**RESEARCH PROPOSAL**

**PREDICTING STUDENT PASSING GRADES USING MACHINE LEARNING**

**Topic**: Education on predicting secondary students' academic success.

2602199582 Fiona Maharani Nugraha Data Science 085311506199

2602190816 Kimberly Kayla Dewi Data Science 0818819779



BINUS University

2024

# TABLE OF CONTENTS

[TABLE OF CONTENTS 2](#_heading=h.30j0zll)

[**CHAPTER 1: INTRODUCTION 3**](#_heading=h.1fob9te)

[1.1 Background 3](#_heading=h.2et92p0)

[1.2 Problem Formulation 4](#_heading=h.tyjcwt)

[1.3 Scope 5](#_heading=h.3dy6vkm)

[1.4 Objectives and Benefits 5](#_heading=h.1t3h5sf)

[1.4.1 Objectives 5](#_heading=h.4d34og8)

[1.4.2 Benefits 5](#_heading=h.2s8eyo1)

[1.5 Methods 6](#_heading=h.17dp8vu)

[1.6 Writing System 6](#_heading=h.3rdcrjn)

[**CHAPTER 2: LITERATURE REVIEW 7**](#_heading=h.26in1rg)

[2.1 Machine Learning 8](#_heading=h.35nkun2)

[2.2 Educational Data Mining 8](#_heading=h.1ksv4uv)

[2.3 Regression and Evaluation Metrics 9](#_heading=h.bh05qbzrw50)

[2.4 Related Works 10](#_heading=h.44sinio)

[**CHAPTER 3: IMPLEMENTATION METHOD 25**](#_heading=h.2jxsxqh)

[3.1 Thinking Framework 25](#_heading=h.3j2qqm3)

[3.2 Research Methods 25](#_heading=h.1y810tw)

[3.2.1 Data Collection and Understanding 26](#_heading=h.4i7ojhp)

[3.2.2 Exploratory Data Analysis 26](#_heading=h.2xcytpi)

[3.2.3 Data Preprocessing 28](#_heading=h.1ci93xb)

[3.2.4 Scaling and Feature Selection 28](#_heading=h.x1moqffpoq7u)

[3.2.5 Model Training 29](#_heading=h.kb0fb5vqvlsk)

[3.2.6 Evaluation and Analysis 29](#_heading=h.pqi5scc80yi2)

[REFERENCES 30](#_heading=h.3as4poj)

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

Education plays a pivotal role in shaping individuals and societies, providing essential skills, knowledge, and opportunities for personal and professional growth. In secondary education, academic achievement is often measured through students’ grades, which serve as key indicators of their understanding and mastery of subjects. Forecasting student performance is an important aspect of educational research, especially for improving teaching strategies, personalized learning experience, and providing early support for students at risk. Traditional methods of assessing student performance, such as standardized tests and teacher evaluations, are often limited by subjective biases and do not fully account for the complex factors influencing academic success. Modern approaches such as machine learning have become popular for their ability to analyze large datasets, process vast amounts of data, and predict student outcomes with greater accuracy.

Machine learning (ML) has demonstrated its potential in the education sector by providing data-driven insights that enhance decision-making processes. By leveraging student data, machine learning models can uncover patterns and relationships that may not be immediately obvious through manual analysis. Several studies have shown that combining demographic, social, and behavioral data in addition to academic performance, can significantly improve the accuracy of predicting students' final grades. For instance, predictive models using data such as parental education, student attendance, and extracurricular activities have been found to yield reliable forecasts of student performance results. This demonstrates the importance of utilizing a wide range of variables beyond academic scores to better understand student success academically.

A common challenge in educational research is the complexity of the factors that affect student performance. Factors such as family background, socioeconomic status, and personal habits all contribute to students' academic outcomes. In this context, our research offers a comprehensive method for analyzing existing research on the application of machine learning in predicting academic success. We unify various studies to provide a broader understanding of the field, identify research gaps, and highlight best practices in using machine learning for educational purposes. By reviewing previous research, we hope that we can develop more effective interventions to support struggling students and enhance overall educational outcomes.

In the field of secondary education, predicting student success is particularly critical as it has long-term implications for a student’s future, especially higher education opportunities and career prospects. A study by Cortez and Silva (2008) introduced a dataset from two Portuguese secondary schools, focusing on student performance in mathematics and Portuguese language subjects. The study highlighted the significance of considering not only the academic grades of students but also their demographic, social, and behavioral attributes. By applying machine learning techniques to the dataset, researchers would be able to model student success under various regression tasks, emphasizing the value of predictive models in education.

This research proposal aims to conduct a project on predicting student passing grades using machine learning, focusing on secondary education. The study will provide a comprehensive analysis of the methods, datasets, and variables used in this research. Additionally, it will examine the effectiveness of machine learning models in predicting student performance, contributing to a deeper understanding of how education systems can benefit from data-driven predictions.

## 1.2 Problem Formulation

Based from the background presented, this research aims to address the following questions:

1. Which machine learning algorithms are most effective in predicting student passing grades in secondary education?
2. What are the key demographic, social, behavioral, and academic features that significantly influence student academic success?
3. To what features can predictive models accurately forecast final grades (G3) without relying on prior period grades (G1 and G2)?

