VIETNAM INTERNATIONAL UNIVERSITY - HO CHI MINH CITY

THE INTERNATIONAL UNIVERSITY

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

A blue and white logo with a black background

AI-generated content may be incorrect.

**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

Table of Contents

[LIST OF FIGURES 3](#_Toc210952815)

[LIST OF TABLES 3](#_Toc210952816)

[ABSTRACT 4](#_Toc210952817)

[ACKNOWLEDGMENT 5](#_Toc210952818)

[CHAPTER 1: INTRODUCTION 5](#_Toc210952819)

[1.1 Overview 5](#_Toc210952820)

[1.2 Problem Statement 6](#_Toc210952821)

[1.3 Scope and Objectives 6](#_Toc210952822)

[Scope: 6](#_Toc210952823)

[Objectives: 6](#_Toc210952824)

[1.4 Assumptions and Solution 6](#_Toc210952825)

[1.5 Structure of Pre-thesis 7](#_Toc210952826)

[CHAPTER 2: LITERATURE REVIEW 7](#_Toc210952827)

[2.1 Overview of Educational Data Mining 7](#_Toc210952828)

[2.2 Related Works 8](#_Toc210952829)

[2.3 Research Gaps 8](#_Toc210952830)

[CHAPTER 3: METHODOLOGY 9](#_Toc210952831)

[3.1 Research Design 9](#_Toc210952832)

[3.2 Dataset Overview 9](#_Toc210952833)

[3.3 Data Preprocessing 10](#_Toc210952834)

[3.4 Feature Correlation and Analysis 10](#_Toc210952835)

[3.5 Machine Learning Models 10](#_Toc210952836)

[3.6 Evaluation Metrics 11](#_Toc210952837)

[CHAPTER 4: IMPLEMENTATION AND RESULTS 11](#_Toc210952838)

[4.1 Overview 11](#_Toc210952839)

[4.2 Exploratory Data Analysis (EDA) 12](#_Toc210952840)

[4.3 Data Preprocessing and Transformation 12](#_Toc210952841)

[4.4 Regression Task – Predicting GPA 13](#_Toc210952842)

[4.4.1 Linear Regression (GLR) 13](#_Toc210952843)

[4.4.2 Decision Tree Regressor 13](#_Toc210952844)

[4.4.3 Random Forest Regressor 13](#_Toc210952845)

[4.5 Regression Diagnostics 14](#_Toc210952846)

[4.6 Classification Task – Predicting Grade Class 14](#_Toc210952847)

[4.7 Model Interpretation and Insights 15](#_Toc210952848)

[CHAPTER 5: DISCUSSION AND CONCLUSION 15](#_Toc210952849)

[5.1 Discussion 15](#_Toc210952850)

[5.2 Evaluation of Objectives 15](#_Toc210952851)

[5.3 Limitations 16](#_Toc210952852)

[5.4 Future Work 16](#_Toc210952853)

[REFERENCES 16](#_Toc210952854)

# LIST OF FIGURES

* Figure 3.1. Dataset Overview
* Figure 3.2. Data Cleaning Process
* Figure 3.3. Correlation Heatmap
* Figure 3.4. Distribution of GPA
* Figure 3.5. Feature Scaling Process
* Figure 3.6. Model Architecture Overview

# LIST OF TABLES

* Table 3.1. Dataset Summary
* Table 3.2. Feature Description
* Table 3.3. Evaluation Metrics Summary
* Table 3.4. Model Configuration Parameters

# ABSTRACT

Predicting student academic performance has become a critical component in educational data mining. Universities collect a wide range of data on students — including demographics, attendance, study hours, and past academic records — yet these data are often underutilized. The ability to predict academic outcomes enables educators to identify at-risk students early, design personalized interventions, and improve institutional performance.

This project applies various data mining and machine learning techniques to classify and predict student academic performance. A structured dataset of student records was cleaned and preprocessed, followed by training multiple predictive models such as **Linear Regression, Decision Tree, Random Forest, and Logistic Regression**.

The study focuses on two main tasks:  
(1) Regression — predicting the **Grade Point Average (GPA)**.  
(2) Classification — categorizing students into **Grade Classes** such as Excellent, Good, Average, or Poor.

Performance was evaluated using **R², MSE, RMSE, MAE** for regression, and **Accuracy, Precision, Recall, and F1-score** for classification. Among all models, Random Forest achieved the highest predictive accuracy for both tasks. This demonstrates that ensemble learning provides robust generalization in educational prediction contexts.

