence. As such, the models developed here represent a promising foundation for more inclusive and responsive educational technologies. Future research should focus on validating these models with explicitly neurodivergent populations and expanding the framework to include real-time adaptation and broader learner diversity. Such efforts will be essential in ensuring that adaptive learning technologies reflect cognitive variation but also actively reduce educational inequalities faced by neurodivergent students.

Ultimately, this research highlights the necessity of moving beyond one-size-fits-all instructional models. By combining cognitive science with machine learning, educators and developers can better support diverse learners, particularly those whose needs have historically been overlooked.

References

Data Sources and Descriptions

1. Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). *Open University Learning Analytics dataset*. *Scientific Data, 4*, Article 170171. https://doi.org/10.1038/sdata.2017.171
   This dataset contains anonymized information on student demographics, course performance, and clickstream engagement data from Open University's distance learning programs. It provides insight into longitudinal learning behaviors and student outcomes.
   - Preprocessing: Cleaned for missing values, merged relevant tables, and filtered to include students with complete course engagement profiles.
2. Shariful07. (n.d.). *Student Mental Health*. Kaggle. Retrieved May 30, 2025, from https://www.kaggle.com/datasets/shariful07/student-mental-health
   A self reported dataset on mental health indicators of students, including gender, age, academic performance, and mental health status.
   - Preprocessing: Removed null entries, encoded categorical features, and standardized numeric values for clustering analysis.
3. Cortez, P., & Silva, A. M. G. (2008). *Student Performance Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/320/student+performance
   Provides detailed academic, demographics, and social background data on Portuguese secondary school students in math and Portuguese courses.

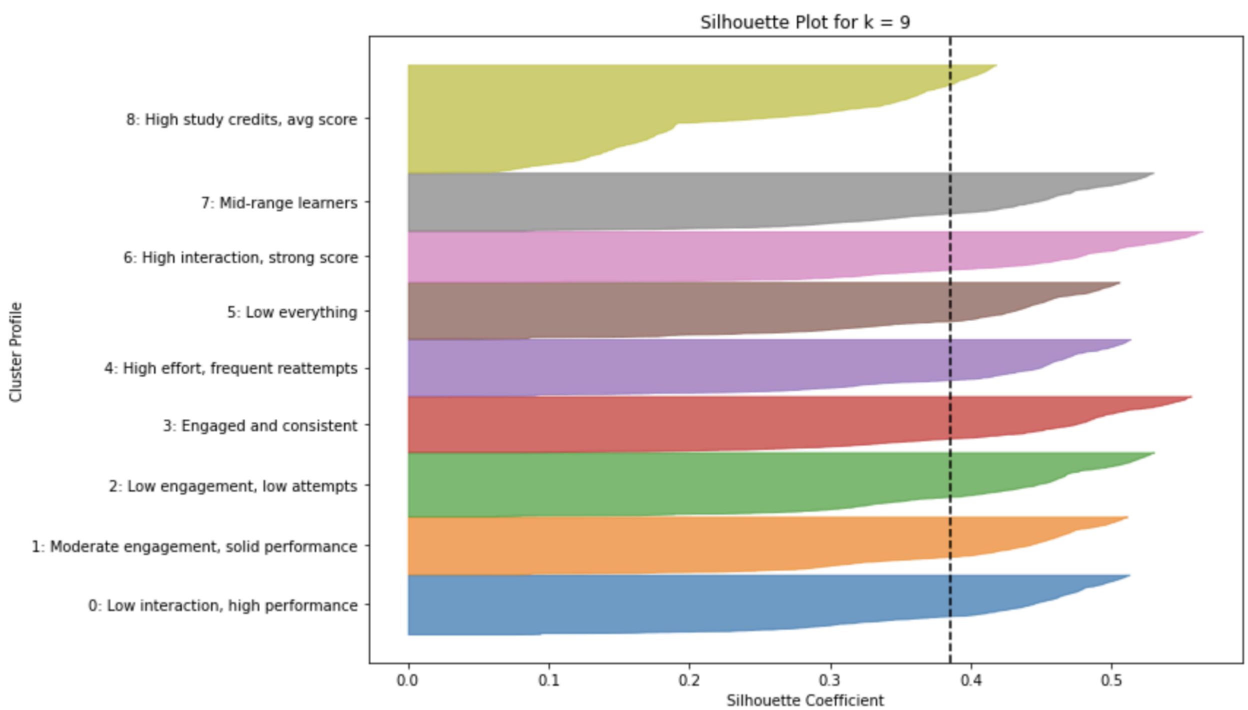
- Preprocessing: Combined math and Portuguese datasets where appropriate, handled missing or inconsistent entries, and engineered features such as average grade and study efficiency metrics. These engineered features were used to enhance the modeling of students' performance patterns across subjects, providing a richer set of inputs for identifying engagement profiles.