## 1.3 Scope

This research focuses on assessing the student performance of three machine learning algorithms—K-Nearest Neighbors (KNN), Decision Tree, and Linear Regression—using the "Student Performance Dataset" from Kaggle, uploaded by Dev Ansodariya. The dataset contains various student-related features, including academic performance, demographic, behavioral, and social factors from Portuguese secondary schools.

The study aims on this specific dataset and these algorithms to explore their effectiveness in predicting student passing grades and focuses on comparing machine learning algorithms for predicting student passing grades within Portugal and does not extend to assessing the adaptability of these methods beyond Portugal or internationally.

For model evaluation, we will use several performance metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R²), ensuring a comprehensive analysis of the models' accuracy and reliability.

## 1.4 Objectives and Benefits

### 1.4.1 Objectives

1. To assess and identify the most effective machine learning algorithm (KNN, Decision Tree, or Linear Regression) for predicting student passing grades in secondary education.
2. To analyze key demographic, social, behavioral, and academic features that significantly influence student performance.
3. To evaluate the predictive accuracy of models in forecasting final grades (G3) without relying on prior period grades (G1 and G2).

### 1.4.2 Benefits

1. Assisting educators and institutions in selecting suitable machine learning models to predict student success, helping in early intervention strategies.
2. Offering insights into the critical factors that affect student academic outcomes, which can be used to design better student support programs.
3. Contributing to the educational data science field by showcasing the application of machine learning for improving academic success prediction.

## 1.5 Methods

This research begins by collecting the "Student Performance Dataset" from the Kaggle platform, which includes student data from Portuguese secondary schools. First, the dataset will be conducted with exploratory data analysis (EDA) to uncover key patterns and relationships within the data. Second, the dataset is cleaned and prepared through data pre-processing steps, such as handling missing values and label encoding. Third, we will select features that have a great impact on the students’ final grades (G3). Fourth, the dataset will then be split into training and testing sets using an 80:20 ratio to ensure an unbiased evaluation of the models.

Fifth, three machine learning models—K-Nearest Neighbors (KNN), Decision Tree, and Linear Regression—will be implemented to predict the students' final grades (G3). Each model will be trained on the training set and evaluated on the test set. Sixth, the performance of these models will be assessed using evaluation metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R²). This methodology ensures a thorough comparison of the models' effectiveness in predicting student performance.

## 1.6 Writing System

1. Title
2. Table of Contents
3. Chapter 1 – Introduction
4. Introduction – A big picture and background about the research topic.
5. Problem Formulation – Problems that will be investigated.
6. Scope – Research limitations and boundaries.
7. Objectives and Benefits – The purpose of the research and what benefits can be taken from the research.
8. Methods – Methods used in the research.
9. Writing System – The sequence or structure of writing scientific papers.
10. Chapter 2 – Literature Review
    1. Literature Review – Explanation of theories relevant to the research topic.
    2. Related Works – Comparison of this research with other similar studies.
11. Chapter 3 - Implementation Method
12. Thinking Framework – The steps that will be taken to solve the problem.
13. Research Methods – Detailed methods that will be used in the research.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Machine Learning

Machine Learning (ML) is a part of artificial intelligence that allows computers to learn from data and make decisions or predictions without being explicitly programmed. It enables machines to automatically improve from experience by identifying patterns in data, making it a powerful tool in predictive modeling and data analysis. ML algorithms fall into three categories, including supervised, unsupervised, and reinforcement learning. The focus of this study is on supervised learning, which uses labeled datasets to train models to predict outcomes based on input features. Popular supervised learning algorithms include Decision Tree, K-Nearest Neighbors (KNN), and Linear Regression (<https://doi.org/10.1088/1742-6596/1142/1/012012>).

Decision Tree (DT) is a widely used algorithm that splits data into branches based on feature value. K-Nearest Neighbors (KNN) makes predictions by analyzing the closest data points in feature space, while Linear Regression models the relationship between input variables and continuous outcomes. These algorithms have proven effective in various fields, including education, where they help predict student performance, behaviors, and outcomes. Machine learning’s ability to obtain insights from educational data helps to better understand students’ learning processes and enhance educational strategies.

## 2.2 Educational Data Mining

Educational Data Mining (EDM) refers to the process of applying machine learning and statistical methods to educational data to gain insights into student learning patterns, behaviors, and performance. With the rise of technology in education, large amounts of data are being collected from learning management systems, online platforms, and academic records. EDM can be used to predict student outcomes, identify at-risk students, and personalize learning experiences based on individual needs (Romero & Ventura, 2020). By using machine learning algorithms, EDM facilitates data-driven decision-making to improve educational practices and outcomes.

The **Student Performance Dataset** used in this research contains data related to student demographics, academic performance, and behavioral indicators. This dataset is ideal for building predictive models that assess factors influencing students’ academic achievements. EDM has previously been applied to similar datasets to identify key predictors of student academic success, enabling institutions to tailor interventions and support systems.

