



**Early detection of academic weakness
in students using artificial intelligence tools**

Prepared By

Hissa Naif Al-Otaibi

252001788

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

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr Yasser Abdel Hamid

College of Informatics, Midocean University, Comoros

[2025]

Aabstract

Communities are concerned with the academic education of their students and its development, which helps equip students with the necessary skills and knowledge throughout their academic journey. They also face several challenges, the most significant of which is student academic failure, and they strive to find the necessary solutions. However, we can identify this problem early on and intervene promptly so that students can improve their academic performance. This is what we focused on in this study: enabling teachers to identify students at risk of academic failure and intervene early to address this issue before final exams. We achieved this by analyzing and interpreting the factors influencing students and predicting their failure or success early on.

In the study, we used a dataset consisting of 13 features and built an artificial intelligence model that predicts the “Categories” column, which consists of 5 classifications for students indicating excellent, good, and poor grades. The most prominent features affecting students’ final results, as identified through the correlation matrix, were midterm exams and assignments, while parental education showed a weak influence on student performance. Through the model, columns that negatively affected model performance were removed, missing data was handled, and text data was converted to numerical values.

We used traditional machine learning methods such as decision trees, K-Nearest Neighbors, and XGBoost. KNN and XGBoost yielded excellent results and outperformed the other models, while the decision tree performed less well because it requires further refinement. Based on these results, teachers can identify students at risk of failing and then take appropriate action to help them overcome the problem.

Thus, we can conclude that artificial intelligence plays an important and effective role in the educational process and its improvement, serving as a tool to assist both teachers and students in achieving desired educational goals and resolving issues related to academic underperformance.

List of Publications

Table of Contents

List of Abbreviations خطأ! الإشارة المرجعية غير معروفة.

List of Figures..... خطأ! الإشارة المرجعية غير معروفة.

Chapter 1: Introduction..... 7

1.1 Problem Definition 7

1.2 Thesis Objectives8

1.3 Thesis Contribution 8

1.4 Thesis Organization..... 8

Chapter 2: Background and Related Work..... 10

2.1 Introduction 10

2.2 Classical Machine Learning Methods 10

2.3 Related Work 14

2.4 Summary 16

2.5 Quantitative Results Summary 17

2.6 Research Gaps and Opportunities18

Chapter 3: Proposed Approach 20

3.1 Data Collection and Description 20

3.2 Data Preprocessing 20

3.3 Feature Engineering21

3.4 Feature Selection23

3.5 Model Selection and Optimization 23

3.6 Model Evaluation 23

3.7 Visualization and Insights 25

Chapter 4: Experimental Results and Discussion	27
4.1 Experimental Setup	27
4.2 Baseline Model Performance	27
4.3 Feature Selection and Model Refinement	28
4.4 The Relationship Between Quiz_1 and Final Exam.....	28
4.5 Analysis of Final Exam Scores.....	29
4.5 Discussion.....	29
Chapter 5: Conclusion and Future Work	31
5.1 Conclusion	31
5.2 Future Work	31
References	32

List of Abbreviations

- AI: Artificial Intelligence
- CNN: Convolutional Neural Network
- GPA: Grade Point Average
- XGBoost: Extreme Gradient Boosting
- DT: Decision Tree
- GBM: Gradient Boosting Machine
- GridSearchCV: Grid Search with Cross-Validation
- IQR: Interquartile Range
- LSTM: Long Short-Term Memory
- LIME: Local Interpretable Model-agnostic Explanations
- ELM: Extreme Learning Machine
- AutoML: Automated Machine Learning
- RNN: Recurrent Neural Network
- SVM: Support Vector Machine
- DNN: Deep Neural Network
- IQR: Interquartile Range
- LightGBM: Light Gradient Boosting Machine
- ML: Machine Learning
- MLP: Multi-Layer Perceptron
- PR: Precision-Recall
- RF: Random Forest
- RFE: Recursive Feature Elimination
- RMSE: Root Mean Square Error
- SHAP: SHapley Additive exPlanations
- SVM: Support Vector Machine
- XAI: Explainable Artificial Intelligence
- XGBoost: Extreme Gradient Boosting
- KNN: K-Nearest Neighbors

List of Figures

Figure 1 Explore the Data.....	20
Figure 2 Box Plot.....	21
Figure 3 Correlation Matrix.....	22
Figure 4 Comparison of Models-Phase 1.....	24
Figure 5 Comparison of Models-Phase 2.....	24
Figure 6 SHAP diagram.....	26
Figure 7 Scatter Plot.....	28
Figure 8 Distribution of Final Exam Mars	29

Chapter 1 :Introduction

1.1 Problem Definition

The Kingdom of Saudi Arabia places great emphasis on education and its importance, and strives to improve the educational process and provide a suitable environment for students. Students are considered a key component of the educational process and must be supported; any issues that hinder their learning must be addressed. Among the most prominent of these issues are poor academic achievement or academic delays. Some researchers have defined academic achievement as

A student's academic achievement is based on their marks throughout the academic year and is measured as a percentage; this tells us whether the student has passed or failed. Furthermore, a student's success does not depend solely on marks; it is also influenced by behavioural, social and psychological factors, as well as intrinsic motivation, resources and the school environment, all of which contribute to academic achievement. (Wang et al. 2025)

The study by **Mensah, R. O. (2024)** I understood that academic failure means a student not achieving the minimum required marks in an exam.

Academic failure is usually due to social or psychological factors, or to a lack of basic study skills among students, such as motivation and time management. A study involving 4,660 students from 24 schools revealed that the strongest predictor of academic achievement is previous grades and extracurricular activities. Based on these findings, we emphasis the importance of early and effective assessment, building on the two previous points, in order to identify risks and intervene to improve academic performance.

A study by Laetitia Cassells (2012) indicated that students at risk of academic failure benefited more from the early detection system and additional lessons than students who were not subject to the system, achieving cumulative grades that were 8% higher.

