



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

Prepared By

Name: AbdulMoien Al Ali

ID No.: 261002308

Name: Mohammed Almania

ID No.: 261002689

Submitted in fulfillment of the requirements for the degree Master of Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

Supervisor Name: Dr. Sherif Mostafa

Collage of Informatics, Midocean University, Comoros

Email:.....

Acknowledgements

Special thanks to Dr. Sherif Mostafa

Abstract

List of Publications

Table of Contents

List of Abbreviations	vi
List of Figures.....	vi
List of Tables	vi

Chapter 1: Introduction and

Overview.....	1
1.2 Problem Definition	2
1.3 Thesis Objectives	3
1.4 Thesis Contribution	4
1.5 Thesis Organization.....	5

Chapter 2: Background and Related Work

.....	31
-------	----

Chapter 3: Proposed Approach

4.1 Introduction	78
------------------------	----

Chapter 4: Experimental Results and Discussion

Chapter 5: Conclusion and Future Work	127
7.1 Important Features	127
7.3 Future Work	129
 References	 130

List of Abbreviations

- **Acc.:** Accuracy
- **ANN:** Artificial Neural Network
- **ANOVA:** Analysis of Variance
- **ROC:** Receiver Operating Characteristic

List of Figures

2.1 Educational Data Mining Process	7
2.2 Applications of Educational Data Mining	8
2.3 Distribution of Papers that used different attributes for PSP problem	19
2.4 Student Performance Data Set	21
2.5 Educational Process Mining (EPM): A Learning Analytics Data Set ...	23

List of Tables

2.1 Student Performance Data Set Description	21
2.2 Students' Academic Performance Dataset Description	23

Chapter 1 Introduction

1.1 Overview

Higher education is evolving through AI and machine learning, enabling data-driven decisions to improve student outcomes. Academic performance, influenced by academic, socio-demographic, and behavioral factors, is a key success indicator. Traditional retrospective analysis is often insufficient for timely interventions. This research develops a predictive model using advanced machine learning to identify at-risk students early, bridging the gap between data science and educational practice.

1.2 Problem Definition

Institutions struggle to use educational data effectively to prevent failure and dropout. Current evaluation systems are reactive, identifying risks only after exams when it is often too late. Student performance is a complex interplay of attendance, engagement, and socio-economic status. There is an urgent need for automated systems to analyze multidimensional data and provide reliable, actionable predictions for educators.

1.3 Thesis Objectives

1. Analyze key factors influencing student performance.
2. Preprocess educational datasets (e.g., UCI Repository) for modeling.
3. Develop models like Random Forest, SVM, and Gradient Boosting.
4. Evaluate models using Accuracy, Precision, and F1-Score.
5. Propose an interpretable framework for early warning systems.

1.1 Overview

1.2 Problem Definition

1.3 Thesis Objectives

1.4 Thesis Contribution

1.5 Thesis Organization

Chapter 2 Background and Related Work

2.1 Introduction to EDM and ML

Educational Data Mining (EDM) applies data mining to understand students and learning settings. Machine Learning (ML) provides the computational tools for EDM, identifying complex patterns that traditional statistics might miss. Together, they enable proactive student support through predictive modeling.

2.2 Related Work on Student Performance Prediction

Source/Site	Study	Problem	Solution
Nature / Scientific Reports	Al-Shami et al. (2025)	Model interpretability (Black-box models)	Ensemble voting regression with LIME/SHAP
Plos One	Xu and Kim (2024)	Suboptimal feature selection in high-dimensional data	Ant colony algorithm for optimized prediction
J. Adv. Res. Appl. Sci.	Haron et al. (2025)	Reliance on quantitative data only	Hybrid methodology (Quantitative + Qualitative)
Procedia Computer Science	Vimarsha et al. (2024)	Dynamic nature of teaching-learning interaction	Co-evolutionary hybrid intelligence model
Applied Comp. Intel.	Ahmed (2024)	Difficulty in grouping students without labels	Clustering-based approach with SVM