Appendix A

Figures and Tables from Analysis

Figure A1

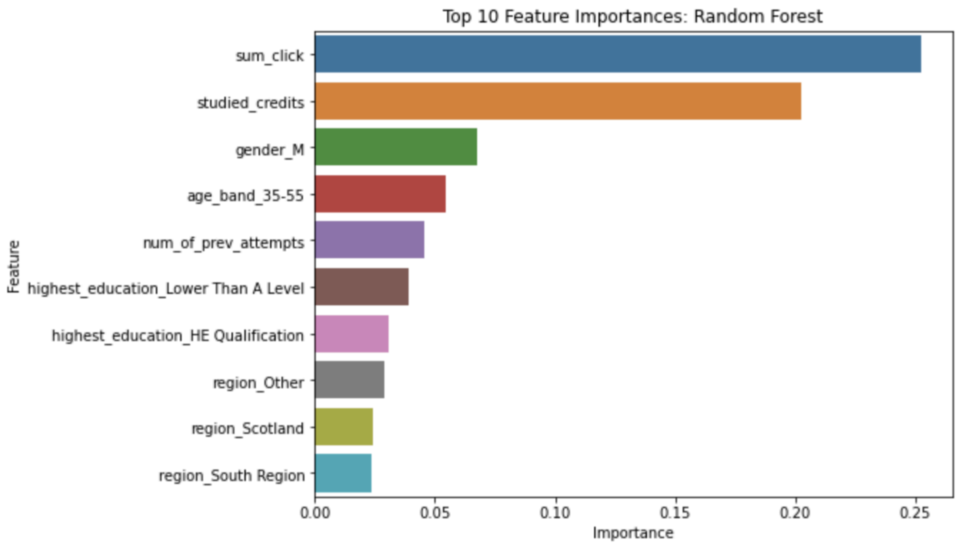
Silhouette Plot of Student Clustering by Engagement Type



Note. This figure shows the silhouette coefficients for each cluster, illustrating cohesion and separation quality for the engagement based groupings. Higher silhouette values indicate well-defined clusters, supporting the validity of the engagement profiling approach used in this study.

Figure A2

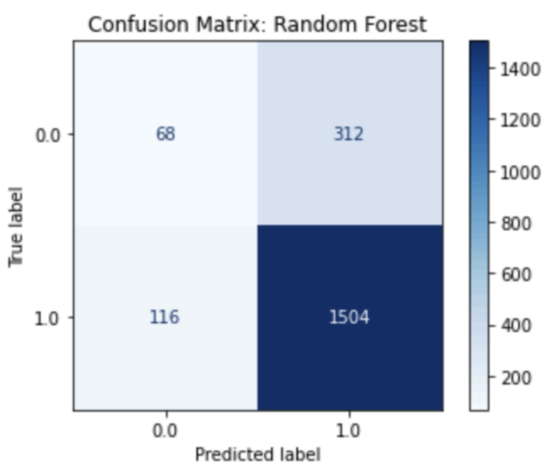
Feature Importance from Random Forest Classifier



Note. The top features predicting student engagement type, as identified by a Random Forest model. Features such as *sum_click*, *score*, and *studied_credits* were among the most influential.