1.2 Thesis Objectives

- Review research studies on the causes of poor academic performance among students, ways to address it, and methods for developing the abilities of high-achieving students
- Use and analyze student data
- Create a tool to help teachers identify students with low grades at an early stage and assist them in passing the course

1.3 Thesis Contribution

This project focuses on school students with the aim of helping to reduce academic failure at the end of the school year, assisting teachers in identifying students at risk of failing a particular subject, and identifying high-achieving students, thereby contributing to improving the quality of the educational process. This is achieved by creating a tool designed using machine learning models and real student data to provide interpretable results, while allowing for future scalability.

1.4 Thesis Organization

Chapter 1 : Introduction and Overview

The first chapter will provide an overview of the study topic, review the problem and objectives, explain how the study contributed to the field of education, and describe the organization of the thesis. reviews traditional methods of machine learning and advanced methods of deep learning. It also explains hybrid models that combine traditional methods with deep learning and reviews quantitative results, research gaps, and opportunities.

Chapter 2 : Background and Related Work

This chapter will provide a comprehensive background on previous studies in the field and identify current trends and their potential for future development.

Chapter 3 : Proposed Approach

This chapter reviews the proposed approach, the techniques and methods used to solve the problem, and the selected features.

Chapter4:Experimental Resultsand Discussion

This chapter presents the results of the models, compares their performance, provides analyses and identifies challenges ,and discusses the models' performance.

Chapter 5: Conclusion and Future Work

This chapter concludes the study with a summary of the main findings and challenges, as well as recommendations for improving the model and advancing the research.

Chapter 2: Background and Related Work

2.1 Introduction

Artificial intelligence models play an important role in education and improve and develop the educational process. Artificial intelligence models are divided into several sections, which we will now review.

2.2 Classical Machine Learning Methods

I have read studies that use the method and show that the method works well for predictions. However the method shows problems when the method has to handle a lot of data. The problem has caused modern deep learning models to appear. The modern deep learning models help schools and big institutions. Examples of models include:

- **K Nearest Neighbour (KNN)**

Definition: A supervised learning algorithm used for classification or prediction based on the proximity of points. It operates on the principle that a test point belongs to the nearest cluster in the case of classification, or to the predicted value in the case of regression. (fernandez-Delgado,et al,2020)

How it works:

1. The distance between the points and the test point we want to classify is calculated.
2. k nearest points to the test point are selected
3. If the goal is classification, the test point belongs to the majority among the nearest points; if the goal is regression, the average of these points is calculated

Advantages:

- A simple and easy-to-understand algorithm
- Learns directly from the data and does not require a prior model for training

Disadvantages:

- Sensitive to noise
- Slow with very large datasets

• **The Random Forest** It is a supervised learning algorithm that uses multiple decision trees to build a robust model. Decision trees are randomly constructed using training data, and the results are then combined to produce the final prediction. (fernandez-Delgado, et al, 2020)

How it works:

1. Decision trees are built using random samples of the data
2. Each tree is trained on a portion of the data and takes a random sample of features at each split point within the tree
3. The results are aggregated to produce the final prediction

Advantages

- Resistant to overfitting
- Prediction speed and accuracy
- Used for regression and classification

Disadvantages

- Trees may not be easily interpretable
- Consumes a lot of time and resources with large datasets

• **Decision Tree** It is a supervised learning algorithm used for classification or regression, which partitions data spaces based on specific features. (de Ville, 2019)

How it works:

1. The data is partitioned according to features, such that the feature most important in determining the outcome is repeatedly selected.
2. The data is repeatedly split until the final node in the tree is reached, at which point a decision is made.
3. When splitting the data, the split is based on the split measure or information gain.

Advantages

- Ease of interpretation
- Does not require data normalization

Disadvantages

- Susceptibility to overfitting if the tree is deep
 - Can be susceptible to noise
- **XGBoost** It is a learning-based algorithm that uses incremental reinforcement techniques and is optimized to be faster and more efficient. Incremental decision trees are constructed, and each tree is added incrementally to correct errors resulting from the previous trees. (Chen & Guestrin, 2016)

How it works:

1. It begins by creating an initial decision tree.
2. The error of this tree is calculated.
3. Another decision tree is built to correct the errors or residuals of the first one.
4. Prediction improves incrementally through the addition of new trees.
5. GXBoost performs regularization until the model is optimized and overfitting is minimized.

Advantages:

- Achieves good results in many competitions
- Powerful for handling large datasets and fast
- Supports parallel processing

Disadvantages:

- Considered more complex than other algorithms
 - Requires careful parameter tuning
 - XGBoost Official.
- **GridSearchCV** is a method for finding the best parameters for machine learning algorithms.

Parameters are tested, and the final model performance is evaluated based on prediction accuracy.(Geron , 2019)

How it works:

- Select a set of potential parameters.
- Each set of parameters is tested on a test dataset.
- Each set is evaluated using cross-validation.
- The best parameters, which yielded the highest accuracy, are selected.
- The final model is trained using the best parameters.

Advantages:

- A comprehensive method for finding the best parameters.
- Works best with algorithms that have a limited number of parameters.

Disadvantages:

- Time- and resource-intensive
- Not suitable for algorithms with a large number of parameters
- “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”

In general, deep learning models have a high capacity to improve prediction accuracy and analyze complex patterns in educational data, making them effective for supporting decision-making and developing intelligent educational systems, especially in the context of predicting academic performance and early detection of at-risk students.

2.3 Related Work

Research has focused on using artificial intelligence and machine learning algorithms to predict student outcomes in educational settings. The goal is to improve understanding of student academic performance, develop systems for early detection of students at risk of academic failure or dropout, and implement appropriate preventive measures. Among the most prominent algorithms used in this field are artificial neural networks (ANNs), semi-supervised learning, random forests, and probabilistic learning. Research and studies on this topic have been diverse.