## 

## 2.3 Regression and Evaluation Metrics

In the context of educational data analysis, regression models are used to predict continuous variables, such as student grades, based on multiple input features. Linear Regression models the relationship between a dependent variable and one or more independent variables. It is commonly applied in educational research to forecast outcomes like final exam scores, GPA, or course completion rates (Han et al., 2021). These models enable educators and researchers to gain insights into the factors influencing academic performance and develop strategies to support student success.

Evaluating the performance of machine learning models in predicting educational outcomes involves various metrics to assess prediction accuracy. Common metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R²), and Mean Absolute Percentage Error (MAPE). MSE and RMSE measure the average squared differences between predicted and actual values, providing insights into the magnitude of errors. MAE quantifies the average magnitude of errors without considering their direction, offering a straightforward interpretation of prediction accuracy. R-squared explains the proportion of variance in the dependent variable accounted for by the model, serving as an indicator of the model's overall explanatory power.

Meanwhile, MAPE calculates the average percentage difference between predicted and actual values. This metric offers a normalized error measure, making it easier to compare models across datasets with different scales. However, it is less effective when the actual values are close to zero, as it can produce excessively high error percentages. Together, these metrics provide a comprehensive framework for evaluating model performance, ensuring that predictions not only align closely with actual outcomes but also generalize effectively to new data.

## 2.4 Related Works

Research related to student performance prediction using machine learning has been extensively explored. Suzan et al. (2021) conducted a study titled *Student Adaptability Level in Online Education Using Machine Learning Approaches*, where they applied various algorithms to classify student adaptability levels in online education. Their research focused on evaluating multiple machine learning algorithms, including Random Forest and Decision Tree. The Random Forest model demonstrated the best performance, achieving an accuracy of 89.63%. This research emphasized the role of machine learning algorithms in improving accuracy, highlighting that certain algorithms, like Random Forest, perform better in predicting adaptability in online education.

Another significant study by M. Wu, G. Subramaniam, D. Zhu, C. Li, H. Ding, and Y. Zhang (2024) titled "Using Machine Learning-based Algorithms to Predict Academic Performance - A Systematic Literature Review." Their findings highlighted that ensemble learning methods, specifically, showed the highest performance, reaching an average accuracy of 87.67%, closely followed by support vector machines (SVM) with 84.30% accuracy. The study underscored the significance of demographic, academic, and behavioral factors in predicting academic success. Additionally, it emphasized the importance of early identification and timely intervention to enhance educational outcomes.

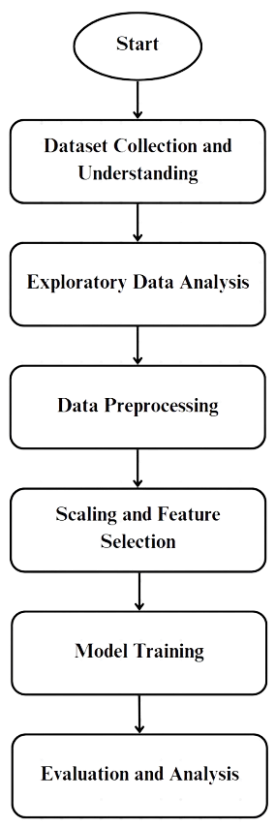
Both studies highlight the importance of feature selection and algorithm performance when applying machine learning to educational data, and their findings to improve adaptive learning techniques.

