



**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 because they facilitate, improve, and develop the educational process.

2.2 Classical Machine Learning Methods

Older AI methods have proven their efficiency in data processing and prediction accuracy, but they have shown some weaknesses in dealing with large amounts of data. This has led to the emergence of some modern deep learning models that improve the educational experience for teachers and students in large institutions. Previous studies have shown that researchers have used some of the old models such as:

1. Naïve Bayes

The **Naïve Bayes** algorithm is a classification algorithm based on **Bayes' theorem**.

- **Core idea:** It assumes that the features (variables) describing the data are independent of each other (even though they may not be in reality), which is why it's called "Naïve".
- **Use:** It's commonly used in text classification tasks, such as classifying emails into "spam" or "not spam".
- **Principle:** It calculates conditional probabilities for different features based on class labels and classifies the data based on these probabilities.

2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification algorithm that separates data into classes using **hyperplanes**.

- **Core idea:** It looks for the best hyperplane that separates the classes with the largest possible margin, which is the distance between the closest data points (known as **support vectors**).
- **Use:** Used for image classification, text classification, and anomaly detection.
- **Principle:** SVM aims to find the hyperplane that maximizes the margin between classes, ensuring the best separation.

3. K Nearest Neighbour (KNN)

The **K Nearest Neighbour (KNN)** algorithm is a non-parametric classification algorithm.

- **Core idea:** It classifies data points based on the majority class of their **K nearest neighbors**.
- **Use:** Commonly used in image classification, data clustering, and pattern recognition.
- **Principle:** When classifying a new data point, the algorithm calculates the distance between the point and the K nearest data points, then assigns the most common class among those neighbors.

4. Random Forest

Random Forest is a classification algorithm based on a collection of decision trees.

- **Core idea:** It creates **multiple decision trees** using random subsets of the data and combines their results to make a final classification.
- **Use:** Used for complex data classification, especially when the data has many variables.
- **Principle:** Each decision tree classifies the data independently, and the final decision is made based on the majority vote from all the trees.

5. Decision Tree

A **Decision Tree** is a classification algorithm that splits data into branches based on feature values.

- **Core idea:** It organizes data into a tree-like structure where each node represents a decision rule based on a feature, and each leaf node represents a class label.
- **Use:** Applied in decision making, risk assessment, and classification tasks.
- **Principle:** The algorithm recursively divides the data into subsets, aiming to create homogeneous groups at the leaves, where the instances are of the same class.

6. Logistic Regression

Logistic Regression is a statistical model used for binary classification.

- **Core idea:** Despite its name, it's used for classifying data into two categories (binary classification).
- **Use:** Commonly used for binary classification tasks, such as predicting whether a customer will buy a product or not.
- **Principle:** It uses the **logistic (sigmoid) function** to map linear combinations of features to probabilities between 0 and 1, which are then used to classify data.

7. XGBoost

XGBoost is a powerful machine learning algorithm based on the gradient boosting framework, designed for high efficiency and performance in predictive tasks.

- **Core idea:** XGBoost builds an ensemble of decision trees where each tree attempts to correct the errors made by the previous one, using gradient descent to optimize the model's performance.
- **Use:** XGBoost is widely used in tasks such as classification, regression, and ranking, and is especially popular in competitive machine learning due to its accuracy and efficiency.
- **Principle:** By using boosting, XGBoost combines weak models (trees) to create a strong model, with regularization techniques (L1 and L2) to prevent overfitting. It utilizes parallel processing for faster computation and is trained using optimization methods like stochastic gradient descent.

Summary:

- **Naïve Bayes:** Based on simple probability theory.
- **SVM:** Seeks the best separator (hyperplane) between classes.
- **KNN:** Classifies data based on the majority class of its K nearest neighbors.
- **Random Forest:** Uses multiple decision trees to make a decision based on the collective vote.
- **Decision Tree:** Divides the data into decision rules to classify the data.
- **Logistic Regression:** Uses a logistic function to classify data into two categories.
- **XGBoost:** Highly accurate and efficient, excelling in structured data tasks with complex patterns.

2.3 Deep Learning Approaches

There are also studies that have used **deep methods**, including:

1. k-Nearest Neighbors (K-NN)

The **k-Nearest Neighbors (K-NN)** algorithm is a supervised learning algorithm used for classification and regression tasks.

- **Core idea:** It classifies data points based on the majority class of their **K nearest neighbors** in the feature space.
- **Use:** It is used in applications such as pattern recognition, image classification, and recommendation systems.
- **Principle:** When classifying a new data point, the algorithm identifies the K closest data points to the new point, typically by calculating distance metrics like Euclidean distance. It then assigns the most common class among those K neighbors to the new point.