A group of these studies focused on predicting student performance and early detection of at-risk students using traditional artificial intelligence techniques. Studies such as **Khan et al. (2021)** collected data from the University of Amman over three semesters, covering 151 cases and ten predictive characteristics of student performance. The study identified the most important factors affecting performance as GPA, exam scores, and attendance, and determined that students with a GPA of 2.75 or higher showed better results and performance than those who obtained averages below 2.75. The study used several models, the most accurate and best in performance being the decision tree. There was also a study by **Lau et al. (2019)** that included 1,000 cases from a Chinese university and three university programs over four years and used 11 variables. The results improved over the years, and the study used the ANN model, which gave results with an accuracy of 84.8%. We also have a study by **Sukhbaatar et al. (2019)**. Data was collected from Kumamoto University as part of a compulsory course for second-year students. The number of students reached 1,110 over six semesters from 2012 to 2017. A multi-layer neural network was used, and the model identified 25% of students at risk of failure after the first short test and after the second test. The midterm test results improved, and 65% of students at risk of failure were identified due to increased data and, consequently, increased model accuracy.

This group focused on students' previous grades, attendance, and tests. One of the most notable findings of this research is that neural networks and decision trees outperformed other models, with prediction accuracy ranging from 86% to 91%, enabling educational institutions to provide remedial plans for struggling students, such as extra lessons and early guidance.

In contrast, another group of studies used probabilistic and advanced models to achieve more accurate early predictions, such as the studies by **Nimy et al. (2023)** used academic data from students at a university in South Africa, such as tests and assignments, in addition to their interaction on the platform, such as the amount of time they spent on the platform and demographic data. The researchers used probabilistic machine learning and AutoML techniques to measure the probability of academic failure at each stage of the semester, rather than just giving a final result. The results showed that these models supported the reliability of early prediction, with an accuracy of 92%.

A bunch of studies also used advanced techniques like recurrent neural networks (RNN) and semi-supervised learning, like in the studies by **Mondal and Mukherjee (2018)** and **Melo and Souza (2023)**, which used data from an educational institution in Brazil for one year, including 5,915 dropouts and 2,096 graduates. Researchers **Mondal, A. & Mukherjee** data from Kaggle, where RNN was successful in extracting temporal patterns, and these models were used to analyze educational data. The models in this group showed great effectiveness in tracking changes in student performance over time and early prediction of academic failure, proving their usefulness in long-term early intervention programs.

We have the study by **Ujkani et al.(2024)** which used explainable artificial intelligence (XAI) techniques to identify students at risk of failure in online learning environments. The study utilized data from 32,593 students and employed SHAP and LIME to understand and interpret the impact of various features on predictions, to improve transparency and accuracy, and to demonstrate how different factors influence model decisions. Among the most notable findings was that interaction with the e-learning environment and early registration dates are significant factors in academic performance. The models used in the study were Random Forest, Gradient Boosting Classifier, K-NNs, Neural Network, LSTM, XGBoost, and CNN, and they demonstrated high predictive ability ranging from 80% to 94%.

As for the study by **George et al. (2025)** the same XAI techniques as the previous study were applied to a dataset collected from several educational institutions. The study sought to use these techniques to understand and explain the factors influencing student failure in order to improve the prediction of student grades and enhance academic success. The models used were Random Forest and XGBoost, and they yielded good results

The latter group, such as the studies **by Deo et al. (2020)**, which relied on six years of student performance data from the University of Southern Queensland, Australia, included several in-person and distance learning courses, and the ELM model outperformed the other models in predicting student grades and identifying those at risk of failure with a very low error rate. It was faster in terms of computational efficiency and pattern capture in continuous assessment data and helps the model improve the quality of teaching and early support. For the study **by Zeineddine et al. (2021)** used data from 461 university students for a specific period between 2016 and 2021. It focused on developing high-performance models using AutoML to predict student grades in specific courses and used student data available before the start of the academic program. It showed high prediction accuracy with low errors, outperforming traditional models such as random forests or traditional neural networks, with an accuracy of 83%.

2.4 Summary

The literature supports the use of artificial intelligence and shows its success in identifying students at risk of failing with high accuracy using different models. Based on these results, we find that the results of the models differ between studies, but the most accurate was XGBoost with 92%, while the most commonly used models are recurrent neural networks (RNN) and Random Forest. Therefore, the appropriate models for the study will be selected and trained on real data after processing and improving feature selection to give higher quality and credibility results, thus contributing academically and scientifically in this field.

2.5 Quantitative Results Summary

References	Model-Used)	Data Set	Results
Khan, I., Ahmed, M., & Ali, S. (2021).	• k-Nearest Neighbors (K-NN) • Multilayer Perceptron (ANN) • Naïve Bayes • Decision Tree (RepTree)	collected from the University of Amman over three semesters, covering 151 cases and ten predictive features of student performance	MCC (Matthews Correlation Coefficient) (K-NN)=0.97 (ANN)=0.98 Naïve Bayes=0.99 Decision Tree=0.97
Deo, R. C., Yaseen, Z. M., Al-Ansari, N., Nguyen-Huy, T., Langlands, T. A. M., & Galligan, L. (2020)	• ELM model • Random Forest (RF) • Volterra model	Six years of student performance data from the University of Southern Queensland, Australia Included several in-person and distance learning courses	prediction error ELM model :0.74% Random Forest (RF) :3.12% Volterra model :1.06%
Ujkani, B., Minkovska, D., & Hinov, N. (2024)	• Random forest • Gradient Boosting Classifier • K-NNs • Neural Network • LSTM • XGBoost • CNN	The data set from the Open University includes data on 32,593 students over a 9-month academic period from 2014 to 2015	Accuracy • First custom Neural Network Model = 94% • RF=86% • XGBoost = 86% • k-NN = 82% • Multi-layer Perceptron=85% • CNN=86%
Zeineddine, H., Braendle, U., & Farah, A. (2021)	• SVM • K-NN • Random Forest • Gradient Boosting	The study used university student data comprising 461 cases for a specific period between 2016 and 2021	Relative Prediction Error • SVM: 3.5% • K-NN:8.0% • Random Forest:6.0% • Gradient Boosting:4.5%
Nimy, E., Mosia, M., & Chibaya, C. (2023)	• Logistic Regression • Support Vector Machine (SVM) • Decision Tree • Random Forest • Naive Bayes	A university in South Africa used student academic data such as tests and assignments, as well as their interaction on the platform, such as the amount of time they spent on the platform, and demographic data.	• Logistic Regression:11.6% • (SVM):10.5% • Decision Tree:12.7% • Random Forest:12% • Naive Bayes :7.19%