| **No.** | **Source (Journal/Conference)** | **References (Author, Year, Title)** | **Research Purpose** | **Dataset/Sample** | **Methods/Model/Algorithm** | **Result** |
| --- | --- | --- | --- | --- | --- | --- |
| 1. | Conference - ICIPTM (International Conference on Innovative Practices in Technology Management) | Wu, M., Subramaniam, G., Zhu, D., Li, C., Ding, H., & Zhang, Y. (2024). *Using Machine Learning-based Algorithms to Predict Academic Performance - A Systematic Literature Review.* | To conduct a systematic review of machine learning applications for predicting academic performance. | 83 indexed research articles (2020-2023). | Ensemble learning, Support Vector Machines (SVM), Neural Networks. | Ensemble learning showed highest accuracy (87.67%), followed by SVM (84.30%). Highlighted the importance of early identification and interventions to improve educational outcomes, aligning with SDG 4 (Quality Education). |
| 2. | Research Article - ResearchGate | Suzan, M., Rahman, M., & Hasan, F. (2021). Student Adaptability Level in Online Education Using Machine Learning Approaches | To evaluate algorithm accuracy on assessing student adaptability in online education. | Survey data from students at school, college, and university levels in Bangladesh | Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN) | Random Forest achieved the highest accuracy (89.63%), demonstrating its effectiveness in predicting student adaptability in online education. |
| 3. | Open Access Repository - MDPI (Multidisciplinary Digital Publishing Institute) | Balqis Albreiki, Hany Alashwal, & Nazar Zaki. (2021). *A Systematic Literature Review of Student’ Performance Prediction Using Machine Learning Techniques* | To understand and overcome challenges, that is predicting students at risk and students drop out prediction. | Data from student colleges/university databases and online learning platforms | Feed-Forward Network, SVM, PESFAM, and SEDM. | ML has the ability to accelerate educational advancement, and it is evident that educational efficiency is increasing dramatically. A weekly tutoring action plan was also recommended by this study review as a way to keep students from dropping out. |
| 4. | Research Article - Research Square | Kannan, R., Abarna, K. T. M., & Vairachilai, S. (2023), *Student Academic Performance Prognosticative Using Optimized Hybrid Machine Learning Algorithms* | To predict student performance in early higher education to prevent dropouts. | Dataset of 4424 students' academic progress. | Traditional classification algorithms, Hybrid ML algorithms, Stacking model. | The hybrid stacking model achieved the highest ROC-AUC scores (dropout: 0.91, enrolled: 0.82, graduate: 0.94). The study suggests using Graph Neural Networks (GNNs) to provide more accurate predictions in large and complex educational datasets. |
| 5. | Conference Proceedings - IOP Publishings (Institute of Physics) | Ali Salah Hashim *et al.* (2020). *Student Performance Prediction Model based on Supervised Machine Learning Algorithms.* | Forecast student success to help teachers prevent dropouts, identify students needing additional help, and improve institutional ranking, and prestige. | Records from the bachelor study programs at the College of Computer Science and Information Technology, University of Basra, 2017-2018, 2018-2019. | Several supervised machine learning algorithms, implemented using the Weka 3.8.0 software environment. | Logistic Regression was the best-performing algorithm. The accuracy was influenced by data cleanliness, feature domain, number of features, dataset size, and the domain of the final class. |
| 6. | Research Article - Research Square | Adil, K., Youness, M., Ahmed, A., & Ahmed, E. (2023), *Hybrid Machine Learning Algorithms at the Service of Student Performance.* | To improve student performance predictions by comparing different machine learning algorithms. | Student performance in various disciplines. | ANN, Decision Tree, ELM, KNN, Logistic Regression, Linear Regression, Naïve Bayes, Random Forest, SVM. | Random Forest has the highest accuracy (86%). |
| 7. | Open Access Journal - Nature Publishing Group (Scientific Reports) | Muhammad Bilal, Muhammad Omar, Waheed Anwar, Rahat H. Bokhari, & Gyu Sang Choi. (2022). *The Role of Demographic and Academic Features in A Student Performance Prediction.* | Predict the final semester performance of students in the Doctor of Veterinary Medicine (DVM) program using pre-admission academic achievements, demographics, and first-semester performance. | Data from 166 students from three sessions (2010-2015, 2012-2017, 2013-2018) of the DVM program at The Islamia University of Bahawalpur, Pakistan. | Various supervised ML algorithms (Decision Tree, Random Forest, SVM, KNN, and Logistic Regression) | SVM algorithm achieved the highest accuracy (92%). Academic features are strong indicators of student performance, while demographic features did not significantly impact performance prediction. |
| 8. | Research Article - Research Square | Tarek Abd El-Hafeez, Ahmed Omar (2022), *Student Performance Prediction Using Machine Learning Techniques.* | To identify students difficulties facing e-learning systems and support decision-making to enhance university e-learning. | Students using e-learning systems. | Logistic Regression, Decision Tree, Random Forest, SGD Classifier, Multinomial NB, K-Neighbors Classifier, Ridge Classifier, Nearest Centroid, Complement NB, Bernoulli NB. | Random Forest has the highest accuracy (0.844) and improves to 0.864 with parameter tuning. Key features affecting performance included visited resources, absence days, raised hands, and announcements view. |
| 9. | Open Access Library - AACE (Association for the Advancement of Computing in Education) | Farouk Ouatik, Mohammed Erritali, Fahd Ouatik, Mostafa Jourhmane. (2022). *Predicting Student Success Using Big Data and Machine Learning Algorithms.* | Some ML methods are not sufficient in time, especially for a big number of students. Big Data technology was used to minimize execution time. | Personal information, academic evaluation, the activities of the students in VLE, psychology, the student environment, practical work, number of student absences. | ML algorithms and educational Data Mining (KNN, C4.5, and SVM) | SVM reached 87.32% recognition rate and using the Big Data HDFS executed well in time. |
| 10. | Research Article - Research Square | Gautam Appasaheb Kudale, Sandeep Singh Rajpoot (2024), *Performance Evaluation of Different Machine Learning Algorithms on Student Dataset Clustered by K-means Algorithm.* | To evaluate ML algorithms on student clusters created by K-Means. | Kaggle student dataset | K-Means, KNN, Neural Network, Random Forest, SVM | Random Forest achieved the highest accuracy (92%), indicating the best performance among tested algorithms. |
| 11. | Open Access Journal - Frontiers Media | Zhang Y, Yun Y, An R, Cui J, Dai H and Shang X (2021). *Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis*. | Predicting student performance to help learners and educators improve their learning and teaching. | Data collection from Traditional Classroom, Online Classroom, Blending Classroom. | ML, Method-Learning methods, EDM, Feature selection methods to improve performance (Lasso, Selection Operator) | The method-learning methods could achieve a good performance, feature selection, both could boost Student Performance Prediction. |
