



Early detection of academic weakness in secondary and middle school students using artificial intelligence tools

Hissa Naif Al-Otaibi

252001788

Master's student in Artificial Intelligence

MIDOCEAN UNIVERSITY

2025

Chapter 1 :Introduction

1.1 Problem Definition

The Kingdom of Saudi Arabia is keen on education and its importance, and seeks to develop the educational process and provide a suitable environment for students. Students are considered an important focus in the educational process and must be supported, and any problems that hinder their learning must be resolved. One of the most prominent problems is poor or delayed academic achievement. Some researchers have defined academic achievement as

“all performance by students in various school subjects that can be measured by test scores, teacher assessments, or both.” [Hadi Shaalan Rabie and Ismail Muhammad al-Ghoul, *The Educational Counselor and His Active Role in Solving Student Problems*, Jordan, Dar Al-Thaqafa, 2006]

Academic weakness is a decline in a student's academic performance and can be caused by various social or psychological factors. It is necessary to address and highlight this problem, identify it early, and treat it so that the student can catch up with the rest of the students and prevent them from falling behind.

According to a study by Casillas, A., Robbins, S. B., Allen, J., Kuo, Y. L., Hanson, M. A., & Schmeiser, C. (2012).

which studied the school environment and included 4,660 students from 24 schools, the results indicate that the strongest indicator is the prediction of previous grades, projects, and activities, in addition to psychological and social factors. From these results, we emphasize the importance of early and effective assessment based 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

- To review research studies on the causes of student academic weakness, methods of treatment, and ways to enrich high achievers
- To collect and analyze student grade data from the first period of the semester
- To create a tool to help teachers identify students with low grades early on and help them pass the course
- To provide teachers with remedial and enrichment plans

1.3 Thesis Contribution

This work focuses on middle and high school students to help reduce end-of-year academic failure and help teachers identify students at risk of failing a course, develop remedial plans for them, identify high-achieving students, and develop enrichment plans for them, thereby enhancing the quality of the educational process. This is done by creating a tool designed with machine learning models and using real student data to provide interpretable results and take into account future expansion.

1.4 Thesis Organization

Chapter 1 : Introduction

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.

Chapter 2 : Overview

This chapter 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 3 : 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 2 : Overview

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

Numerous studies have used this method and proven its accuracy and efficiency in making predictions, but some weaknesses have emerged when dealing with large amounts of data. This problem has led to the emergence of some modern deep learning models that help educational and large institutions. Examples of older models include:

- **Naïve Bayes** This assumes that all data variables are independent of each other, which is why it is called “naïve.” It is often used in text classification, such as spam classification, and relies on the principle of calculating the conditional probabilities of features based on category classification, then classifying the data according to these probabilities.
- **Support Vector Machine (SVM)** There is the largest possible margin or hyperplane separating the categories, and it is used in image and text classification and anomaly detection.
- **K Nearest Neighbour (KNN)** Classifies data based on the most common points among its K nearest neighbors and is used to classify images, aggregate data, and recognize patterns.
- **Random Forest** Relies on creating a set of decision trees that are aggregated to arrive at the final classification and is used to classify complex data containing many variables.
- **Decision Tree** It divides data into branches based on feature values and is commonly used in decision making, risk assessment, and classification tasks. It works on the principle of repeatedly dividing data until it reaches a homogeneous set of leaves where all cases are of the same class.
- **Logistic Regression** Classifies data into two categories (binary classification) and is used in binary predictions, such as whether a customer will buy a product or not. It uses a logistic function to convert the linear combination of features into probabilities between 0 and 1 to classify the data.
- **XGBoost** This model builds a set of decision trees where each tree is better than the previous one and has fewer errors using gradual regression to improve the model's performance. It is used in classification, regression, and ranking and is characterized by its high accuracy and efficiency in educational competitions and is excellent for structured and complex data.

2.3 Deep Learning Approaches

- One of the most important artificial intelligence models (**Artificial Neural Networks – ANN**) It is used in complex data analysis and aims to simulate the human brain through a set of interconnected artificial neurons. It is characterized by its ability to represent non-linear data between variables and is suitable for prediction, classification, and regression tasks. It is trained using a backpropagation algorithm to reduce the error between actual and predicted values. occur.
- **Multilayer Perceptron (MLP)** This is the most common type of neural network and consists of an input layer, one or more hidden layers, and an output layer. The hidden layers help to learn complex patterns in the data and have proven effective in many educational applications, particularly academic prediction for students.
- **Convolutional Neural Networks (CNN)** Designed to process spatially structured data such as images and visual data, these networks rely on convolutional layers capable of automatically extracting features. CNNs are commonly used in the fields of computer vision and pattern recognition.

They can also analyze multimedia educational data.