Lau, E. T., Sun, L., & Yang, Q. (2019)	ANN neural networks	The study included 1,000 cases from a Chinese university and three university programs over four years and used 11 variables	ANN neural networks: 84.8
Mondal, A. & Mukherjee, J. (2018)	- Recurrent Neural Network (RNN) - Artificial Neural Network (ANN) - Deep Neural Network (DNN)	Data from Kaggle was used	prediction error (RNN):85% (ANN):82% (DNN):83%
Embarak, O. H., & Hawarna, S. (2023).	- Artificial Neural Networks (ANN) - Random Forest - XGBoost algorithm - Support Vector Machine (SVM)	The study collected data from 450 university students from different universities. The data included academic, behavioral, and psychological data	Accuracy - ANN:82.22% - Random Forest:84.44% - XGBoost algorithm:87.78% (SVM):85%
Melo, E. C., & Souza, F. S. H. (2023).	- Semi-Supervised Learning - Logistic Regression - K-Nearest Neighbors (KNN) - Decision Tree - XGBoost	From an educational institution in Brazil for one year, including 5,915 cases of dropouts 2,096 cases of graduates	Accuracy - Semi-Supervised Learning:90% - Logistic Regression:85.5% (KNN):87.68% - Decision Tree:88.05% - XGBoost 92.05%
Ben George ,E., Senthilkumar,R., Al-Junaibi, F., & Al-Shuaibi, Z. (2025)	- Random Forest - XGBoost algorithm	- Data from 500 students collected from several educational institutions was used, an 10 features were utilized.	- Random Forest=85% - XGBoost algorithm=84%

2.6 Research Gaps and Opportunities

Connections:

- The studies relied on real student data.
- The studies were conducted within a specific scope within a single educational institution.
- The researchers used more than one algorithm in the studies and selected the best one.
- The models showed remarkable accuracy in the results with a negligible error rate.

Research gap:

- The need to expand the scope of the study so as not to affect the generalizability of the results.
- Adding the largest possible segment of students to contribute to increasing the efficiency and accuracy of the model.
- Adding multiple features, different educational materials, and varying educational levels.
- Need to focus on understanding the model's recommendations to help teachers make decisions
- Need to study the practical application of prediction to guide real educational interventions, determine the effectiveness of the model, and measure its impact on students

Chapter 3 : Proposed Approach

The curriculum is based on rigorous data analysis and the optimization of machine learning models to obtain realistic and accurate results, which are used to identify at an early stage students at risk of academic failure.

3.1 Data Collection and Description

A dataset from Kaggle relating to students' academic performance was used. This dataset contains data on 520 students, detailing their marks in tests, assignments, projects and presentations, as well as other information such as gender, parents' educational attainment and final exam results. The data was categorized to analyze student performance and classify them into five groups that determine each student's level and predict areas of academic weakness.

	Student Id	gender	Quiz_1	Midterm_1	Assignment_1	Assignment_2	Assignment_3	Project	Presentation	Final_Exam_Marks	parental level of education	Total	Grade	Categories
0	1	male	2.34	6.78	4.56	3.24	4.56	5.67	7.89	24	some college	59.04	C+	Pass
1	2	female	8.17	12.43	2.87	1.87	2.89	2.34	2.34	29	some high school	61.91	B-	Pass
2	3	female	5.92	3.21	0.45	4.56	0.76	8.9	5.67	30	associate's degree	59.47	C+	Pass
3	4	female	1.45	9.87	3.12	2.34	3.01	1.23	1.23	33	associate's degree	55.25	C+	Pass
4	5	female	9.78	5.34	0.98	0.98	1.98	9.45	9.01	31	bachelor's degree	68.52	B	Distinction

Figure 1 Explore the Data

3.2 Data Preprocessing

After collecting the data, I applied several data processing steps, such as

- **replacing commas with periods:** The dataset contains decimal commas that can only be processed after converting them to periods
- **Converting text columns to numbers:** I found some columns such as gender, parental level of education, Grade, and Categories, all of which contained text data that was converted to numerical data appropriate for their type
- **Handling missing data**

- **Missing values were replaced with the mean values**
- **Removing some columns:** To achieve accurate results, we had to remove certain columns such as Student Id because it provides no benefit to the predictions. The Grade and Total columns were also removed because they resulted in extremely high prediction rates, leading to unrealistic results; the model interpreted them as a reflection of the student's level, with accuracy rates reaching 100% for some models. This adjustment aims to prevent overfitting and ensure the data is balanced and well-suited for model training
- **Detecting outliers:** Outliers were detected and visualized in the dataset using the interquartile range and box plots. These values were handled using the Interquartile Range (IQR) to identify values outside acceptable limits, which were then removed to ensure the accuracy of the data used.

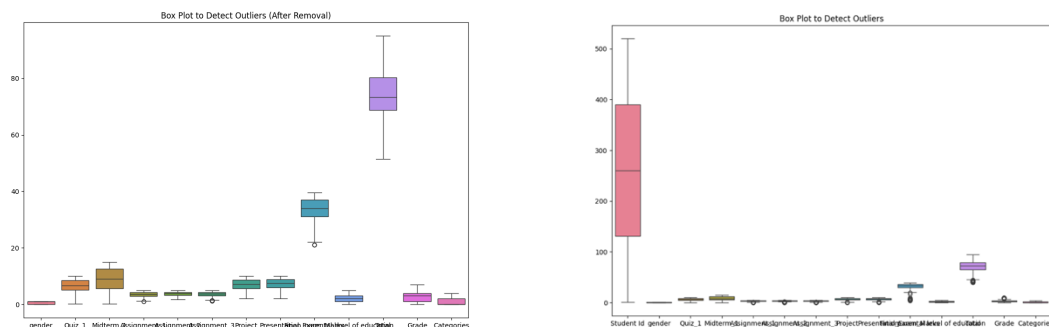


Figure 2 Box Plot

3.3 Feature Engineering

The features available for training were optimized based on the original data, such as extracting interactive features between different test scores.