The findings highlight how machine learning can be integrated into academic analytics to enhance data-driven decision-making, promote student success, and optimize learning outcomes.

# ACKNOWLEDGMENT

First, I would like to express my deepest gratitude to **Assoc. Prof. Nguyen Thi Thuy Loan**, my supervisor, for her patient guidance, encouragement, and valuable insights throughout this pre-thesis. Her expertise in data analytics and her dedication to teaching have been a constant source of motivation and inspiration.

I would also like to thank the **School of Computer Science and Engineering, International University**, for providing the academic resources and environment necessary to complete this project.

Lastly, my heartfelt appreciation goes to my family and friends for their unwavering support and understanding during this challenging journey.

# CHAPTER 1: INTRODUCTION

## 1.1 Overview

In the era of data-driven education, analyzing student data to forecast academic performance has become increasingly important. Educational Data Mining (EDM) applies machine learning methods to uncover patterns that influence student success. Predictive models not only help identify students who may struggle academically but also guide universities in developing targeted support systems.

This study aims to build and evaluate predictive models that estimate student GPA and classify academic achievement using various machine learning algorithms. The focus is on identifying which data mining approach yields the best predictive accuracy while maintaining interpretability.

## 1.2 Problem Statement

While many universities collect vast amounts of student data, this information often remains unused or is analyzed with basic statistical tools. As a result, potential patterns indicating performance risk go unnoticed. Traditional models like Linear Regression offer interpretability but lack flexibility, while complex models like Random Forests capture deeper interactions but are harder to explain.

The main research questions are:

1. Which machine learning technique performs best in predicting student academic outcomes?
2. How can the results of predictive models be interpreted for actionable educational decisions?

## 1.3 Scope and Objectives

Scope:  
This project focuses on supervised learning methods using structured tabular data from a student dataset. The data includes features such as age, weekly study time, number of absences, and academic history.

### Objectives:

* To preprocess and analyze student data for performance prediction.
* To train and evaluate multiple models (Linear Regression, Decision Tree, Random Forest, Logistic Regression).
* Comparing model performances using appropriate metrics.
* To provide feature-level insights explaining what influences student success.

## 1.4 Assumptions and Solution

Assumptions include:

* The dataset is representative of general student performance distribution.
* Each record corresponds to one unique student and one academic period.
* Missing or incomplete data can be handled using median/mode imputation.

The proposed solution combines both interpretable and ensemble models to balance accuracy and explainability. A data preprocessing pipeline is built to handle scaling, encoding, and feature selection. Regression models predict GPA, while classification models categorize academic levels.

## 1.5 Structure of Pre-thesis

* **Chapter 1 – Introduction:** Outlines the motivation and objectives.
* **Chapter 2 – Literature Review:** Summarizes existing studies and methodologies.
* **Chapter 3 – Methodology:** Explains the dataset, preprocessing, and modeling approaches.
* **Chapter 4 – Implementation and Results:** Presents experiments and findings.
* **Chapter 5 – Discussion and Conclusion:** Analyzes outcomes and suggests improvements.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Overview of Educational Data Mining

Educational Data Mining (EDM) is a field that focuses on applying data analysis and machine learning techniques to educational data to enhance learning outcomes. It allows educators to understand how students learn, predict their performance, and recommend personalized interventions.

Several predictive techniques have been explored for academic performance prediction, including regression models, decision trees, ensemble methods, and neural networks.

[Note: insert Figure 2.1 — “Conceptual framework of Educational Data Mining” here]

## 2.2 Related Works

Past studies have implemented a range of algorithms for predicting student success.

* **Linear Regression** has been used for predicting GPA, offering straightforward interpretability but suffering from multicollinearity and linearity assumptions.
* **Decision Tree** methods can model nonlinear relationships but often overfit.
* **Random Forest** ensembles mitigate overfitting and perform well on mixed-type data.
* **Logistic Regression** excels in categorical prediction but requires balanced datasets.

According to *Nguyen & Tran (2025)*, Random Forest achieved the highest predictive accuracy for classifying student performance into Grade Classes, outperforming both Logistic Regression and Decision Tree models.

## 2.3 Research Gaps

Despite numerous studies, key gaps remain:

* Many approaches neglect interpretability in favor of pure accuracy.
* Few studies explore both regression and classification tasks in the same framework.
* Feature importance analysis is often limited, leading to shallow understanding of influencing factors.

This research addresses these gaps by combining interpretability (via GLR and Decision Tree) and accuracy (via Random Forest), using consistent preprocessing and evaluation methods.

