SAAT 2.0: A Vectorial and Unsupervised Learning Framework for Academic Analytics and Early Warning Systems

Author

Johnny Villalobos Murillo

jvillalobos@una.ac.cr

Abstract

Early Warning Systems (EWS) in higher education traditionally rely on predefined indicators and supervised predictive models. Although effective in specific scenarios, these approaches often fail to discover emerging behavioral patterns and hidden academic trajectories.

This paper proposes SAAT 2.0 (Academic Analytics and Early Warning System 2.0), a mathematical framework that combines hierarchical vector representations, temporal analytics, unsupervised learning, similarity analysis, and academic trajectory mining.

The model extends traditional academic analytics by introducing two additional spaces: an Academic Embedding Space and a Pattern Discovery Space. Through these extensions, students are represented not only by observed variables, indicators, and risks, but also by latent behavioral structures discovered directly from institutional data.

The proposed framework enables clustering, similarity search, trajectory analysis, emerging risk detection, and explainable academic intelligence.

Keywords—Learning Analytics, Early Warning Systems, Academic Embeddings, Unsupervised Learning, Educational Data Mining, Student Trajectories.

I. INTRODUCTION

Higher education institutions continuously seek mechanisms capable of identifying students at academic risk before adverse outcomes occur.

Traditional Early Warning Systems generally focus on predefined indicators such as grades, attendance, or enrollment patterns. These approaches often require expert-designed risk models and may overlook latent structures present in institutional data.

SAAT 2.0 addresses this limitation by introducing a multi-layer vectorial architecture capable of discovering unknown academic patterns through unsupervised learning.

The central hypothesis is:

"Higher-order academic behaviors emerge naturally from student trajectories and can be identified through vector-space transformations and clustering techniques without requiring predefined labels."

II. THEORETICAL FOUNDATIONS

SAAT 2.0 integrates concepts from:

A. Learning Analytics

B. Educational Data Mining

C. Vector Space Models

D. Temporal Data Analysis

E. Unsupervised Machine Learning

F. Graph Theory

G. Representation Learning

III. SAAT 1.0 FOUNDATION

The original SAAT architecture defines three hierarchical spaces:

Variables

$V \in \mathbb{R}^{16}$

Indices

$I \in \mathbb{R}^{8}$

Risks

$R \in \mathbb{R}^{4}$

connected through two transformation matrices:

$I = A \cdot V$

$R = B \cdot I$

thus:

$R = B \cdot A \cdot V$

This architecture produces explainable risk estimations while preserving full traceability.

IV. SAAT 2.0 EXTENDED ARCHITECTURE

The proposed extension introduces two new analytical spaces.

$V \rightarrow I \rightarrow R \rightarrow E \rightarrow C$

where:

V = Observed Variables

I = Academic Indicators

R = Academic Risks

E = Academic Embeddings

C = Discovered Clusters

The resulting architecture becomes:

$\mathbb{R}^{16} \rightarrow \mathbb{R}^{8} \rightarrow \mathbb{R}^{4} \rightarrow \mathbb{R}^{d} \rightarrow \mathbb{C}$

where d is the embedding dimension.

V. ACADEMIC EMBEDDING SPACE

Each student is represented by a vector:

$E_e(t)$

defined as:

$E_e(t) = \Phi(V_e(t), I_e(t), R_e(t))$

where Φ denotes an embedding transformation.

The embedding space satisfies:

$E_e(t) \in \mathbb{R}^{d}$

with d chosen according to analytical requirements.

The embedding captures latent academic characteristics that cannot be observed directly through individual indicators.

VI. STUDENT SIMILARITY MODEL

Given two students:

E_a

and

E_β

their similarity is defined by:

$sim(a,b)=\cos(E_a,E_\beta)$

or alternatively:

$d(a,b)=||E_a-E_\beta||$

This enables:

• Student recommendation.

• Academic twin discovery.

• Similar-case retrieval.

• Historical intervention matching.

VII. TEMPORAL TRAJECTORY MODEL

Each student trajectory is defined as:

$T_e=\{S_e(t_1),S_e(t_2),\ldots,S_e(t_n)\}$

where:

$S_e(t)=$
$(V_e(t),I_e(t),R_e(t),E_e(t))$

The trajectory forms a path through the academic state space.

VIII. TRAJECTORY DISTANCE

Two trajectories are compared using:

$DTW(Ta,Tb)$

Dynamic Time Warping

or

Hausdorff Distance

allowing the identification of students with similar academic evolution patterns.

IX. UNSUPERVISED PATTERN DISCOVERY

A clustering function:

$C=f(E)$

partitions the embedding space into groups.

Possible algorithms include:

- K-Means
- DBSCAN
- HDBSCAN
- Spectral Clustering
- Hierarchical Clustering

Each cluster represents a latent academic profile.

X. EMERGENT RISK DISCOVERY

Traditional systems define risks a priori.

SAAT 2.0 introduces:

R^*

where:

$R^* = g(C)$

representing risks discovered directly from data.

Examples include:

- Progressive disengagement.
- Hidden academic burnout.
- Enrollment instability.
- Financial vulnerability patterns.

These risks may not be visible within predefined institutional indicators.

XI. ACADEMIC GRAPH MODEL

A graph is defined as:

$G=(N,E)$

where:

N = students

E = similarity relationships

An edge exists if:

$sim(a,b) > \theta$

The graph enables:

• Community detection.

• Centrality analysis.

• Risk propagation studies.

• Academic influence modeling.

XII. VECTOR DATABASE LAYER

SAAT 2.0 introduces a vector database layer.

For every student:

Embedding ID

Student ID

Timestamp

Embedding Vector

This structure enables:

Nearest Neighbor Search

Top-K Similar Students

Historical Case Retrieval

Semantic Academic Search

The architecture becomes compatible with modern AI systems.

XIII. MATHEMATICAL DEFINITION OF SAAT 2.0

$SAAT_2 =$

(V,A,I,B,R,T,E,C,G)

where:

V = Variable Space

A = Indicator Transformation Matrix

I = Indicator Space

B = Risk Transformation Matrix

R = Risk Space

T = Temporal Space

E = Embedding Space

C = Cluster Space

G = Academic Similarity Graph

XIV. FUNDAMENTAL EQUATION

$R(t)=B\cdot A\cdot V(t)$

$E(t)=\Phi(V(t),I(t),R(t))$

$C(t)=f(E(t))$

$G(t)=h(E(t))$

Thus:

$SAAT_2=$
$(V\rightarrow I\rightarrow R\rightarrow E\rightarrow C\rightarrow G)$

XV. SCIENTIFIC CONTRIBUTIONS

The proposed framework contributes:

1. A hierarchical vector representation of academic behavior.
2. Temporal trajectory modeling.
3. Student similarity analytics.
4. Academic embeddings.
5. Unsupervised profile discovery.
6. Emergent risk identification.
7. Graph-based academic intelligence.
8. Explainable educational analytics.

XVI. CONCLUSIONS

SAAT 2.0 transforms traditional Early Warning Systems into a comprehensive Academic Intelligence Platform.

Rather than limiting analysis to predefined risks, the framework discovers hidden structures, latent profiles, and emerging trajectories directly from institutional data.

The resulting architecture provides a foundation for next-generation educational analytics, combining explainability, predictive capabilities, similarity reasoning, and unsupervised discovery within a unified mathematical framework.