A correlation matrix shows the relationship between variables in the data; its purpose is to measure the degree of correlation between the characteristics represented by the columns. To represent positive and negative correlations, the colors red and blue are used, where red indicates a strong positive correlation and blue indicates a negative correlation or no correlation.

Depending on the degree of correlation, the color becomes darker; the correlation coefficient ranges from +1 to -1.

+1 indicates a positive correlation, meaning that as the value of the first variable increases, the value of the second variable also increases.

-1 indicates a negative correlation, meaning that as the value of the first variable increases, the value of the second variable decreases, and vice versa. 0 means there is no correlation between the two variables.

From this matrix for our dataset, it is clear that there is a positive correlation (.44) between Quiz_1 and Midterm_1, meaning that if a student scores well on Quiz_1, they are more likely to score high on Midterm_1, and there is also a strong correlation (.70) between Final Exam and the total score, indicating that final exam scores account for the largest share of the total score. There is a positive correlation (0.68) between Midterm-1 and the total score, while weak and negative correlations were observed between Assignment

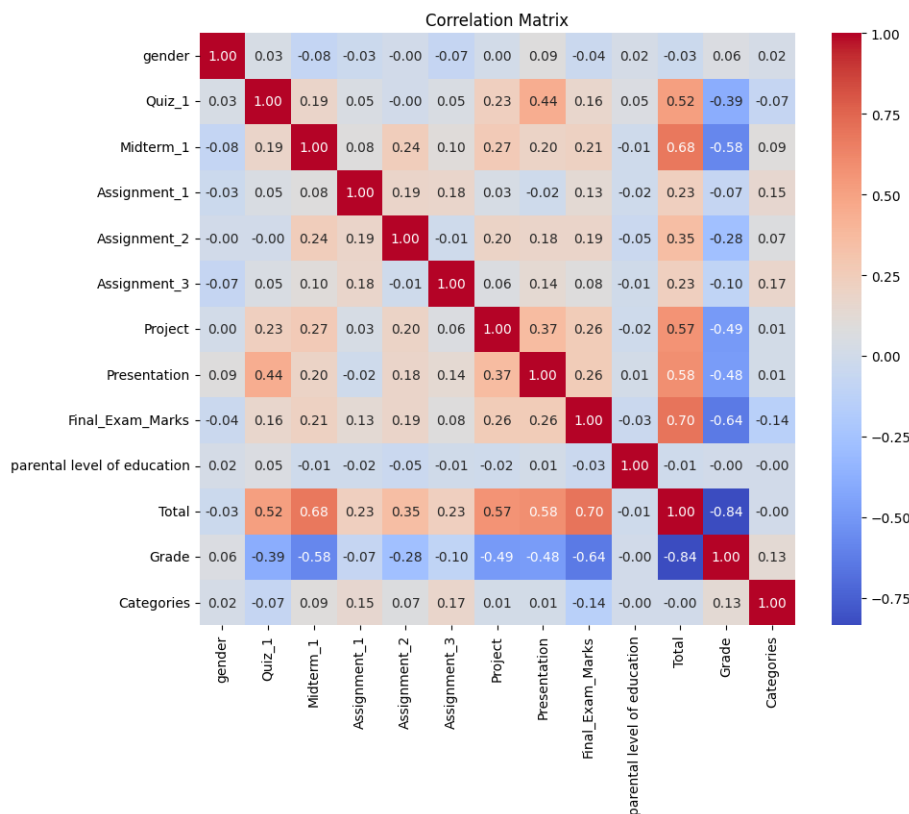


Figure 3 Correlation Matrix

3.4 Feature Selection

Features with a strong influence on the prediction of academic underachievement were selected by examining the previous correlation matrix, analyzing the relationships between features, choosing those most closely associated with academic underachievement, and removing features that cause overfitting—such as Grade and Total—to improve class distribution. The column that the models will predict is “Categories,” so that we have five classifications.

3.5 Model Selection and Optimization

The machine learning models used are suitable for the task of predicting academic underachievement, including KNN, Decision Tree, Random Forest, and XGBoost

They were optimized using GridSearchCV to obtain the best parameters, such as the number of trees in Random Forest and XGBoost and the tree depth in Decision Tree. We also used optimization techniques such as cross-validation to reduce overfitting and ensure that the models are capable of good generalization

3.6 Model Evaluation

The results were evaluated using various evaluation metrics, such as:

Accuracy: the proportion of correct predictions in the model .

Precision: the number of correct positive predictions relative to the total number of positive predictions made.

Recall: the number of correct positive predictions relative to the total number of actual positive instances.

F1 score: The harmonic mean of recall and precision, and an important measure of the balance between them.

The study presented results in two phases. The first phase included all features in the dataset, and the results were unrealistic and extremely high due to the presence of “Grade” and “Total,” with accuracy reaching 100%, indicating “overfitting.”

