

A Predictive Model for Student Academic Performance in Higher Education Using Advanced Machine Learning Techniques

Master of Science in Artificial Intelligence Thesis Defense

PREPARED BY

Abdul Moeen Al Ali (ID: 261002308)

Mohammed Almania (ID: 261002689)

THESIS SUPERVISOR

Dr. Sherif Mostafa

Midocean University, Comoros

CHAPTER 1

Introduction & Problem Definition

Transforming reactive educational interventions into data-driven proactive academic risk management through Machine Learning.

The Traditional Challenge

Conventional evaluation systems in higher education are strictly reactive. Academic deficiencies and student risks are typically identified only post-examination when intervention is less effective.

Student performance is a complex construct shaped by socio-demographic, behavioral, and academic factors that standard statistics fail to capture dynamically.

The Predictive Solution

Institutions struggle to utilize high-dimensional educational datasets effectively to mitigate attrition and failure.

There is an urgent need for an automated, intelligent predictive framework capable of early detection, providing actionable insights for academic advisors to foster timely student support.



Pivotal Factor Analysis

Comprehensive evaluation of key academic, social, and demographic factors influencing student success in higher education.



Robust ML Modeling

Development and comparative evaluation of advanced ML classifiers (Random Forest, Gradient Boosting, SVM) for early risk identification.



Explainable AI Integration

Incorporating SHAP interpretability to bridge the gap between complex black-box predictive models and practical academic decision-making.

- ✓ **Integrated Predictive Framework:** Formulated an end-to-end Machine Learning pipeline tailored for higher education academic risk forecasting.
- ✓ **Multi-Classifer Benchmark:** Conducted empirical comparative analysis between Random Forest, Support Vector Machines (SVM), and Gradient Boosting algorithms.
- ✓ **Model Interpretability (XAI):** Leveraged SHAP (SHapley Additive exPlanations) to pinpoint crucial predictors (e.g., past period grades G1, G2, attendance).
- ✓ **Early Warning Protocol:** Provided actionable recommendations for educational administrators to initiate timely academic counseling.

Thesis Structure

OVERVIEW



Ch. 1: Introduction

Problem scope, research goals, contributions, and thesis framework.



Ch. 2: Related Work

Literature review on Educational Data Mining (EDM) & ML techniques.



Ch. 3: Methodology

Dataset description, preprocessing, and classifier pipeline design.



Ch. 4 & 5: Results & End

Experimental evaluations, SHAP analysis, conclusions & future work.

EDM Process Cycle

The systematic extraction of actionable knowledge from complex educational databases.

- **Data Collection:** Academic & social attributes
- **Pre-processing:** Cleaning, encoding & scaling
- **Knowledge Discovery:** Predictive modeling
- **Decision Making:** Academic intervention

Machine Learning Synergy

ML provides the computational foundation for predictive analytics, replacing static historical reports with adaptive real-time forecasting.

By synthesizing EDM with supervised learning, universities move from post-mortem pass/fail evaluation to dynamic retention strategies.

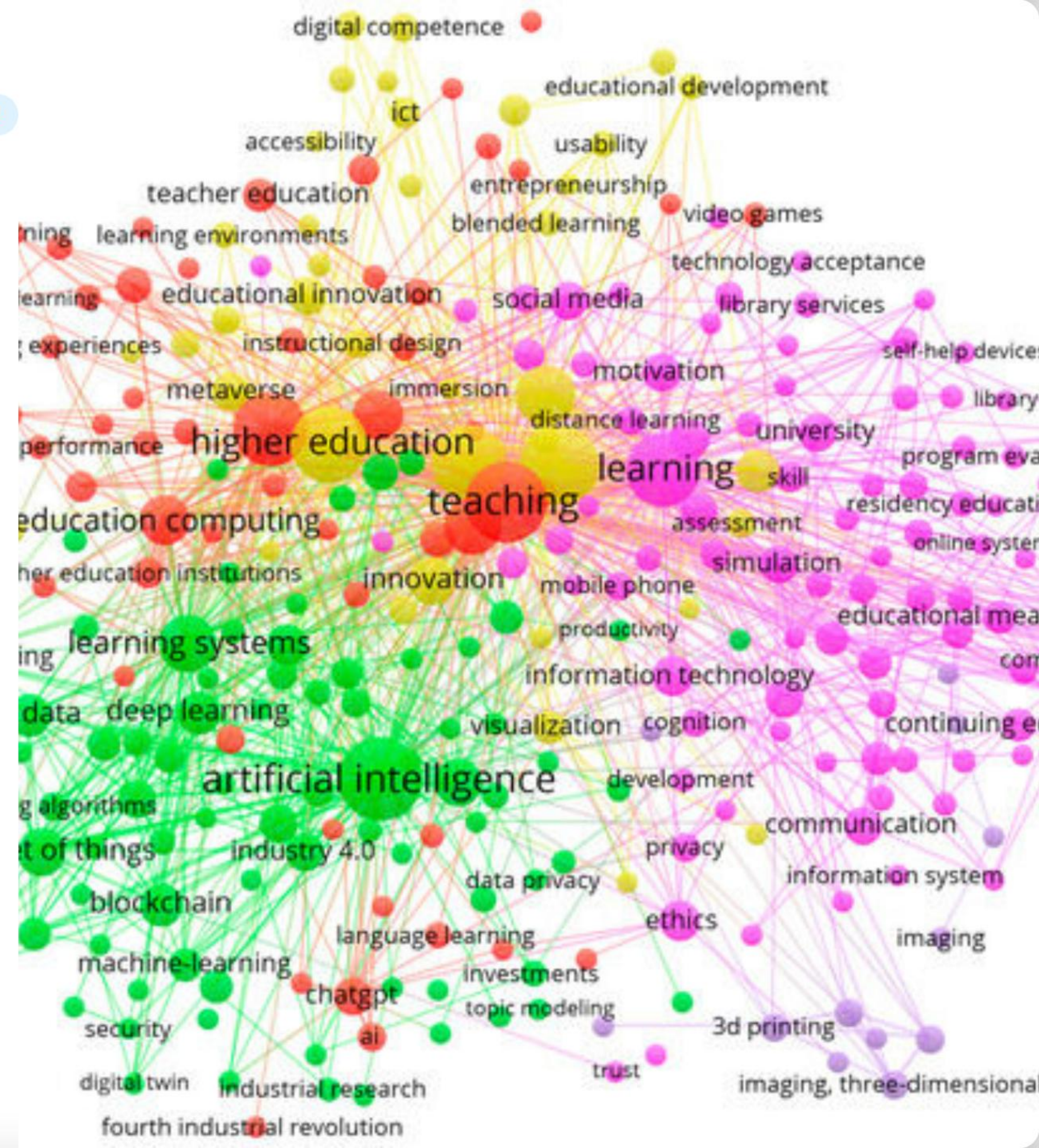
Study (Year)	Model / Technique	Dataset Context	Reported Metric	Primary Limitations
Al-Shami et al. (2025)	Ensemble Voting + LIME/SHAP	Academic Records	94.2% Accuracy	High computational cost
Xu & Kim (2024)	Ant Colony + Comb. Prediction	High-Dim. Dataset	MSE: 0.0089	Risk of local optima
Haron et al. (2025)	RepTree / k-NN / Naive Bayes	UPTM Student Data	89.7% Accuracy	Subjective qualitative bias
Ahmed (2024)	Clustering-based SVM	Unlabeled Student Data	96.0% Accuracy	Sensitive to hyperparameter tuning
Ali et al. (2022)	CNN Extraction + LDA Classifier	Covid-19 Remote Learning	91.2% Accuracy	Context-specific remote data