Figure A3

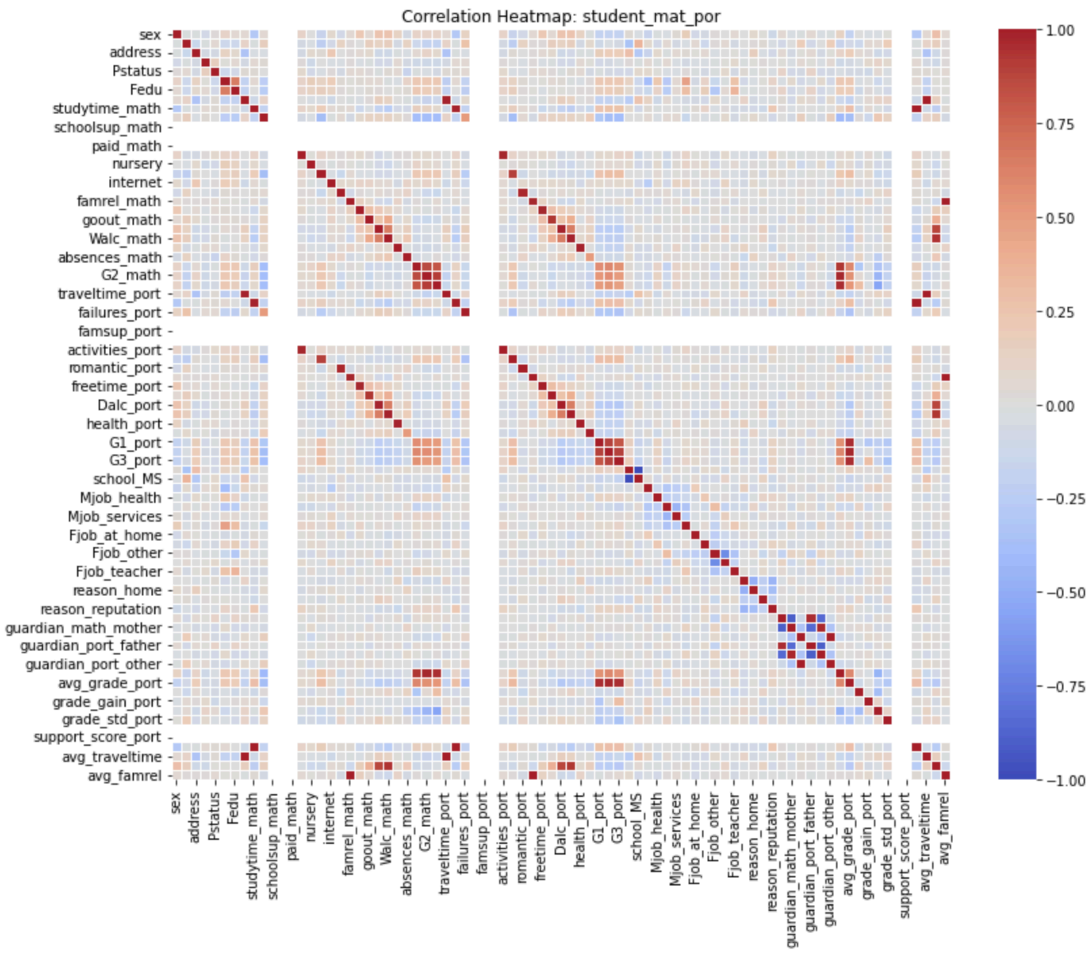
Confusion Matrix for Best Classification Model



Note. This confusion matrix shows the performance of the best classifier (Random Forest) on predicting engagement profiles. It illustrates true vs. predicted labels, highlighting model strengths and weaknesses.