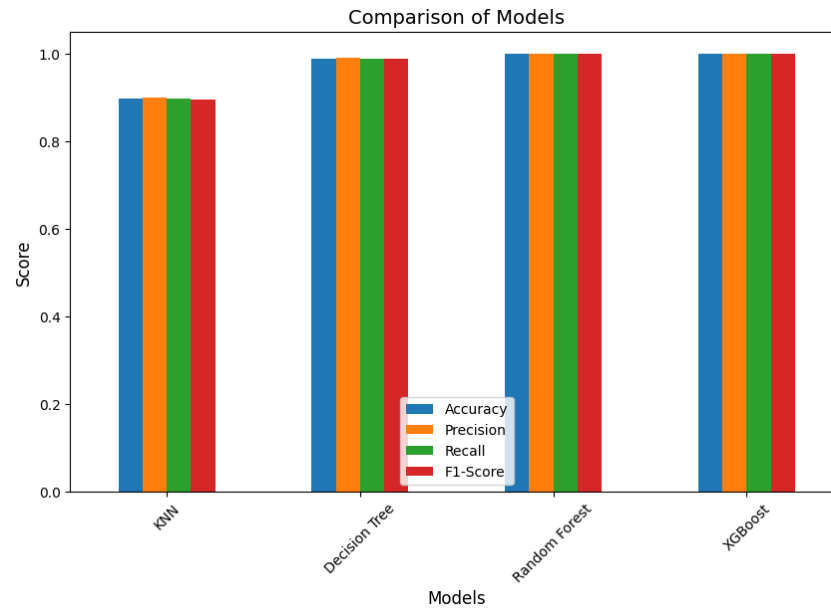


Figure 4 Comparison of Models-Phase 1

As for the second phase the features causing high accuracy were removed because the models treated them as redundant, and the results became more realistic

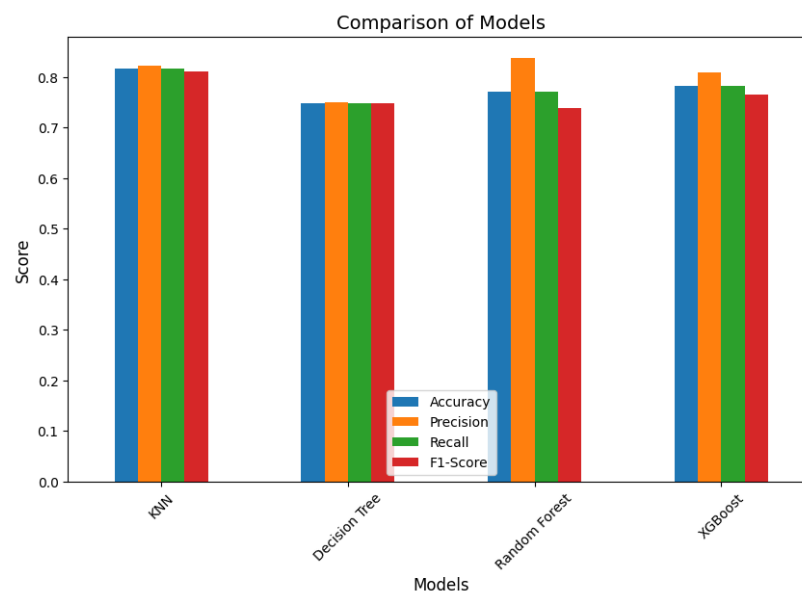
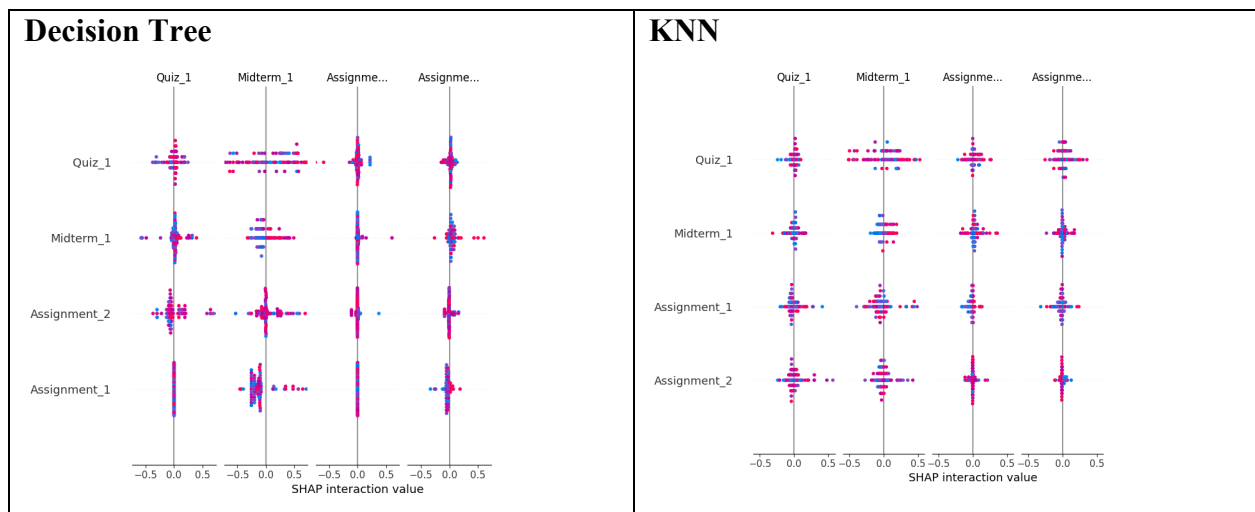


Figure 5 Comparison of Models-Phase 1

Here, the models' results varied and were more realistic. The best model in terms of overall performance is KNN, as precision, recall, and F1-score are more balanced in this model, while Random Forest also achieved high precision but failed to improve the F1 -Score, and thus did not outperform KNN in terms of balance, making KNN the best model when considering the balance between accuracy and recall. As for the XGBoost model, its performance was average compared to the rest of the models, and the decision tree model performed the worst among them and needs parameter tuning.

3.7 Visualization and Insights

After using SHAP diagrams, it becomes clear that certain features have a significant impact on the final decisions regarding models.



