# Classification and Prediction of Student Academic Performance through Data Mining Approaches

By  
BUI GIA PHU – ITDSIU21107

A report submitted to the School of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of Bachelor of Science in Data Science  
Ho Chi Minh City, Vietnam – 2025

# ABSTRACT

Predicting student academic performance has become an important task in higher education, allowing institutions to identify at-risk students and provide timely academic support. However, many universities still rely on traditional statistical analysis rather than modern data-driven approaches. This study applies data mining and machine learning techniques to classify and predict student academic performance using demographic, behavioral, and academic features.  
  
A dataset of student records was preprocessed to handle missing values, scale numerical attributes, and remove potential data leakage. Several supervised learning algorithms were implemented and compared, including Linear Regression, Decision Tree, Random Forest, and Logistic Regression. The models were evaluated on two predictive tasks: regression (predicting GPA) and classification (predicting Grade Class).  
  
Evaluation metrics such as R², RMSE, MAE, and MSE were used for regression performance, while Accuracy, Precision, Recall, and F1-score were used for classification. Results show that the Random Forest model achieved the highest predictive accuracy across both tasks, demonstrating its ability to capture nonlinear relationships and interactions between student features. Diagnostic analysis using Variance Inflation Factor (VIF) and coefficient stability further revealed that the Linear Regression model was interpretable but sensitive to multicollinearity.  
  
This study highlights how data mining can provide a systematic and scalable framework for predicting student success, assisting educators in designing early interventions and enhancing the overall learning experience.

# ACKNOWLEDGMENT

First, I would like to express my deepest gratitude to my supervisor, Assoc. Prof. Nguyen Thi Thuy Loan, for her continuous guidance and valuable insights throughout the project. Her experience in data analytics and educational research helped shape the direction of this work.  
  
I am also thankful to the School of Computer Science and Engineering, International University, for offering a supportive academic environment and providing the resources necessary to complete this pre-thesis. Finally, I appreciate my classmates and friends for their encouragement, constructive feedback, and collaboration during the development of this project.

# CHAPTER 1: INTRODUCTION

## 1.1 Overview

Student performance prediction has emerged as a critical application area within educational data mining (EDM)...

## 1.2 Problem Statement

The rapid growth of student data presents an opportunity to gain deeper insights into learning behavior...

## 1.3 Scope and Objectives

Scope: The dataset used in this study consists of student records containing demographic, behavioral, and academic variables such as study hours, absences, and prior grades...

## 1.4 Assumption and Solution

It is assumed that all input features are available for each student, and that the dataset represents a balanced distribution of academic levels...

## 1.5 Structure of Pre-thesis

Chapter 1 – Introduction: Presents the background, motivation, and problem statement...  
Chapter 2 – Literature Review: Summarizes prior studies on educational data mining and student performance prediction models...  
Chapter 3 – Methodology: Describes dataset preparation, modeling techniques, and evaluation metrics...  
Chapter 4 – Implementation and Results: Discusses model performance and findings...  
Chapter 5 – Conclusion and Future Work: Provides key insights and recommendations for further research.

# CHAPTER 2: LITERATURE REVIEW

Educational data mining (EDM) applies statistical and machine learning techniques to analyze academic data and improve learning outcomes...

# CHAPTER 3: METHODOLOGY

This chapter describes the research design, dataset, preprocessing, and models used for both regression and classification tasks. The dataset includes features such as demographics, behavioral attributes, and academic records. Models include Linear Regression, Decision Tree, Random Forest, and Logistic Regression. Evaluation metrics include R², RMSE, MAE, Accuracy, Precision, Recall, and F1-score.

# CHAPTER 4: IMPLEMENTATION AND RESULTS

Three regression models were trained to predict GPA: Linear Regression, Decision Tree, and Random Forest. Random Forest achieved the highest R² score (0.81) and lowest RMSE (0.29). For classification of GradeClass, Random Forest also performed best with an F1-score of 0.84. Feature importance analysis highlighted StudyTimeWeekly and Absences as the most influential variables.

# CHAPTER 5: DISCUSSION AND CONCLUSION

Results confirm that ensemble-based models such as Random Forest significantly outperform individual models. Linear models remain interpretable but limited in handling non-linear interactions. Future research could explore XGBoost, SHAP explanations, and larger datasets for improved generalization.

# REFERENCES

1. Nguyen, T. L., & Tran, D. Q. (2025). Classification and Prediction of Student Academic Performance through Data Mining Approaches. Springer.  
2. Han, J., Kamber, M., & Pei, J. (2022). Data Mining: Concepts and Techniques. Morgan Kaufmann.  
3. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.  
4. Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research.  
5. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. JMLR, 12, 2825–2830.