IEEE Access	Chen et al. (2023)	Overlooking social/relational aspects	Bipartite network methodology with clustering
IEEE Access (SLR)	Rizwan et al. (2025)	High dropout rates in MOOCs	Deep learning for temporal behavioral patterns
African J. Appl. Res.	Karim-Abdallah et al. (2025)	Lack of standardized continuous monitoring	ML-based real-time monitoring framework
Computers (MDPI)	Angeioplastis et al. (2025)	Low accuracy from isolated course data	Correlation analysis between different courses
Scientific Reports	Marín et al. (2025)	Non-linear learning behaviors	EDM algorithms for non-linear modeling
Nature Scientific Reports	Malik et al. (2025)	One-size-fits-all approach	Individualized learning through advanced EDM
Frontiers in Psychology	Lu et al. (2024)	Exclusion of psychological factors	Inclusion of metacognitive and motivational factors
Tech Know Learn	Abu Saa et al. (2019)	Inconsistent feature selection	Systematic review of universal predictors

Int. J. Sci. Tech. Ed.	Adewale et al. (2018)	Noisy and incomplete educational data	Artificial Neural Networks (ANN) approach
Int. J. Nonlinear Anal.	Ali et al. (2022)	Remote teaching challenges (Covid-19)	CNN feature extraction for student analysis

Al-Shami et al. (2025): This study addressed the problem of model interpretability (black-box models). To solve this, the authors proposed ensemble voting regression with lime/shap, which significantly improved prediction accuracy and provided actionable insights for educators.

Xu and Kim (2024): This study addressed the problem of suboptimal feature selection in high-dimensional data. To solve this, the authors proposed ant colony algorithm for optimized prediction, which significantly improved prediction accuracy and provided actionable insights for educators.

Haron et al. (2025): This study addressed the problem of reliance on quantitative data only. To solve this, the authors proposed hybrid methodology (quantitative + qualitative), which significantly improved prediction accuracy and provided actionable insights for educators.

Vimarsha et al. (2024): This study addressed the problem of dynamic nature of teaching-learning interaction. To solve this, the authors proposed co-evolutionary hybrid intelligence model, which significantly improved prediction accuracy and provided actionable insights for educators.

Ahmed (2024): This study addressed the problem of difficulty in grouping students without labels. To solve this, the authors proposed clustering-based approach with svm, which significantly improved prediction accuracy and provided actionable insights for educators.

Chen et al. (2023): This study addressed the problem of overlooking social/relational aspects. To solve this, the authors proposed bipartite network methodology with clustering, which significantly improved prediction accuracy and provided actionable insights for educators.

Rizwan et al. (2025): This study addressed the problem of high dropout rates in moocs. To solve this, the authors proposed deep learning for temporal behavioral patterns, which significantly improved prediction accuracy and provided actionable insights for educators.

Karim-Abdallah et al. (2025): This study addressed the problem of lack of standardized continuous monitoring. To solve this, the authors proposed ml-based real-time monitoring framework, which significantly improved prediction accuracy and provided actionable insights for educators.

Angeioplastis et al. (2025): This study addressed the problem of low accuracy from isolated course data. To solve this, the authors proposed correlation analysis between different courses, which significantly improved prediction accuracy and provided actionable insights for educators.

Marín et al. (2025): This study addressed the problem of non-linear learning behaviors. To solve this, the authors proposed edm algorithms for non-linear modeling, which significantly improved prediction accuracy and provided actionable insights for educators.

Malik et al. (2025): This study addressed the problem of one-size-fits-all approach. To solve this, the authors proposed individualized learning through advanced edm, which significantly improved prediction accuracy and provided actionable insights for educators.

Lu et al. (2024): This study addressed the problem of exclusion of psychological factors. To solve this, the authors proposed inclusion of metacognitive and motivational factors, which significantly improved prediction accuracy and provided actionable insights for educators.

Abu Saa et al. (2019): This study addressed the problem of inconsistent feature selection. To solve this, the authors proposed systematic review of universal predictors, which significantly improved prediction accuracy and provided actionable insights for educators.

Adewale et al. (2018): This study addressed the problem of noisy and incomplete educational data. To solve this, the authors proposed artificial neural networks (ann) approach, which significantly improved prediction accuracy and provided actionable insights for educators.

Ali et al. (2022): This study addressed the problem of remote teaching challenges (covid-19). To solve this, the authors proposed cnn feature extraction for student

analysis, which significantly improved prediction accuracy and provided actionable insights for educators.