- **Recurrent Neural Networks (RNN)** Developed to process temporal and sequential data, RNNs rely on feedback and allow previous information to be retained. They are suitable for measuring student performance over time, but can suffer from gradient vanishing when dealing with long sequences. This has led to the emergence of
 - Long Short-Term Memory (LSTM) networks as an improved version of RNN, allowing important information to be retained for a long period of time.
 - Gated Recurrent Units (GRU) are an advanced model of RNN and are simpler than LSTM and require less training time while maintaining similar performance. They are used in applications that require higher computational efficiency.

- **Deep Neural Networks (DNN)** contain a large number of hidden layers and are used in advanced fields such as voice recognition and analysis of huge amounts of data. They require large amounts of data and high computational capabilities to achieve better performance

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.

Of these methods, it is clear that we can use one algorithm or a combination of them, such as:

- Decision Tree
- Naïve Bayes
- Logistic Regression
- Support Vector Machine (SVM)
- Recurrent Neural Network (RNN)
- Artificial Neural Network (ANN)

to predict student grades early on and identify weaknesses. Data will be collected from several schools in Saudi Arabia over a period of two years, and noise, loss, and unbalanced distribution will be processed to ensure the effectiveness of the model and the accuracy and comprehensiveness of the predictions to help teachers identify students who are prone to failure and intervene to address their weaknesses. Then comes the role of measuring the impact on students following the intervention after the prediction and determining the extent to which the study's objective has been achieved

2.4 Use of Real-world Datasets

Collect student grade data from the first and second semesters of the previous two years, along with data from the first semester of this year, including student information (midterm exams, participation grades, assignments, and projects) from several schools, in addition to a database from Kaggle and Zenodo. It is necessary to address the difference in grade distribution between databases and analyze grades to identify student patterns and understand the relationship between them. Noise, missing data, and unbalanced distribution will be addressed to ensure the effectiveness of the model and the accuracy and comprehensiveness of predictions to help teachers identify students who are at risk of failure and intervene to address their weaknesses. Next comes the role of measuring the impact on students after the intervention following the prediction and determining the extent to which the study's objective has been achieved.

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)=0.917 (ANN)=0.980 Naïve Bayes=0.999 Decision Tree=0.937
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%
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%

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 Related Work

Research has focused on the use of artificial intelligence and machine learning algorithms to predict student outcomes in educational settings. The goal is to improve our understanding of student academic performance, develop early detection systems for students at risk of academic failure or dropout, and implement preventive measures. Among the most prominent algorithms used are artificial neural networks (ANNs).

semi-supervised learning, random forests, and probabilistic learning. Research and studies on this topic have varied.

A group of them focused on predicting student performance and early detection of at-risk students and used traditional artificial intelligence techniques. Studies such as **Khan et al. (2021)** and **Lau et al. (2019)** and **Sukhbaatar et al. (2019)** relied on traditional artificial neural networks and decision trees to analyze student data. This group focused on students' previous grades, attendance, and exams. One of the most notable findings of this research was that neural networks and decision trees outperformed other models, with an accuracy of 86% -91%, enabling educational institutions to provide remedial plans for struggling students, such as extra lessons and early counseling.

Another group of studies used probabilistic and advanced models for more accurate early prediction, such as **Nimy et al. (2023)** and **Embarak & Hawarna (2023)**. The researchers used Probabilistic Machine Learning and AutoML to measure the probability of academic failure at each stage of the semester, rather than giving only a final result. The results showed that these models increased the reliability of early predictions and allowed for timely intervention, with the models reaching an accuracy of 92%.

A bunch of studies used advanced techniques like recurrent neural networks and semi-supervised learning, like

Mondal & Mukherjee (2018) and **Melo & Souza (2023)**, where RNN and semi-supervised learning were used to analyze data. The models showed particularly effective performance in tracking performance changes over time and predicting early academic failure, and the results prove that they are useful in long-term early intervention programs.

The last group, **Deo et al. (2020)** and **Zeineddine et al. (2021)**, they focused on developing high-performance models using AutoML and Extreme Learning Machine (ELM) to predict student grades in specific courses. These models showed high prediction accuracy with low errors and were better than traditional models such as Random Forest or traditional neural networks.

Ultimately, we can agree that the use of artificial intelligence models is capable of accurately predicting student performance and enabling academic institutions to intervene early. The difference between them is the type of algorithm used and the prediction goal. Traditional neural networks and decision trees showed good performance for direct prediction and interactive support, while probabilistic models and AutoML have demonstrated their strength in estimating uncertainty for early intervention. We also have RNN and Semi Supervised Learning models, which were ideal for analyzing sequential and incomplete data to improve long-term predictions.

Conclusion

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.

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>