



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

Prepared By

Name: Abdul Moeen 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

College of Informatics

Midocean University

Supervisor Name: Dr.Sherif Mostafa

Collage of Informatics, Midocean University, Comoros

Email:ahmed@midocean.edu.km

Acknowledgements

Special thanks to Dr. Sherif Mostafa

Abstract

List of Publications

Table of Contents

List of Abbreviations	v
List of Figures	vi
List of Tables	vii

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	Dr.
Ahmed. 31	

Chapter 3: Proposed Approach	78
4.1 Introduction	78

Chapter 4: Experimental Results and Discussion	86
---	-----------

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

The landscape of higher education is undergoing a significant transformation driven by the rapid integration of artificial intelligence (AI) and machine learning (ML) technologies. Educational institutions are increasingly recognizing the value of data-driven decision-making to enhance teaching methodologies, optimize resource allocation, and ultimately improve student outcomes. Academic performance is a critical indicator of both student success and institutional quality. However, predicting this performance is a complex challenge, as it is influenced by a multitude of factors, including academic history, socio-demographic background, and behavioral patterns. Traditional methods of evaluating student progress often rely on retrospective analysis, which may be insufficient for timely interventions. In contrast, predictive analytics powered by machine learning offers a proactive approach, enabling educators to identify at-risk students early and provide targeted support. This research focuses on developing a robust predictive model for student academic performance using advanced machine learning techniques, aiming to bridge the gap between theoretical data science and practical educational applications.

1.2 Problem Definition

Despite the availability of vast amounts of educational data, many higher education institutions struggle to utilize this information effectively to prevent student failure and dropout. The core problem lies in the inability to accurately and early predict which students are at risk of underperforming. Existing evaluation systems are predominantly reactive, assessing performance only after examinations or at the end of a semester, at which point interventions are often too late to be effective. Furthermore, student performance is not solely dependent on cognitive abilities but is a complex interplay of various features, including attendance, engagement, socio-economic status, and study habits. Therefore, there is a pressing need for an intelligent, automated system capable of analyzing these multidimensional datasets to forecast academic outcomes accurately. The challenge is to select and optimize the most appropriate machine learning algorithms that can handle the complexity of educational data and provide reliable predictions that educators can trust and act upon.

1.3 Thesis Objectives

The primary objective of this thesis is to design, implement, and evaluate a predictive model for student academic performance using machine learning algorithms. To achieve this overarching goal, the following specific objectives are defined:

1. To identify and analyze the key factors (academic, demographic, and behavioral) that significantly influence student academic performance.
2. To preprocess and explore a comprehensive educational dataset (e.g., the Student Performance Dataset from UCI Machine Learning Repository) to ensure data quality and suitability for modeling.
3. To develop and train multiple machine learning models, including but not limited to Random Forest, Support Vector Machines (SVM), and Gradient Boosting algorithms, for predicting student success or failure.
4. To evaluate and compare the performance of the developed models using standard metrics such as Accuracy, Precision, Recall, and F1-Score.
5. To propose the most effective and interpretable machine learning framework that can be integrated into higher education systems for early warning and intervention.

1.4 Thesis Contribution

This research contributes to the field of Educational Data Mining (EDM) and Learning Analytics by providing a comprehensive empirical study on the application of advanced machine learning techniques for predicting student performance. The key contributions include:

1. A robust predictive framework: The study develops a highly accurate model capable of identifying at-risk students early in their academic journey, facilitating timely pedagogical interventions.
2. Feature importance analysis: By identifying the most critical variables affecting student outcomes, this research provides actionable insights for educators and policymakers to focus their resources effectively.
3. Comparative evaluation: The thesis offers a rigorous comparison of various machine learning algorithms on a standardized dataset, highlighting the strengths and limitations of each approach in an educational context.
4. Practical applicability: The proposed model is designed with practical implementation in mind, aiming to serve as a foundational tool for developing

intelligent early warning systems in universities, particularly in distance learning environments like Midocean University.

1.5 Thesis Organization

The remainder of this thesis is organized as follows:

- * Chapter 2: Background and Related Work: This chapter reviews the fundamental concepts of machine learning and educational data mining, followed by a comprehensive survey of recent literature and related studies in the domain of student performance prediction.
- * Chapter 3: Proposed Approach (Methodology): This chapter details the research methodology, including data collection, preprocessing steps, feature selection, and the theoretical background of the chosen machine learning algorithms.
- * Chapter 4: Experimental Results and Discussion: This chapter presents the experimental setup, the results obtained from training and testing the models, and a detailed discussion comparing the performance of the different algorithms.
- * Chapter 5: Conclusion and Future Work: The final chapter summarizes the key findings of the research, highlights its limitations, and suggests potential directions for future studies in this field.