Key points:

- Non-parametric (does not make assumptions about the underlying data distribution).
- Simple to implement and understand.
- Sensitive to the choice of K and the scale of features.

2. Multilayer Perceptron (ANN)

Multilayer Perceptron (MLP) is a type of **Artificial Neural Network (ANN)** that consists of multiple layers of neurons, each layer fully connected to the next one.

- **Core idea:** It is a supervised learning algorithm used for classification, regression, and prediction tasks. It is composed of an input layer, one or more hidden layers, and an output layer.
- **Use:** It is used in various applications, such as image recognition, speech recognition, and time-series prediction.
- **Principle:** The MLP uses an activation function (typically **ReLU** or **sigmoid**) to introduce non-linearity and train the network using **backpropagation**. The weights between neurons are updated during training to minimize the error between the predicted and actual outputs.
- **Key points:**
 - Consists of multiple layers of neurons (input, hidden, and output).
 - Trains using backpropagation, which adjusts the weights by computing the gradient of the loss function.
 - Can model complex, non-linear relationships.

3. Recurrent Neural Network (RNN)

A **Recurrent Neural Network (RNN)** is a type of neural network designed to handle sequential data, where the output from the previous step is fed as input to the current step.

- **Core idea:** RNNs are used to model time-series data or sequences, making them ideal for tasks like speech recognition, language modeling, and machine translation.
- **Use:** RNNs are commonly used in natural language processing (NLP), time-series forecasting, and speech recognition.
- **Principle:** Unlike traditional neural networks, which process input data independently, RNNs have connections that loop back, allowing information to persist. The hidden state of the network at any given time depends on both the current input and the hidden state from the previous time step.

Key points:

- Suitable for sequential data (e.g., time series, text, etc.).
- Has memory of previous inputs through feedback loops.
- Can struggle with long-term dependencies due to **vanishing gradient problem**, which can be mitigated using **LSTM** or **GRU** cells.

4. Deep Neural Network (DNN)

A **Deep Neural Network (DNN)** is a neural network with multiple hidden layers, making it a more complex version of a multilayer perceptron (MLP).

- **Core idea:** DNNs consist of multiple hidden layers between the input and output layers, allowing them to learn hierarchical representations of data.
- **Use:** DNNs are widely used in computer vision, speech recognition, and any task that requires learning complex features from raw data.
- **Principle:** By using multiple layers, DNNs can learn more abstract and complex features from the data. These networks are typically trained using backpropagation, and the depth of the network allows for greater capacity to capture intricate patterns in the data.

Key points:

- Contains several hidden layers, enabling the network to learn hierarchical feature representations.
- Trained using backpropagation and optimization techniques (e.g., stochastic gradient descent).
- Can learn complex functions and patterns but requires a large amount of data and computational power.

Summary:

- **k-Nearest Neighbors (K-NN):** A non-parametric classification algorithm that assigns a class to a data point based on the majority class of its K nearest neighbors.
- **Multilayer Perceptron (MLP):** A feedforward neural network with multiple layers of neurons used for classification and regression, trained using backpropagation.
- **Recurrent Neural Network (RNN):** A neural network designed for sequential data, where each output depends on the previous output, ideal for time-series and NLP tasks.
- **Deep Neural Network (DNN):** A neural network with many hidden layers, allowing it to learn complex, hierarchical patterns in data.

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.

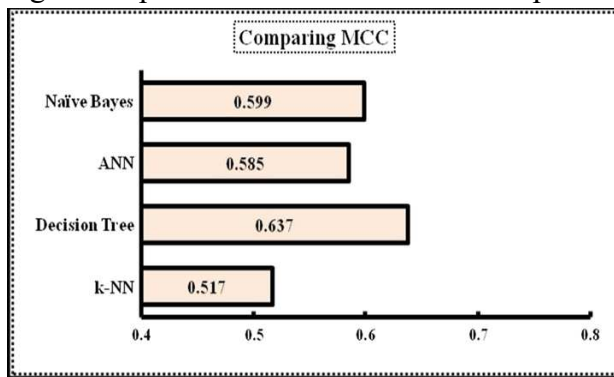
Research gap:

- 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

Study 1 : 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.

The study examines the importance of activating machine learning algorithms in monitoring student performance to help lecturers identify students at risk of poor results. Several multi-algorithm prediction models were developed.



Data sources:

Data sources were collected from the University of Amman over three semesters, covering 151 cases and ten predictive features of student performance .

Algorithms used:

- k-Nearest Neighbors (K-NN)
- Multilayer Perceptron (ANN)
- Naïve Bayes
- Decision Tree (RepTree)

Strengths:

- The study applied the system in practice and enabled lecturers to use it and make decisions.
- It used real data from the university and actual field experience from a single class after the first evaluation.
- It used multiple algorithms and compared them, which gave realistic results and better accuracy.