| 12. | Research Article - Research Square | Thilanka Seneviratne, Supun Manathunga (2023), *Machine learning predicts student exam performance with high sensitivity allowing personalized interventions.* | To predict student exam performance for targeted interventions. | 583 medical students’ exam scores (F, S1, S2). | Multiple Linear Regression, Random Forest, KNN, SVM | Linear model performs better than the rest of the algorithms. It predicted S2 with 100% sensitivity and 64.7% specificity, AUC 88% for test data, allowing early interventions. |
| 13. | Open Access Repository - MDPI (Multidisciplinary Digital Publishing Institute) | Marina Segura, Jorge Mello, & Adolfo Hernández. (2022). *Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?* | To predict university student dropout rates at an early stage. Identifying key factors that influence dropout across different program areas. | Dropout candidates at the end of the students’ first year of the UCM. Includes information from five major program areas. | Feature Selection Process, Machine Learning Models (SVM, Decision Trees, ANN, Logistic Regression) | First-semester performance, academic performance, course preference influences the dropout predictions. Logistic Regression worked well as a baseline. Prediction success varied by program area. |
| 14. | Research Article - Research Square | Korchi Adil, Fayçal Messaoudi, Abatal Ahmed, Manzali Youness (2023), *Machine Learning and Deep Learning based Students’ Grades Prediction.* | Predict student grades using machine learning techniques. | Student personal data and grades. | Decision Tree, Random Forest, Linear Regression, K-Nearest Neighbors, XGBoost, Deep Neural Network | Deep Neural Network outperformed other algorithms with low error metrics ( R²=99.97%, MAE=0.45, MSE=0.05, RMSE=1.13). |
| 15. | Open Access Repository - MDPI (Multidisciplinary Digital Publishing Institute) | Diego Opazo, Sebastián Moreno, Eduardo Álvarez-Miranda, & Jordi Pereira. (2021). *Analysis of First-Year University Student Dropout through*  *Machine Learning Models: A Comparison between Universities* | Predict first-year dropout of engineering students and to identify key factors influence dropout risk, and to understand whether dropout prediction varies across different universities. | First-year engineering students of Universidad Adolfo Ibáñez and Universidad de Talca, include variables such as entrance exam scores, and other enrollment-related data. | ML Models (SVM, Decision Tree, Random Forest, Gradient-Boosting Decision Tree, Naive Bayes, Logistic Regression, ANN) | It is better to develop separate models for each university. Gradient-boosting decision trees performed best. Higher entrance test scores reduce dropout risk, higher language tests slightly increase dropout probability. |
| 16. | Research Article - Research Square | Sudais, M., Safwan, M., Khalid, M. A., & Ahmed, S. (2021), *Students’ Academic Performance Prediction Model Using Machine Learning.* | To predict and identify students who might fail semester exams to provide additional assistance from teachers. | Transcript data from a university, including CGPA and grades in all courses. | Naïve Bayes, Neural Network, Support Vector Machine (SVM), Decision Tree. | Naïve Bayes achieved the highest accuracy, followed by other algorithms with overall accuracies below 70%, indicating high error rates and unreliable predictions. Suggested future work includes expanding dataset attributes for better accuracy. |
| 17. | Institutional Repository - NM-AIST (Nelson Mandela African Institution of Science and Technology) | Neema Mduma, Khamisi Kalegele, & Dina Machuve. (2019). *Machine Learning Approach for Reducing Students Dropout Rates.* | Predict student dropout rates particularly in developing countries using ML algorithms. | Uwezo Annual Learning Assessment dataset for Tanzania. | ML Models (Logistic Regression, Random Forest, KNN, MLP) | Combining LR and MLP enhances predictive performance, combining models can effectively predict student dropout rates. |
| 18. | Research Article - Research Square | Ahmed, A., Tolera, D., & Shebera, W. (2023), *Multi-Category Prediction of Students’ Academic Performance Using Machine Learning: For Students Joining Higher Educational Institution in Ethiopia.* | To predict student academic performance in Ethiopian higher education and analyze significant predictive features. | Dataset from 3 semesters of Bule Hora University students | SVM, Random Forest, KNN, Gradient Boosting, Decision Tree | The Decision Tree model has the best performance with 97.3% testing accuracy. Significant predictive features are academic factors (entrance result, study time, attendance) and socio-demographic factors (age, gender, family background). Features like extra classes and guidance counseling had less impact. The study suggests future inclusion of additional pre-university data, such as exit exam results for better accuracy prediction. |
| 19. | Digital Library - IEEE Xplore (Institute of Electrical and Electronics Engineer) | Siti Dianah Abdul Bujang, Ali Selamat, Roliana Ibrahim, Ondrej Krejcar, Enrique Herrera-Viedma, Hamida Fujita, & Nor Azura MD. Ghani. (2021). *Multiclass Prediction Model for Student Grade Prediction Using Machine Learning.* | Improve predictive performance in student grade prediction by evaluating various ML techniques and to handle imbalance data effectively. | Two core courses of 1.282 real student course grades collected from the first-semester course from the Department of Information and Communication Technology at one of the Malaysia Polytechnics. | ML Models (Decision Tree, SVM, Naive Bayes, KNN, Logistic Regression, Random Forest), Multi-class Prediction Model (SMOTE, feature selection) | RF has the best performance from the combination of SMOTE and feature selection, it significantly improves the predictive accuracy and performance for multi-class imbalance student grade prediction. |
| 20. | Research Article - Research Square | Huiling Zhang, Huatao Wu, Zhengde Li, Wenwen Gong, Yan Yan (2024), *Machine Learning based Analysis of the Effect of Team Competition on College Students’ Academic Performance.* | To examine the impact of team competition on college students’ academic performance, determine effective competition design methods for improving academic performance. | Freshmen enrolled in a college English course. | XGBoost (Extreme Gradient Boosting), LGBM (Light Gradient Boosting Machine), Lasso, Ridge. | The study reduced prediction error by up to 30%. It provided insights into team competition's impact on academic performance, highlighting effective strategies for teaching design in college English courses. Future recommendations will include further field experiments and exploring additional courses for broader insights. |