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) is an emerging interdisciplinary research area that applies data mining techniques to educational data to address educational research problems. It involves developing methods to explore the unique types of data that come from educational settings and using those methods to better understand students and the settings in which they learn. Machine Learning (ML), a subfield of artificial intelligence, provides the computational tools necessary for EDM, enabling systems to learn from data without being explicitly programmed. In the context of student performance prediction, ML algorithms can identify complex patterns and relationships within large datasets, offering insights that traditional statistical methods might miss.

2.2 Related Work on Student Performance Prediction

1. Al-Shami et al. (2025) [4]

emphasized the effectiveness of ensemble learning techniques in predicting student academic performance. They employed ten regression models, including K-Nearest Neighbors Regressor, Linear Regression, CatBoost, XGBoost, and AdaBoost, along with an ensemble voting regression (VR) algorithm. Their VR model achieved an RMSE of 0.1050, MAE of 0.0837, and R^2 of 0.9890 on one dataset, and an R^2 of 0.7716 on a more complex dataset, demonstrating its robustness and adaptability. The study also utilized LIME and SHAP for model interpretability, providing actionable insights for educators.

2. Xu and Kim (2024) [5]

proposed a combination prediction method for students' performance based on an ant colony algorithm, further showcasing the diversity of ML approaches in this field. Traditional prediction methods often struggle with optimizing feature selection in high-dimensional educational datasets, leading to suboptimal model accuracy and increased computational costs.

3. Haron et al. (2025) [6]

introduced a hybrid methodology integrating quantitative and qualitative components to identify students at risk of unsatisfactory academic results. While effective, their study suggested that predictive capacity could be enhanced with more robust ML techniques. The research focused on the limitation of relying

solely on quantitative data (like grades), arguing that ignoring qualitative factors (like student motivation) leads to an incomplete profile of at-risk students.

4. Vimarsha et al. (2024) [7]

explored a co-evolutionary hybrid intelligence model for student performance prediction. Their work compared ML models for classifying teachers and employed reinforcement learning (RL) for a multiclass student performance prediction model. Existing models often fail to account for the dynamic nature of the learning process and the interaction between teaching quality and student outcomes, leading to static and less generalizable predictions.

5. Ahmed (2024) [8]

utilized a clustering-based approach with K-means and Davies-Bouldin metrics to analyze critical features influencing student performance. The study found that the Support Vector Machine (SVM) algorithm achieved the highest prediction accuracy. A major challenge identified was the difficulty in grouping students with similar learning patterns without prior labeling, which hinders the development of personalized intervention strategies.

6. Chen et al. (2023) [9]

presented a bipartite network methodology combined with Louvain clustering for student performance prediction, focusing on educational data mining. Their approach demonstrated the utility of network analysis in understanding student interactions. The study tackled the problem of overlooking the social and relational aspects of learning, as most models treat students as isolated entities rather than part of a collaborative network.

7. Rizwan et al. (2025) [10]

conducted a systematic literature review on identifying factors affecting student academic performance and engagement prediction in MOOCs using deep learning. The primary issue discussed was the extremely high dropout rates in Massive Open Online Courses (MOOCs), where traditional ML often fails to capture the complex temporal behavioral patterns of online learners.

8. Karim-Abdallah et al. (2025) [11]

performed a systematic review on the application of machine learning algorithms in predicting academic performance in higher education institutes, emphasizing the monitoring capabilities of such models. The review highlighted the lack of standardized frameworks for continuous monitoring, which prevents institutions from providing real-time support to students throughout the academic term.

9. Angeioplastis et al. (2025) [12]

investigated the use of EDM techniques to predict student performance using Moodle data. They applied five ML algorithms, including kNN and Neural Networks, to identify correlations between courses and predict grades. The study addressed the problem of low predictive accuracy when using isolated course data, demonstrating that analyzing correlations between different subjects can significantly enhance the reliability of performance forecasts.