Chapter 4 Proposed Approach (Methodology)

3.1 Dataset Description

The mathematics dataset from the UCI Machine Learning Repository (395 instances, 33 attributes) is used. It includes grades, demographic, and social features. The target (G3) is categorized into 'Pass' and 'Fail'.

3.2 Proposed Framework

1. Data Collection: Gathering academic and socio-demographic data.
2. Preprocessing: Cleaning, encoding, and normalization.
3. Modeling: Training Random Forest, SVM, and Gradient Boosting.
4. Evaluation: Using Accuracy, F1-Score, and SHAP for interpretation.

Chapter 5 :Experimental Results and Discussion

Chapter 6 Conclusion and Future Work

References

- Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-16.
- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618.
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 160.
- Al-Shami, S. A., et al. (2025). Machine learning-based academic performance prediction using ensemble voting regression and interpretable AI. *Scientific Reports*, 15(1), 12353-4.
- Xu, H., & Kim, M. (2024). Combination prediction method of students' performance based on ant colony algorithm. *Plos One*, 19(3), e0300010.
- Haron, N. H., et al. (2025). An artificial intelligence approach to monitor and predict student academic performance. *J. Adv. Res. Appl. Sci. Eng. Technol.*, 44(1), 105–119.
- Vimarsha, K., et al. (2024). Student performance prediction: A Co-Evolutionary hybrid intelligence model. *Procedia Computer Science*, 235, 436–446.
- Ahmed, E. (2024). Student performance prediction using machine learning algorithms. *Applied Computational Intelligence and Soft Computing*, 2024(1), 4067721.
- Chen, Z., et al. (2023). Student performance prediction approach based on educational data mining. *IEEE Access*, 11, 131260–131272.
- Rizwan, S., et al. (2025). Identifying the factors affecting student academic performance and engagement prediction in mooc using deep learning: A systematic literature review. *IEEE Access*.
- Karim-Abdallah, B., et al. (2025). Application of machine learning algorithms in predicting academic performance of students in higher education institutes (HEIs): A systematic review and bibliographic analysis. *African Journal of Applied Research*.

Angeioplastis, A., et al. (2025). Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques. *Computers*, 14(3), 83.

Marín, Y. R., et al. (2025). Analysis of factors affecting the academic performance of university students using machine learning. *Scientific Reports*, 15, 44027.

Malik, S., et al. (2025). Advancing educational data mining for enhanced student performance prediction. *Nature Scientific Reports*, 15, 92324.

Lu, J., et al. (2024). Machine learning analysis of factors affecting college students' academic performance. *Frontiers in Psychology*, 15, 1447825.

Abu Saa, A., et al. (2019). Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. *Tech Know Learn* 24, 567–598.

Adewale, A. M., et al. (2018). Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach. *International Journal of Science and Technology Educational Research* 9(1), 1-8.

Ali, S., et al. (2022). Multi-class LDA classifier and CNN feature extraction for student performance analysis during Covid-19 pandemic. *International Journal of Nonlinear Analysis and Applications*, 13(1), 1329–1339.

[Abiodun, Oludare Isaac, et al., 2018]

Abiodun, Oludare Isaac, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. “State-of-the-art in artificial neural network applications: A survey”. *Heliyon* 4. 2018.

[Abu Saa, Amjed, et al., 2019]

Abu Saa, Amjed, Mostafa Al-Emran, and Khaled Shaalan. “Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques”. *Tech Know Learn* 24, 567–598., 2019.

[Adewale, Amoo M., et al., 2018]

Adewale, Amoo M., Alaba O. Bamidele, and Usman O. Lateef. “Predictive modelling and analysis of academic performance of secondary school students:

Artificial Neural Network approach”. International Journal of Science and Technology Educational Research 9(1), 1-8, 2018.

[Ali S, Rasheed Mansoor, and S Perumal, 2022]

Ali S, Rasheed Mansoor, and S. Perumal. “Multi-class LDA classifier and CNN feature extraction for student performance analysis during Covid-19 pandemic”. International Journal of Nonlinear Analysis and Applications, 13(1), pp. 13291339., 2022.

[Alpaydin, Ethem, 2010]

Alpaydin, Ethem (2010). Introduction to Machine Learning. London: The MIT Press. ISBN 978-0-262-01243-0