| 21. | Digital Library - SpringerLink | Shah Hussain & Muhammad Qasim Khan. (2021). *Student‑Performulator: Predicting Students’ Academic Performance at Secondary and Intermediate Level Using Machine Learning.* | Predict students’ grades and marks using ML methods to improve educational quality and planning. | 30 selected attributes related to student performance from the Board of Intermediate & Secondary Education Peshawar in Pakistan. Includes historical academic data from seven regions. | EDM, supervised machine learning method (regression model, Decision Tree Classifier). | Regression model and DT-classifier are effective in predicting student performance in terms of both grade and marks. |
| 22. | Research Article - Research Square | Eka Miranda, Mediana Aryuni, Mia Ika Rahmawati, Siti Elda Hiererra, Dian Sano (2023), *Machine learning's model-agnostic interpretability on The Prediction of Students' Academic Performance in Video-Conference-Assisted Online Learning During the Covid-19 Pandemic.* | Predicting student performance in video-conference online learning during Covid-19. | 361 students (Sept 2022 - Jan 2023). | Random Forest, SVM, Gaussian Naive Bayes. | Random Forest achieved the highest accuracy (60.27%) and AUC (87%). |
| 23. | Open Access Repository - MDPI (Multidisciplinary Digital Publishing Institute) | Juan L. Rastrollo-Guerrero, Juan A. Gómez-Pulido, & Arturo Durán-Domínguez. (2020). *Analyzing and Predicting Students’ Performance by Means of Machine Learning: A Review* | Review and analyze modern techniques AI that are used to predict student performance, to provide insights that can support more effective academic strategies. | Literature review of approximately 70 research papers that cover techniques and applications in student performance prediction. | ML models, collaborative filtering and recommender systems, ANN. | The aforementioned methods influence data analysis effectively, bringing promising results in predicting student performance. |
| 24. | Research Article - Research Square | Hanieh Zehtab Hashemi, Rezvan Rahimi, Parvane Parvasideh, Zahra Bahrevar (2022), *Prediction of Students’ Performance in a National Medical Exam Using Machine Learning Techniques.* | Predicting students’ performance on a national medical exam. | National medical exam dataset. | Neural Network, Deep Learning, Random Forest. | Random Forest achieved the highest accuracy (96.16%), followed by Deep Learning (95.64%) and Neural Network (85.6%). |
| 25. | Open Access Repository - MDPI (Multidisciplinary Digital Publishing Institute) | Diego Buenaño-Fernández , David Gil, & Sergio Luján-Mora. (2019). *Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study.* | Improve educational quality in alignment with sustainable development goals (SDGs), focusing on predicting students’ finals grades (FGs). | Historical dataset of student grades, academic records of 335 students in a total of 68 subjects into seven knowledge areas.) from computer engineering degree program at an Ecuadorian University. | Data Collection and Preprocessing, Grouping Students, Supervised Learning ML Models, Experimental Process. | ML techniques are effective in predicting students’ academic performance. The writer claims that they plan to design an architecture that uses big data tools since the data they use are large. |
| 26. | Research Article - Research Square | Rumana Rois, Manik Ray, Atikur Rahman, Swapan K. Roy (2021), *Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms.* | To predict stress prevalence and identify risk factors of Bangladeshi university students. | 355 Bangladeshi university students. | Decision Tree, Random Forest, SVM, Logistic Regression, Boruta algorithm. | Random Forest achieved the highest accuracy (89.7%) with key features: pulse rate, blood pressure, sleep, smoking, academic background. |
| 27. | Digital Library - SpringerLink | Mustafa Yağcı. (2022). *Educational data mining: prediction of students’ academic performance using machine learning algorithms.* | Develop ML model to predict the final exam grades of undergraduate students, to identify the most effective ML algorithms for early prediction of student performance. | Academic achievement grades of 1.854 students that took Turkish Language-I course at a state university in Turkey in 2019-2020. Includes information such as a midterm exam, grades, department data, and faculty data. | ML Models (Random Forest, Nearest Neighbour, SVM, Logistic Regression, Naive Bayes, KNN) | The aforementioned ML model achieved a classification accuracy of 70-75% accuracy. It concludes that the midterm grades, department and faculty data are helpful parameters for predicting final exam performance. |
| 28. | Research Article - Research Square | Ebem, D. U., Ikegwu, A. C., Ezugwu, C. J., Ikpo, C. V., & Ogbunude, F. O., (2024), *Machine Learning-Based Real-Time Feedback Assessment System for Student Performance Prediction in Tertiary Institution*. | Predict student performance in real-time to enhance learning and engagement. | Students’ attendance, personal information, and assessment history. | K-Nearest Neighbor (KNN). | KNN achieved 78% accuracy in predicting student engagement and performance. |
| 29. | Digital Library - IEEE Xplore (Institute of Electrical and Electronics Engineer) | Aya Nabil, Mohammed Seyam, & Ahmed Abou-Elfetouh. (2021). *Prediction of Students’ Academic Performance*  *Based on Courses’ Grades Using Deep*  *Neural Networks.* | Identify students at risk of failure early semester using EDM, to explore the efficiency of deep learning. | Public 4-year university, includes students’ academic data (grades in previous courses of the first academic year). | ML models (DNN, Decision Tree, Random Forest, Gradient Boosting, Logistic Regression, SVC, KNN) and resampling techniques for imbalanced dataset (SMOTE, ADASYN, ROS, SMOTE-ENN). | DNN model reached the highest accuracy of 89% in predicting students’ performance in a data structures course and identifying students at risk of failure early in the semester. |
| 30. | Open Access Repository - MDPI (Multidisciplinary Digital Publishing Institute) | Janka Kabathova & Martin Drlik. (2021). *Towards Predicting Student’s Dropout in University Courses Using Different Machine Learning Techniques.* | Explore the effectiveness of ML classifiers in predicting student dropout, identify suitable features with limited data available to teachers. | 4-academic years from an e-learning course, includes features that were available to teachers during the course run. | ML classifiers models (Decision Trees, Random Forest, SVM, Logistic Regression, KNN) using performance metrics. | Accuracy result varied between 77% and 93% on unseen data from the following academic year. It is important to carefully select features and performance metrics when working with limited educational data. |
| 31. | Research Article - Research Square | Kandula, N., & Kumar, R. (2023). *A Deep Dive into Academic Excellence: Using Deep Learning to Evaluate and Improve Engineering Students' Performance*. | Analyze and predicting student performance for placements, support, and dropout intervention. | 60,000 students from a university, including academic and extracurricular data. | LSTM model with Adam optimizer and binary cross-entropy loss function, Logistic Regression, Decision Tree, and Random Forest. | LSTM achieved 99.9% accuracy, making it effective in predicting student performance and supporting educational decision-making. |