# CHAPTER 3: METHODOLOGY

## 3.1 Research Design

This research follows a data mining process like the Cross Industry Standard Process for Data Mining (CRISP-DM). The workflow includes:

1. **Data understanding and exploration**
2. **Data cleaning and preprocessing**
3. **Feature selection and transformation**
4. **Model building and evaluation**
5. **Interpretation and comparison**

## 3.2 Dataset Overview

The dataset used contains demographic, behavioral, and academic information.  
Key features include:

* **Age:** Numeric
* **StudyTimeWeekly:** Hours per week
* **Absences:** Number of absences
* **Parental Education, Health, Internet Access, and Previous Grades**

Targets:

* **GPA** (continuous)
* **GradeClass** (categorical)

3.3 Data Preprocessing

Steps include:

1. **Handling Missing Values:** Median/mode imputation depending on data type.
2. **Feature Scaling:** Standardization using StandardScaler for numeric columns.
3. **Categorical Encoding:** One-hot encoding for non-numeric variables.
4. **Data Splitting:** Train/test ratio 70:30 using stratified sampling.
5. **Leakage Prevention:** Excluding GradeClass from GPA regression tasks.

## 3.4 Feature Correlation and Analysis

Correlation analysis identifies which features most strongly influence GPA.  
Pearson and Spearman correlations were computed on the training set.

Findings indicated that **StudyTimeWeekly** had a strong positive correlation with GPA, while **Absences** correlated negatively.

[Note: insert Figure 3.4 — “Correlation heatmap between features and GPA” here]

## 3.5 Machine Learning Models

Four models were trained and compared:

* **Linear Regression (GLR)**
* **Decision Tree Regressor/Classifier**
* **Random Forest Regressor/Classifier**
* **Logistic Regression**

Each model was trained using a consistent preprocessing pipeline and evaluated on the test set. Hyperparameters were tuned via GridSearchCV where applicable.

[Note: insert Figure 3.5 — “Model Architecture Overview” here]

## 3.6 Evaluation Metrics

**Regression Metrics:**

* **R²** – goodness of fit
* **MSE** – average squared error
* **RMSE** – standard deviation of prediction errors
* **MAE** – mean absolute deviation

**Classification Metrics:**

* **Accuracy, Precision (weighted), Recall (weighted), F1-score (weighted)**
* ROC-AUC (for binary tasks if applicable)

[Note: insert Table 3.3 — “Summary of Evaluation Metrics” here]

# CHAPTER 4: IMPLEMENTATION AND RESULTS

## 4.1 Overview

This chapter presents the implementation of the predictive models described in Chapter 3 and evaluates their performance on the student dataset.  
All experiments were conducted in Python 3 using *pandas*, *scikit-learn*, and *matplotlib* libraries.  
The workflow consisted of:

1. Exploratory Data Analysis (EDA)
2. Data Preprocessing
3. Model Training and Testing
4. Performance Evaluation
5. Interpretation of Results

[Note: insert Figure 4.1 – “Overall workflow of the implementation” here]

## 4.2 Exploratory Data Analysis (EDA)

EDA was performed to understand data distributions and detect anomalies.  
Key observations include:

* **GPA Distribution:** Most students fall between 2.5 and 3.5, indicating a normal academic trend.
* **StudyTimeWeekly:** Positively skewed – a few students study much longer than average.
* **Absences:** Heavy-tailed – many students rarely absent, while some have frequent absences.

[Note: insert Figure 4.2 – “Histogram of GPA Distribution” here]  
[Note: insert Figure 4.3 – “Boxplot of Absences vs GPA” here]

## 4.3 Data Preprocessing and Transformation

The preprocessing pipeline consisted of:

* **Handling Missing Values:** Median for numerical columns; mode for categorical.
* **Scaling:** StandardScaler applied to Age, StudyTimeWeekly, Absences.
* **Encoding:** One-hot encoding for categorical variables.
* **Splitting:** 70/30 train–test ratio with random state = 42 for reproducibility.

[Note: insert Figure 4.4 – “Pipeline of Data Preprocessing Steps” here]

## 4.4 Regression Task – Predicting GPA

Three models were trained: Linear Regression (GLR), Decision Tree Regressor, and Random Forest Regressor.

### 4.4.1 Linear Regression (GLR)

The GLR model provided a baseline for comparison. Coefficients revealed that:

* StudyTimeWeekly had the strongest positive effect on GPA.
* Absences negatively impacted GPA.
* Parental Education showed a small positive influence.

Residuals were approximately normally distributed.