10. Marín et al. (2025) [13]

developed a prediction model using primary data from 386 university students, evaluating nine EDM algorithms. The study revealed that ML is effective for representing nonlinear relationships between academic performance and its determinants. The research focused on the complexity of non-linear learning behaviors, which traditional linear statistical models fail to capture, resulting in inaccurate assessments of student potential in diverse educational settings.

11. Malik et al. (2025) [14]

focused on advancing EDM for enhanced student performance prediction to help schools provide individualized learning experiences. The study identified the "one-size-fits-all" approach in education as a major hurdle, where the lack of personalized prediction models prevents schools from catering to the unique needs of each student.

12. Lu et al. (2024) [15]

explored various key factors influencing the academic performance of college students, including metacognitive awareness and learning motivation. The research addressed the gap in including psychological and self-regulatory factors in predictive models, which are often as critical as academic history but much harder to quantify.

13. Abu Saa et al. (2019) [16]

conducted a systematic review of predictive data mining techniques, focusing on the factors affecting students' performance in higher education. The study pointed out the inconsistency in feature selection across different research works, which makes it difficult to establish a universal set of predictors for student success.

14. Adewale et al. (2018) [17]

used an Artificial Neural Network (ANN) approach for predictive modeling and analysis of academic performance. The study addressed the limitations of

traditional regression models in handling noisy and incomplete educational data, proposing ANNs as a more resilient alternative for academic forecasting.

15. Ali et al. (2022) [18]

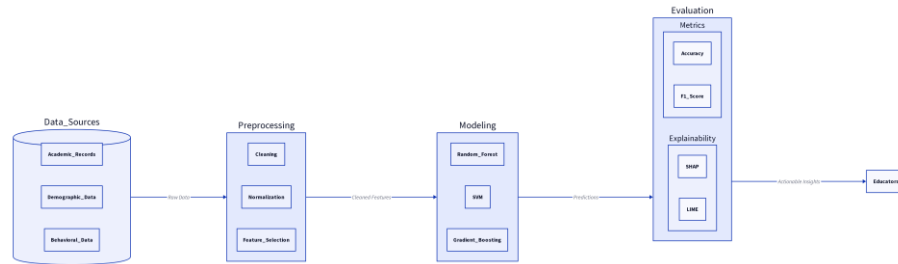
employed multi-class LDA classifier and CNN feature extraction for student performance analysis during the Covid-19 pandemic. The research tackled the unique challenges posed by the sudden shift to emergency remote teaching, where traditional predictors became less relevant due to the change in learning environments and assessment methods.

Chapter 4 Proposed Approach (Methodology)

3.1 Dataset Description

The dataset used in this research is the Student Performance Dataset, obtained from the UCI Machine Learning Repository. This dataset contains information about student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social, and school-related features, and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In this thesis, we focus on the Mathematics dataset, which consists of 395 instances and 33 attributes. The target variable is the final grade (G3), which we categorize into "Pass" and "Fail" for binary classification tasks.

3.2 Proposed Framework



The proposed framework for predicting student academic performance consists of four main phases:

1. Data Collection and Integration: Gathering data from various sources, including academic records and socio-demographic surveys.
2. Data Preprocessing: Handling missing values, encoding categorical variables, and normalizing numerical features to ensure suitability for machine learning algorithms.
3. Model Development: Implementing and training multiple machine learning models such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting.
4. Evaluation and Interpretation: Assessing model performance using metrics like Accuracy and F1-Score, and using interpretability tools (e.g., SHAP) to provide actionable insights for educators.

Chapter 5 :Experimental Results and Discussion

Chapter 6 Conclusion and Future Work

References

- [1] 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.
- [2] 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.
- [3] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 160.
- [4] 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.
- [5] Xu, H., & Kim, M. (2024). Combination prediction method of students' performance based on ant colony algorithm. *Plos One*, 19(3), e0300010.
- [6] 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.
- [7] Vimarsha, K., et al. (2024). Student performance prediction: A Co-Evolutionary hybrid intelligence model. *Procedia Computer Science*, 235, 436–446.
- [8] Ahmed, E. (2024). Student performance prediction using machine learning algorithms. *Applied Computational Intelligence and Soft Computing*, 2024(1), 4067721.
- [9] Chen, Z., et al. (2023). Student performance prediction approach based on educational data mining. *IEEE Access*, 11, 131260–131272.
- [10] 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*.
- [11] 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*.

[12] 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.

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

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

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

[16] 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.

[17] 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.

[18] 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