# CHAPTER 3

# IMPLEMENTATION METHOD

## Thinking Framework

****

Picture 1 – Methodology Flowchart

## Research Methods

This study follows systematic 6 processes to predict secondary students' academic success using machine learning. It begins by collecting the **Student Performance Dataset** and exploring its data to understand the key features influencing student grades. The process includes performing exploratory data analysis to uncover patterns and relationships within the data, followed by data preprocessing to ensure quality. Then, feature selection identifies the most important attributes. The dataset is split into training and testing sets. Each model is trained, evaluated, and compared based on various metrics. Lastly, evaluations and analysis for future research are made from the findings.

### Data Collection and Understanding

The dataset used in this study is titled "**Student Performance Dataset**". This dataset is obtained from Kaggle, uploaded by Dev Ansodariya who is a student at San Jose State University and an experienced software engineer. The dataset contains information on student achievement in secondary education from two Portuguese schools- Gabriel Pereira and Mousinho da Silveira. It includes student grades, demographic, social, behavioral, and school-related attributes, collected through school reports and questionnaires (Cortez & Silva, 2008).

### Exploratory Data Analysis

Exploratory Data Analysis (EDA) is conducted to explore the relationships between features, summary statistics, and detect anomalies which contain 395 rows and 33 columns. This dataset includes demographic, social, behavioral, school-related, and academic performance data for two subjects: Mathematics and Portuguese. The following is a description of each variable:

| No. | Variable | Description |
| --- | --- | --- |
| 1. | school | Student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira). |
| 2. | sex | Student's sex (binary: 'F' - female or 'M' - male). |
| 3. | age | Student's age (numeric: from 15 to 22). |
| 4. | address | Student's home address type (binary: 'U' - urban or 'R' - rural). |
| 5. | famsize | Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3). |
| 6. | Pstatus | Parent's cohabitation status (binary: 'T' - living together or 'A' - apart). |
| 7. | Medu | Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education). |
| 8. | Fedu | Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education). |
| 9. | Mjob | Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other'). |
| 10. | Fjob | Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other'). |
| 11. | reason | Reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other'). |
| 12. | guardian | Student's guardian (nominal: 'mother', 'father' or 'other'). |
| 13. | traveltime | Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour). |
| 14. | studytime | Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours). |
| 15. | failures | Number of past class failures (numeric: n if 1<=n<3, else 4). |
| 16. | schoolsup | Extra educational support (binary: yes or no). |
| 17. | famsup | Family educational support (binary: yes or no). |
| 18. | paid | Extra paid classes within the course subject (Math or Portuguese) (binary: yes or no). |
| 19. | activities | Extra-curricular activities (binary: yes or no). |
| 20. | nursery | Attended nursery school (binary: yes or no). |
| 21. | higher | Wants to take higher education (binary: yes or no). |
| 22. | internet | Internet access at home (binary: yes or no). |
| 23. | romantic | In a romantic relationship (binary: yes or no). |
| 24. | famrel | Quality of family relationships (numeric: from 1 - very bad to 5 - excellent). |
| 25. | freetime | Free time after school (numeric: from 1 - very low to 5 - very high). |
| 26. | goout | Going out with friends (numeric: from 1 - very low to 5 - very high). |
| 27. | Dalc | Workday alcohol consumption (numeric: from 1 - very low to 5 - very high). |
| 28. | Walc | Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high). |
| 29. | health | Current health status (numeric: from 1 - very bad to 5 - very good). |
| 30. | absences | Number of school absences (numeric: from 0 to 93). |
| 31. | G1 | First period grade (numeric: from 0 to 20). |
| 32. | G2 | Second period grade (numeric: from 0 to 20). |
| 33. | G3 | Final grade (numeric: from 0 to 20, output target). |