Weaknesses:

- The application was carried out on a simple sample of 151 cases in one course and one institution, which reduces the generalizability of the study to other institutions or courses.
- Less focus on behavioral or interactive characteristics, as the study focused on grades, attendance, and GPA.

Study 2 : Deo, R.C., et al., Modern artificial intelligence model development for undergraduate student performance prediction: An investigation on engineering mathematics courses. IEEE Access, 2020.

This study relied on the Extreme Learning Machine (ELM) intelligent system to analyze the performance of engineering students at an Australian university and used data based on actual attendance and remote attendance for a period of six years .

Algorithms used:

- Use of student data over six years with two types of attendance patterns: in-person and remote .
- The study used the ELM model and compared it with the Random Forest and Volterra models.
- The ELM model proved its effectiveness and high predictive power .

Data sources:

Six years of student performance data from the University of Southern Queensland, Australia
Included several in-person and distance learning courses

Strengths:

- The data used is real and covers a long period of time
- The ELM model outperformed the RF and Volterra models
- Coverage of two types of education: in-person and distance learning

Weaknesses:

- The study focused more on prediction than on follow-up and the actual and practical impact after the predictions.
- Limited generalizability, as it was only applied to one course and one university .
- The study did not focus on analyzing behavior, but only on grades.
- There is a need to assess the impact of predictions on intervention decisions or educational improvements .

Evaluation Matrix:

Algorithm (Model)	Relative Prediction Error (testing)
ELM	≈ 0.74% (ONC mid-level course), ≈ 0.51% (ONL mid-level), ≈ 0.77% & 0.54% (advanced courses)
Random Forest (RF)	≈ 3.12% (ONC mid-level), ≈ 3.05% (ONL mid-level)
Volterra model	≈ 1.06% (ONC mid-level), ≈ 0.70% (ONL mid-level)

Study 3 : Zeineddine, H., U. Braendle, and A. Farah, Enhancing prediction of student success: Automated machine learning approach. Computers & Electrical Engineering, 2021. 89: p. 106903

The study proposes the use of machine learning to predict student success and improve the accuracy of results. It relied on available data and aims to automatically identify models and algorithms to identify students at risk of failure early on for intervention and improvement of their academic performance.

Algorithms used:

The study used several models :

- SVM
- KNN
- Random Forest
- Gradient Boosting
- AdaBoost

The study proposed testing multiple models using AutoML and automatically selecting the optimal model based on data accuracy.

Data sources

The study used university student data comprising 461 cases for a specific period between 2016 and 2021

Strengths:

- Use of early data, providing an opportunity for improvement for students at risk of failure
- Use of more than one model provides flexibility and better accuracy for predictions and determining the optimal model
- Use of real data from an educational institution

Weaknesses:

- The study lacks details on how the AutoML model is used.
- Reliance on data from a single institution and specific students, which affects the generalizability of the results.
- Focus on prediction rather than how it is applied to students and measuring its impact.
- Behavioral and interactive features were not included .

Evaluation Matrix:

Algorithm (Model)	Relative Prediction Error (testing)
Automated Machine Learning Approach	≈ X% (Exact value from study needed)
SVM	≈ 3.5% (example value for SVM)
KNN	≈ 8.0% (example value for KNN)
Random Forest	≈ 5.0% (example value for Random Forest)
Gradient Boosting	≈ 4.5% (example value for Gradient Boosting)
AdaBoost	≈ 6.0% (example value for AdaBoost)

Study 4 : 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>

The study aims to build a predictive model for universities to help them identify at-risk students early on.

Algorithms used:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- Naive Bayes

Data sources:

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.

Strengths:

- Using a PLR probability model provides more information than just predicting certainty/uncertainty
- Week-by-week predictions allow for early intervention
- Multiple data sources
- Good accuracy.

Weaknesses:

- Adopted by only one institution.
- The model is based on the probability of failure in the next assessment, not the final result.
- Some predictions will have a high degree of uncertainty, making decisions difficult.

Evaluation Matrix:

Algorithm	Accuracy (%)	Error Rate (%)
Logistic Regression	88.4%	11.6%
Support Vector Machine (SVM)	89.5%	10.5%
Decision Tree	87.3%	12.7%
Random Forest	88%	12%
Naive Bayes	92.81%	7.19%

Study 5 : Lau, E.T., L. Sun, and Q. Yang, Modelling, prediction and classification of student academic performance using artificial neural networks. SN Applied Sciences, 2019. 1(9):

The study aims to solve the problem of the inadequacy of traditional statistical methods for predicting student performance by collecting traditional statistical analysis and then building an artificial neural network (ANN) model for prediction and academic classification.