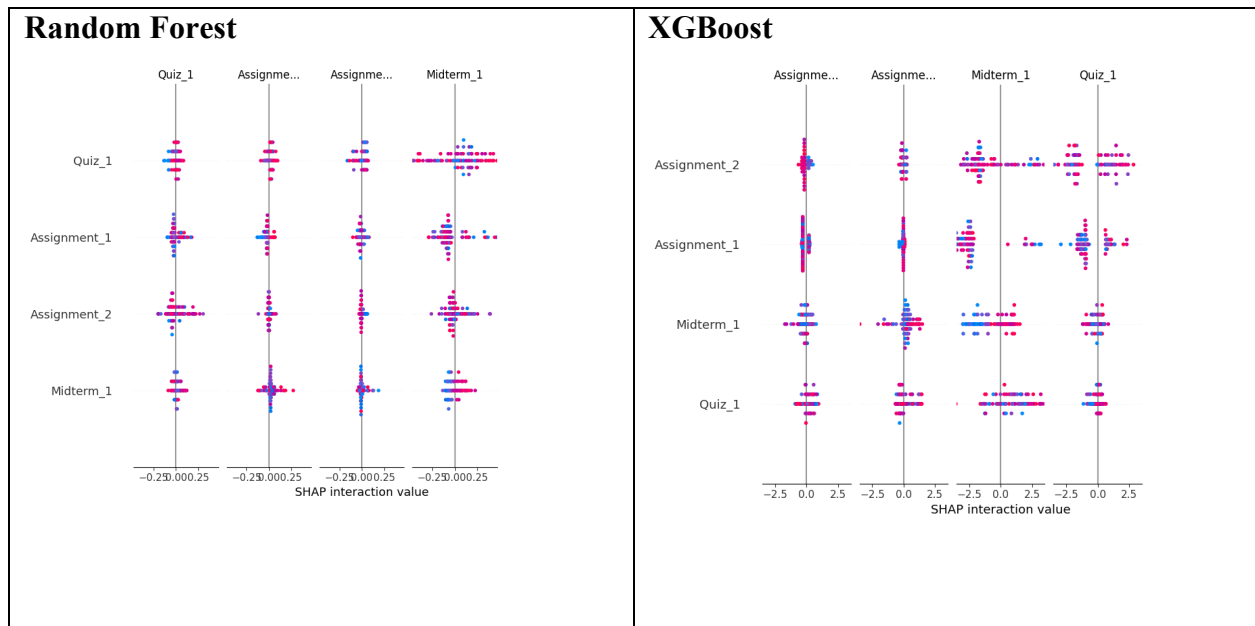


Figure 6 SHAP diagram

In the Decision Tree + KNN models, we see that the features Midterm 1 and Quiz_1 have a significant impact on the final result and play a key role in determining academic success. The SHAP values show strong interactions with these features, where red dots indicate a positive impact and blue dots indicate a negative impact. Meanwhile, the features Assignment_1 and Assignment_2 have a lesser impact. In the RF model, the feature with the greatest impact on the results was Quiz_1 compared to the other features, whereas in XGBoost, the features Assignment_1 and Assignment_2 had the greatest impact.

Chapter 4

Experimental Results and Discussion

4.1 Experimental Setup

All experiments were implemented in Python using several libraries:

- pandas
- numpy
- matplotlib.pyplot
- seaborn
- sklearn

The baseline dataset consists of 521 records. Outliers were handled, and text data was converted to numerical data. The dataset contained 14 columns; columns that caused overfitting issues were removed, and the data was split into 80% for training and 20% for testing

4.2 Baseline Model Performance

All features were used to train the models initially, and the results showed very high accuracy, indicating overfitting. The performance results were as follows:

(Model)	(Accuracy)	(Precision)	(Recall)	F1 (F1-Score)
KNN	0.90	0.90	0.90	0.89
Decision Tree	0.99	0.99	0.99	0.99
Random Forest	1.00	1.00	1.00	1.00
XGBoost	1.00	1.00	1.00	1.00

4.3 Feature Selection and Model Refinement

We removed columns that were irrelevant to the prediction process, such as

student ID

, and removed features that caused overfitting, such as

Total: which provides the total score in numerical form, reflecting the student's level or classification in Categories

Grade: converts the previous total into letter grades, which also indicate the student's level

Therefore, it was necessary to remove these features, and the results became more realistic. After adjusting them with GridSearchCV, the performance metrics appeared as follows

(Model)	(Accuracy)	(Precision)	(Recall)	F1 (F1-Score)
KNN	0.82	0.83	0.82	0.81
Decision Tree	0.75	0.75	0.75	0.75
Random Forest	0.77	0.84	0.77	0.74
XGBoost	0.78	0.81	0.78	0.76

5.4 The Relationship Between Quiz_1 and Final Exams The scores indicate that the correlation between them is very weak; it is not necessary for a student to receive a high score on the midterm exam in order to receive a high score on the final exam.

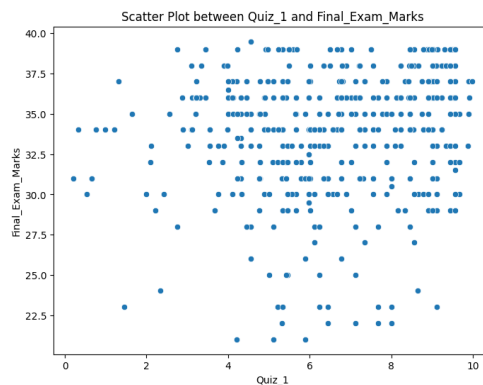


Figure 7 Scatter Plot

4.5 Analysis of Final Exam Scores

We note that most students achieved good results and demonstrated a high level of performance, while a small number received low scores; most scores fell between 35 and 37.5

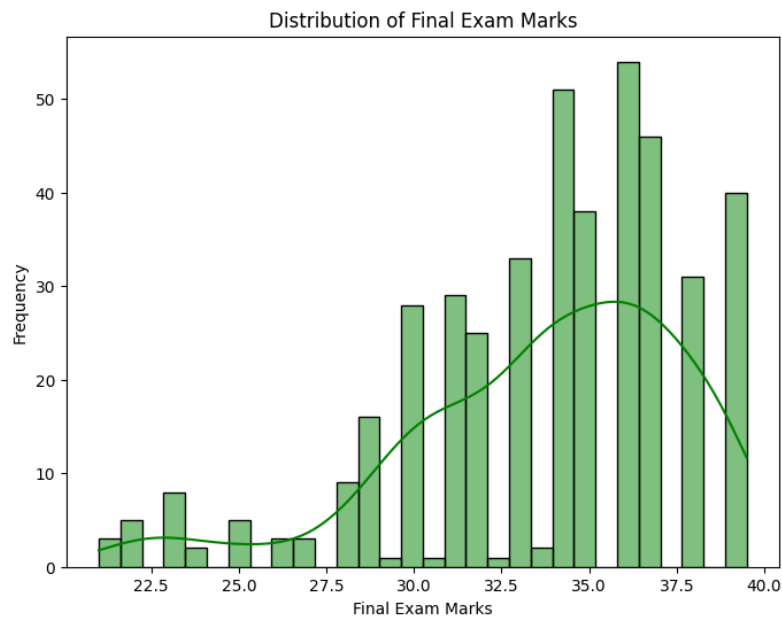


Figure 8 Distribution of Final Exam Marks

4.6 Discussion

The use of artificial intelligence in predicting student performance is important for improving their final results and identifying underperforming students early on so that teachers can intervene and address their weaknesses. The models yielded good results after a careful selection of appropriate features. The most significant findings were as follows:

- A student's performance on projects and presentations has a strong correlation with test scores
- Students who performed well on the midterm exam achieved high scores on the final exam

- Simplifying the model improved results by removing several features and focusing on those most influential on student performance
- The KNN model showed more balanced and stable results and achieved the highest accuracy value, reaching 0.82
- The RF model achieved the highest Precision value, meaning it is best at reducing false positives
- The model with the lowest performance is DT compared to the other models

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Early prediction of student performance is one of the most important factors in addressing academic underachievement, and artificial intelligence has proven its effectiveness and accuracy in identifying this group, while the models demonstrated varying degrees of accuracy, some of which were quite good. It became clear to us from the data set that students' performance on assignments and projects plays a significant role in improving their grades, as students who are engaged in their work perform well on exams, unlike those who are not. Furthermore, data analysis revealed that students who scored high on the midterm exam were also able to perform well on the final exam, which highlights the importance of closely monitoring students before final exams and ensuring they complete all assignments correctly. Therefore, it is important to incorporate artificial intelligence into education and use it as a tool to assist both teachers and students in achieving good results, reducing academic underperformance among students, and reducing the teacher's workload.

5.2 Future Work

- Use of deep learning techniques: We can expand the scope of the curriculum to incorporate deep learning techniques, such as neural networks, to improve predictions
- Adding new features: We can introduce a set of new features to improve predictions, such as student online behavior, previous years' grades, or interaction with educational platforms
- Model optimization: We can optimize the models using other techniques such as Gradient Boosting and LightGBM to achieve better performance.
- The number of categories can be reduced from five to just two to improve the models' performance.

References

1. A. Casillas, S. B. Robbins, J. Allen, Y.-L. Kuo, M. A. Hanson, and C. Schmeiser, "Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior," *Journal of Educational Psychology*, vol. 104, no. 2, pp. 407-420, 2012, doi: 10.1037/a0027180.
2. L. Cassells, "The effectiveness of early identification of 'at risk' students in higher education institutions," University of Pretoria, 2012
3. Hadi Shaalan Rabie and Ismail Muhammad al-Ghoul, *The Educational Counselor and His Active Role in Solving Student Problems*, Jordan, Dar Al-Thaqafa, 2006
4. S. M. Lange and B. Thompson, "Early identification and interventions for children at risk for academic failure," *International Journal of Special Education*, vol. 21, no. 3, pp. 108–117, 2006.
5. A. Rashid, R. Yasmeen, R. Ahmed, and K. Jawed, "Factors leading to the academic failure of undergraduate medical students
6. Khan, I., et al., An artificial intelligence approach to monitor student performance and devise preventive measures. *Smart Learning Environments*, 2021. 8(1): p. 1-18.
7. Deo, R.C., et al., Modern artificial intelligence model development for undergraduate student performance prediction: An investigation on engineering mathematics courses. *IEEE Access*, 2020.
8. Zeineddine, H., U. Braendle, and A. Farah, Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, 2021. 89: p. 106903
9. Hashim, A.S., W.A. Awadh, and A.K. Hamoud. Student performance prediction model based on supervised machine learning algorithms. in *IOP conference series: materials science and engineering*. 2020. IOP Publishing.
10. Lau, E.T., L. Sun, and Q. Yang, Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, 2019.
11. Pallathadka, H., et al., Classification and prediction of student performance data using various
12. machine learning algorithms. *Materials today: proceedings*, 2023.
- Sukhbaatar, O., T. Usagawa, and L. Choimaa, An artificial neural network based early prediction of failure-prone students in blended learning course. *International Journal of Emerging Technologies*

in Learning (iJET), 2019. 14(19)

13. Burkov, Andriy. The Hundred-Page Machine Learning Book. Andriy Burkov, 2019. ISBN: 978-1999579500.

14. Melo, E. C., & Souza, F. S. H. (2023). Improving the prediction of school dropout with the support of the semi-supervised learning approach. iSys: Brazilian Journal of Information Systems, 16(1). <https://doi.org/10.5753/isys.2023.2852>

15. Mondal, A. and J. Mukherjee, An Approach to predict a student's academic performance using Recurrent Neural Network (RNN). Int. J. Comput. Appl, 2018

16. Nimy, E., Mosia, M., & Chibaya, C. (2023). Identifying At-Risk Students for Early Intervention — A Probabilistic Machine Learning Approach. Applied Sciences, 13(6), 3869. <https://doi.org/10.3390/app13063869>

17. Ujkani, B., Minkovska, D., & Hinov, N. (2024). Course Success Prediction and Early Identification of At-Risk Students Using Explainable Artificial Intelligence. Journal/Conference, Volume(Issue)

18. Ben George, E., Senthilkumar, R., Al-Junaibi, F., & Al-Shuaibi, Z. (2025). *Explainable AI methods for predicting student grades and improving academic success*. Journal of Information Systems Engineering and Management, 10(23s), 117–126. <https://doi.org/10.52783/jisem.v10i23s.3680>.

19. Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2020). A survey of Random Forests. *Journal of Machine Learning Research*, 21, 1-37. <https://www.jmlr.org/papers/volume21/20-016/20-016.pdf>