Table 1 – Data Description

One of the goals is finding the strong correlation between the final year grade (G3) and besides the earlier period grades (G1 and G2), as G1 and G2 have a big influence on the final grade. Predicting G3 without G1 and G2 is more challenging, but it is more valuable for real-world applications. Additionally, EDA highlights how non-academic factors like family background and social support influence student academic performance.

With EDA, it helps researchers to understand data distributions and detect any anomalies or outliers. It can also build summary statistics and visualizations. Visualizations such as correlation heatmaps are used to assess patterns and relationships among the variables. This thorough exploration sets the foundation for further analysis and helps identify the most relevant features for predictive modeling.

### Data Preprocessing

Data preprocessing involves cleaning the dataset, handling missing values, and removing outliers to prepare the data for model training. This ensures that the data is in a suitable format for machine learning algorithms.

### Scaling and Feature Selection

Feature selection is conducted to identify the most critical features that influence students' final grades (G3), enhancing the model's predictive accuracy by focusing on the most relevant data. In this study, numerical data were standardized using the StandardScaler from sklearn, ensuring all features are on the same scale and minimizing bias in models sensitive to magnitude differences, such as Linear Regression and K-Nearest Neighbors. Feature selection was performed using SelectKBest with the f\_regression scoring method, revealing G2 as the most influential feature with a score of 4554.71, followed by G1 with a score of 1297.22. Failures had a lower score of 30.87, indicating a weaker but still relevant impact. By concentrating on these selected features, the model's complexity is reduced, improving both interpretability and predictive performance.

### Model Training

Each selected machine learning model is trained using the training dataset to learn the relationships between the input features and the target variable (final grade) to make predictions.

### Evaluation and Analysis

The trained models are evaluated on the testing set using metrics to assess prediction accuracy and reliability. The performance of the models is then compared to identify the most effective algorithm for predicting students' academic success. Based on the analysis, conclusions are drawn regarding the best-performing model, and recommendations for future research are provided, including such as expanding the dataset, adding features, and suggesting practical applications to enhance the educational environment.

### 

# REFERENCES

C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," \*IEEE Transactions on Systems, Man, and Cybernetics\*, 2020. Available: <https://www.researchgate.net/publication/224160756_Educational_Data_Mining_A_Review_of_the_State_of_the_Art>

D. Ansodariya and P. Pathak, "Exploring the Relationship between Students' Engagement and Self-Regulated Learning: A Case Study using OULAD Dataset and Machine Learning Techniques," ResearchGate, 2023. Available: <https://www.researchgate.net/publication/377358413_Exploring_the_Relationship_between_Students_Engagement_and_Self-Regulated_Learning_A_Case_Study_using_OULAD_Dataset_and_Machine_Learning_Techniques>

G. James, D. Witten, T. Hastie, and R. Tibshirani, \*An Introduction to Statistical Learning with Applications in R\*, Springer, 2021. Available: <https://www.springer.com/gp/book/9781071614174>

"Introduction to Machine Learning." Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012>

J. Han, J. Pei, and M. Kamber, \*Data Mining: Concepts and Techniques\*, Elsevier, 2021.

M. Suzan, M. Rahman, and F. Hasan, "Student Adaptability Level in Online Education Using Machine Learning Approaches," ResearchGate, 2021. Available: <https://www.researchgate.net/publication/355891881_Students'_Adaptability_Level_Prediction_in_Online_Education_using_Machine_Learning_Approaches>

M. Wu, G. Subramaniam, D. Zhu, C. Li, H. Ding, and Y. Zhang, "Using Machine Learning-based Algorithms to Predict Academic Performance - A Systematic Literature Review," \*IEEE Xplore\*, 2024. Available: <https://ieeexplore.ieee.org/document/10563566>

P. Cortez and A. Silva, "Using Data Mining to Predict Secondary School Student Performance," in \*Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)\*, Porto, Portugal, 2008.

S. Pal, "Mining Educational Data to Reduce Dropout Rates of Engineering Students," \*International Journal of Information Engineering and Electronic Business\*, vol. 4, no. 2, pp. 1-7, 2012.

Z. Ibrahim, N. Omar, and M. Aziz, "Student Performance Prediction Using Machine Learning Techniques," \*International Journal of Advanced Computer Science and Applications\*, vol. 10, no. 8, pp. 1-5, 2019.