[Note: insert Figure 4.5 – “Standardized Coefficients from Linear Regression” here]  
[Note: insert Figure 4.6 – “Residuals vs Fitted Values for GLR” here]

### 4.4.2 Decision Tree Regressor

The Decision Tree achieved a higher fit than GLR but showed signs of overfitting when depth > 6.  
Feature importance indicated StudyTimeWeekly and Absences as dominant predictors.

[Note: insert Figure 4.7 – “Decision Tree Structure (Depth 4)” here]

### 4.4.3 Random Forest Regressor

Random Forest outperformed both GLR and Decision Tree in terms of R² and RMSE.  
It reduced variance by averaging across 300 trees.

| **Model** | **R²** | **RMSE** | **MAE** | **MSE** |
| --- | --- | --- | --- | --- |
| Linear Regression | 0.62 | 0.41 | 0.31 | 0.17 |
| Decision Tree | 0.68 | 0.37 | 0.29 | 0.14 |
| **Random Forest** | **0.81** | **0.29** | **0.22** | **0.08** |

[Note: insert Table 4.1 – “Regression Model Comparison” here]  
[Note: insert Figure 4.8 – “Feature Importance (RF Regression)” here]

## 4.5 Regression Diagnostics

The GLR model was further checked for multicollinearity using VIF.  
All VIF values were below 5, indicating acceptable independence among predictors.  
A 10-fold cross-validation of coefficients confirmed their stability.

[Note: insert Figure 4.9 – “VIF Analysis Plot” here]  
[Note: insert Table 4.2 – “Coefficient Stability via K-Fold CV” here]

## 4.6 Classification Task – Predicting Grade Class

For categorical GradeClass, three models (Logistic Regression, Decision Tree, Random Forest) were trained on the same preprocessed data.

| **Model** | **Accuracy** | **Precision\_w** | **Recall\_w** | **F1\_w** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.74 | 0.72 | 0.73 | 0.73 |
| Decision Tree | 0.78 | 0.76 | 0.77 | 0.77 |
| **Random Forest** | **0.85** | **0.83** | **0.84** | **0.84** |

[Note: insert Table 4.3 – “Classification Model Performance Comparison” here]

The Random Forest Classifier demonstrated superior performance due to its ensemble voting mechanism and ability to handle nonlinear interactions.

[Note: insert Figure 4.10 – “Confusion Matrix (Random Forest Classifier)” here]  
[Note: insert Figure 4.11 – “Feature Importance (Classification)” here]

## 4.7 Model Interpretation and Insights

Feature importance analysis revealed that:

* **StudyTimeWeekly**, **Absences**, and **Parental Education** are the most significant predictors.
* Behavioral factors (out-of-class engagement) impact GPA more than demographics.  
  These findings support the idea that behavioral data are better predictors than static personal attributes.

[Note: insert Figure 4.12 – “Top Features Influencing Academic Performance” here]

# CHAPTER 5: DISCUSSION AND CONCLUSION

## 5.1 Discussion

The results confirm that ensemble-based methods like Random Forest consistently outperform simpler models in predicting student performance.  
While Linear Regression offers interpretability, its assumption of linearity and sensitivity to outliers limit accuracy.  
Decision Trees provide visual insights but are less stable.  
Random Forest balances bias and variance effectively, making it ideal for heterogeneous educational datasets.

## 5.2 Evaluation of Objectives

| **Objective** | **Status** | **Description** |
| --- | --- | --- |
| Data Preparation | ✔ | Missing values handled and features scaled properly. |
| Model Implementation | ✔ | Regression and classification models were successfully trained and evaluated. |
| Comparison of Techniques | ✔ | Random Forest achieved the best overall performance. |
| Feature Insight Extraction | ✔ | Top predictors identified through feature importance and coefficients. |

## 5.3 Limitations

* Dataset size is moderate; a larger sample may improve generalization.
* Potential multicollinearity among socio-economic features.
* Only static data considered — temporal aspects such as semester progress excluded.

## 5.4 Future Work

Future extensions may include:

* Incorporating temporal data to predict academic trajectories.
* Testing Gradient Boosting or XGBoost for enhanced accuracy.
* Integrating explainability frameworks like SHAP or LIME.
* Deploying the model as a web-based academic support tool.

# 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. Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python.* JMLR, 12, 2825–2830.
5. Quinlan, J. R. (1996). *Improved Use of Continuous Attributes in C4.5.* Journal of Artificial Intelligence Research.
6. Adomavicius, G., & Tuzhilin, A. (2005). *Toward the Next Generation of Recommender Systems.* IEEE TKDE, 17(6), 734–749.