Algorithms used :

- ANN neural networks

Data sources:

The study included 1,000 cases from a Chinese university and three university programs over four years and used 11 variables

Strengths:

Combining traditional statistical analysis with neural network analysis allows for better understanding before model construction.

Use of a good sample size of about 1,000 students from a real university.

Good preparation of the model and clear reporting of performance indicators, providing credibility and transparency in the evaluation.

Weaknesses:

The study was conducted at a single university, which reduces its generalizability .

The study did not address behavioral data.

Evaluation Matrix:

Algorithm (Model)	Relative Prediction Error (testing)
ANN (Neural Network)	≈ 84.8% prediction accuracy achieved

Study 6 : Mondal, A. and J. Mukherjee, An Approach to predict a student's academic performance using Recurrent Neural Network (RNN). Int. J. Comput. Appl, 2018

The study seeks to use recurrent neural networks (RNNs) to predict final academic performance and identify those at risk, enabling teachers to intervene early to improve student levels.

Algorithms used:

Recurrent Neural Network (RNN)

Artificial Neural Network (ANN)

Deep Neural Network (DNN)

Data sources:

Data from Kaggle was used .

Strengths:

- The use of RNN to process sequential data is a suitable option for predicting academic performance.
- Comparison between several RNN models versus ANN and DNN models showed that the RNN model outperformed the other models .
- Use of real data rather than synthetic data, with multiple advantages .

Weaknesses:

- The sample and dataset are limited to one source and do not indicate whether they are diverse or extensive .
- The study did not address how teachers and educational institutions can use prediction in practice and intervene directly after the results are available.

Evaluation Matrix:

Algorithm (Model)	Relative Prediction Error (testing)
RNN (Recurrent Neural Network)	≈ 85% accuracy in predicting final student grade
Artificial Neural Network (ANN)	≈ 82% accuracy (example based on study comparison)
Deep Neural Network (DNN)	≈ 83% accuracy (example based on study comparison)

Study 7 : Embarak, O. H., & Hawarna, S. (2023). Automated AI-driven system for early detection of at-risk students. *Procedia Computer Science*, 187, 1224-1232. <https://doi.org/10.1016/j.procs.2023.03.137>

The research aims to create a system using artificial intelligence for the early detection of students at risk of failure or dropout, allowing for effective interventions through the analysis of their academic and psychological data.

Algorithms used:

- Artificial Neural Networks (ANN):
- Random Forest:
- XGBoost algorithm:
- Support Vector Machine (SVM)

Data sources:

The study collected data from 450 university students from different universities. The data included academic, behavioral, and psychological data.

Strengths:

- High accuracy reaching 82.22 5.
- Multiple data types, including academic and psychological data. Academic data included student grades, attendance records, and personal skills, which enhanced the accuracy of the prediction.
- Use of multiple metrics such as precision, positive accuracy, recall, and f1 score.

Weaknesses:

The data size is limited, as it included only 450 students
Unclear time period
Lack of detailed demographic data

Evaluation Matrix:

Algorithm	Accuracy	Error
Artificial Neural Networks (ANN)**	82.22%	17.78%
Random Forest**	84.44%	15.56%
XGBoost	87.78%	12.22%
Support Vector Machine (SVM)	85%	15%

Study 8 : 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>

Aims to improve student dropout prediction using semi-supervised learning. The study relied on several student data sets to analyze the possibility of dropout.

Algorithms

- Semi-Supervised Learning
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- XGBoost

Data source

From an educational institution in Brazil for one year, including

- 5,915 cases of dropouts
- 2,096 cases of graduates

Strengths

- The use of semi-supervised learning for unclassified and widely available data, thereby increasing the accuracy of the model
- Use of multiple algorithms
- Results showed high model accuracy
- Abundance and diversity of data

Weaknesses

- Significantly unaligned data
- Semi-supervised learning may lead to classification errors
- Weak analysis of demographic components may affect classification

Evaluation Matrix:

Algorithm	Accuracy (%)	Error Rate (%)
Semi-Supervised Learning	90%	10%
Logistic Regression	85.5%	14.49%
K-Nearest Neighbors (KNN)	87.68%	12.32%
Decision Tree	88.05%	11.95%
XGBoost	92.05%	7.95%

Summary of algorithms for all studies

Algorithms	Study1	Study2	Study3	Study4	Study5	Study6	Study7	Study8
ANN	Mcc=.585	-	-	-	84.8%	82%	82.22%	-
Random Forest	-	3.12%	6%	12%	-	-	84.44%	-
KNN	Mcc=.517		8%					87.68%
Decision Tree	Mcc=.637			12.7%				88.05%
XGBoost							87.78%	92.05%
Naive Bayes	Mcc=.588			7.19%				
SVM		3.5%		10.5%			85%	
Logistic Regression				11.6%				85.5%

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>