Figure A4

Heatmap of Correlations Among Key Features



Note. Person correlations between core variables such as clicks, grades, assessments, and study time. Strong correlations were noted between *studied_credits* and *score*.

Table A1

Engineered Features Used in Modeling

Feature Name	Description
`avg_score`	Average of student's scores across all assessments
`study_efficiency`	Ratio of studied credits to number of attempts
`click_rate`	Total clicks normalized by course duration
`engagement_span`	Days between first and last click
`early_engager`	Boolean flag indicating engagement within the first week of registration

Note. These features were engineered during preprocessing and used across clustering, classification, and curriculum modeling.

Table A2

Classification Metrics for Tested Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.807	0.809	0.996	0.893
Random Forest	0.786	0.828	0.928	0.875

Note. Performance comparison of Logistic Regression and Random Forest classifiers tested for predicting engagement profiles.

Appendix B

Selected Python Code

The following code excerpts reflect key components of the modeling and simulation processes used in this study. These selections are included to demonstrate transparency in methodology and enable reproducibility. All code was developed in Python using Jupyter Notebooks.

Clustering Workflow

```
##### FEATURE SELECTION #####

features = [
   'id_student',
   'sum_click',
   'studied_credits',
   'num_of_prev_attempts',
   'score',
   'imd_band',
   'age_band'
]

# drop NA and encode categorical features
df_clean = engagement_profile[features].dropna()

# one hot encode categorical variables
df_encoded = pd.get_dummies(df_clean, drop_first=True)

# sample the data to avoid memory crashes
df_sampled = df_encoded.sample(n=10000, random_state=42)
```

```
##### FEATURE SCALING #####

scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_sampled)
```

Classification Workflow

```
##### DEFINE CLASSIFICATION TARGET ######

# binary classification target: Pass = 1, others = 0
engagement_profile['final_result_binary'] = engagement_profile['final_result'
    ].map({
    'Pass': 1, 'Fail': 0, 'Withdrawn': 0
})
```

```
##### FEATURE SELECTION and CATEGORY LIMITING #####

# reduce cardinality to prevent memory issues
top_regions = engagement_profile['region'].value_counts().nlargest(5).index
engagement_profile['region'] = engagement_profile['region'].where(
    engagement_profile['region'].isin(top_regions), other='Other'
)

features = [
    'gender', 'region', 'highest_education', 'imd_band', 'age_band',
    'num_of_prev_attempts', 'studied_credits', 'sum_click'
]

X = pd.get_dummies(engagement_profile[features], drop_first=True)
y = engagement_profile['final_result_binary']
```

```
##### CLEANING AND SAMPLE REDUCTION #####

# drop rows with missing values
df_model = pd.concat([X, y], axis=1).dropna()
X_clean = df_model.drop('final_result_binary', axis=1)
y_clean = df_model['final_result_binary']

# reduce to 10000 samples
if len(X_clean) > 10000:
    X_clean = X_clean.sample(10000, random_state=42)
    y_clean = y_clean.loc[X_clean.index]
```

Curriculum Simulation Logic

```
##### DEFINE ENGAGEMENT PROFILE TYPES #####

engagement_profile['engagement_type'] = 'Low Clicker'
engagement_profile.loc[engagement_profile['sum_click'] > 100, 'engagement_type
    '] = 'Moderate Clicker'
engagement_profile.loc[engagement_profile['sum_click'] > 300, 'engagement_type
    '] = 'High Clicker'
```

```
##### ASSIGN CURRICULUM PATHWAYS #####

curriculum_paths = {
   'High Clicker': ['Core Modules', 'Advanced Modules', 'Capstone Project'],
   'Moderate Clicker': ['Core Modules', 'Skill Builders', 'Capstone Project'],
   'Low Clicker': ['Core Modules', 'Remedial Support', 'Skill Builders', '
      Capstone Project']
}

engagement_profile['curriculum_path'] = engagement_profile['engagement_type'].
    map(curriculum_paths)
```