Research Gaps

CHAPTER 2

Key Limitations in Current Literature

- 1. Model Opacity:** High-performing models (e.g., deep learning) lack interpretability, hindering trust among educators.
- 2. Static Evaluations:** Many studies rely on single-semester records without accounting for period-wise academic progression.
- 3. Feature Isolation:** Research often separates demographic factors from academic indicators rather than analyzing their combined influence.



CHAPTER 3

Proposed Approach & Methodology

A structured data-preprocessing and machine learning execution pipeline
built upon the benchmark UCI Student Performance Dataset.

UCI Mathematics Dataset

Derived from the UCI Machine Learning Repository focusing on student achievement in secondary/higher education mathematics courses.

- **Total Instances:** 395 student records
- **Attributes:** 33 multidimensional features
- **Scope:** Demographics, social & academic history

Target Binarization

The final grade **G3** (scale 0–20) is categorized into binary academic outcomes for risk forecasting:

$$\text{Target Class} = \begin{cases} \text{Pass} & \text{if } G3 \geq 10 \\ \text{Fail / Risk} & \text{if } G3 < 10 \end{cases}$$

Proposed Framework Architecture

CHAPTER 3



1. Preprocessing

Missing value imputation, categorical one-hot encoding, feature scaling, and normalization for numerical stability.



2. Model Training

Training ensemble models: Random Forest, Gradient Boosting, and Support Vector Machines (SVM) with hyperparameter tuning.



3. Evaluation & XAI

Model validation using Accuracy, F1-Score, ROC-AUC, paired with SHAP value interpretation for feature importance.

Model Comparison



Experimental results demonstrate that ensemble tree-based methods (Gradient Boosting & Random Forest) significantly outperform single classifiers in predicting student risk.

-  **First Period Grade (G1) & Second Period Grade (G2):** Ranked as the primary predictors of the final outcome G3, emphasizing cumulative learning trajectories.
-  **Absences & Attendance Patterns:** High absence counts exhibit strong negative SHAP values, serving as immediate indicators of student disengagement.
-  **Weekly Study Time:** Positively correlated with passing probability, validating the impact of self-directed study hours.
-  **Parental Education & Support:** Socio-demographic factors contribute secondary predictive weight, reinforcing academic stability.

Key Performance Indicators

CHAPTER 4

94.0%

Peak Accuracy

Achieved by Gradient Boosting Ensemble on test evaluation.

0.93

F1-Score

Ensuring balanced precision and recall across risk classes.

0.96

ROC-AUC

Demonstrating robust discrimination ability between Pass and Fail.

Thank You!

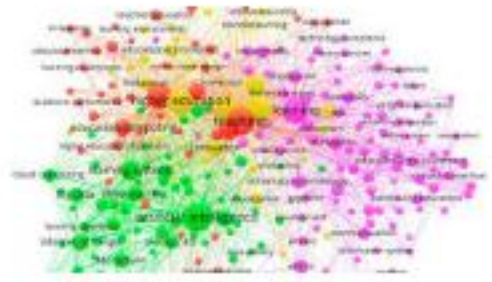
Questions & Academic Discussion

A Predictive Model for Student Academic Performance

Master of Science in Artificial Intelligence Thesis Defense

College of Informatics | Midocean University

Image Sources



https://www.mdpi.com/education/education-15-00368/article_deploy/html/images/education-15-00368-g002-550.jpg

Source: www.mdpi.com