Predicting Future Grades of High School Students Using Machine Learning Techniques

*Nok Yau Bruno Chen**

**The Advanced Technology Research Lab, Tsinglan School, Dongguan, Guangdong, 523808, China*

Abstract

This paper investigates the use of machine learning techniques to predict the future grades of high school students which could potentially be translated into a practical application that can benefit high school students is a guide on course selection. The study utilizes a dataset from a Chinese International School, comprising courses from AP, A-Level, iGCSEs, and internal courses. Various machine learning models and techniques, including decision trees, random forests, K-nearest neighbours, multilayer perceptron, long short-term memory, and convolution neural networks, are employed to approach this problem from multiple perspectives. The performance of each model is evaluated using mean absolute error and root mean square error. The Random Forest model demonstrates the best predictive performance, with the lowest MAE and RMSE values among all models tested. The study's findings provide a tool for educators and school administrators to identify students who may be at risk of falling behind in their studies and implement targeted interventions to support their academic progress. However, the study's results are limited by the relatively small dataset, and future studies should explore other deep learning techniques and model architectures, as well as integrating additional features into the models to enhance their predictive capabilities.

Introduction

The academic performance of high school students is often a source of anxiety, both for students themselves and their educators. This concern is heightened by the uncertainty surrounding grades in different courses and the potential impact on prospects. Consequently, educators are constantly seeking ways to identify students who may be at risk of falling behind and provide them with appropriate support.

This study focuses on utilizing machine learning containing deep learning techniques to predict the future grades of grade 9 students by employing data from past students to train the model and evaluated its performance using various metrics. The dataset used in this study was obtained from a Chinese International School and comprised of courses from College Board Advance Placement, Cambridge International AS & A Levels, Cambridge iGCSEs, and school internal courses.

This study aims to approach this problem from multiple perspectives by utilizing diverse machine-learning models and techniques, which could potentially be translated into a practical application that can benefit high school students. A software could be built based on the research to guide and help students on their future course selection.

Related Work

Previous studies have utilized various machine learning approaches to predict academic performance, including linear regression, decision trees, and neural networks. However, few studies have explored the effectiveness of deep learning techniques, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), in predicting high school students' future grades. [1]

According to a study conducted by Akhilesh Patil et al., which utilized data from multiple years, a Long Short-Term Memory Deep Learning model is more accurate than other deep learning or traditional machine learning models. The study focused on predicting and classifying students from a secondary school in Portugal using multiple classifying models, including deep learning models. The study's findings suggest that the Long Short-Term Memory Deep Learning model is the most effective in predicting student outcomes. [2]

Another Study by Sujan Poudyal et al. gives a method to use a Hybrid 2D CNN model to predict the academic performance of students, which resulted from a very high accuracy and outperformed other deep learning models such as CNN and LSTM while using the same dataset. [3]

Data Preprocessing and Model Training

The dataset of this study is derived from an International School located in mainland China. It comprises 1638 lines of data containing the Student ID, Course name, Grade achieved, and School Year attended.

Student No.	Course	Grade	School Year
4	C-Humanities	A	G10
4	Physical Education	A+	G10
4	IT (Year 1)	A+	G10
4	AP 2D Art and Design	A+	G9
4	IGCSE English Literature (Year 1)	A-	G9
4	IGCSE Math (Year 1)	A-	G9

Table 1, A Snippet of the Dataset

After preprocessing the data, which involved removing less popular courses and some electives, the dataset was divided into two matrices. The first matrix represents Grade 9 and consists of 79 students and 24 courses. The second matrix represents Grade 10 and includes 79 students and 24 courses. In both matrices, each column corresponds to a specific course, while each row represents a student.

To provide an example, the courses in the Grade9 matrix include Chinese Culture, Drama I, English I, IGCSE Biology (Accelerated), IGCSE Combined Science (Year 1), IGCSE Economics (Year 1), and IGCSE English Literature (Year 1). The grades achieved by students in these courses range from A+ to D-, with each student having their unique score for each course. While N represents that the student had not attended the course

Subject No	**Chinese Culture**	**Drama I**	**English I**	**IGCSE Biology (Accele rated)**	**IGCSE Combined Science (Year 1)**	**IGCSE Economic s (Year 1)**	**IGCSE English Literature (Year 1)**
1	A+	N	D-	C	N	N	N
2	A+	N	A-	N	A-	N	N
3	A+	A-	N	B+	N	B+	A-
4	A+	N	N	A+	N	N	A
5	A	N	N	N	A-	N	C-
6	A+	N	A+	N	A	B+	N

Table 2, A Snippet of the processed dataset from the 79 students in Grade 9

Before the training of models, the training set and the test set was split on a 70:30 basis. The input and output grades will be transformed into numeric scores from 0 to 100 using the table below.

Letter Grade	Numerical Grade
A+	97
A	93
A-	90
B+	87
B	83
B-	80
C+	77
C	73
C-	70
D+	67
D	63
D-	60
F	0

Table 2, Grade dictionary from Letter Grade to Numerical Grade

For training, a separate model was trained for each subject in Grade 10 using the corresponding subjects in Grade 9 as input. This approach was adopted to develop models that are specifically tailored to the unique characteristics of each subject. As a result, this approach led to improved performance in predicting grades for individual subjects.
Two metrics are being used to evaluate the models: mean absolute error (MAE) and root mean square error (RMSE). MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. The following equation is used to find the root mean square error.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Where y_i represents the predicted value x_i represents the actual value and n represents the total number of data points.

RMSE measures the square root of the average squared differences between predicted and actual values. The following equation is used to find the root mean square error.

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}}$$

Where y_i represents the predicted value, x_i represents the actual value, and n represents the total number of data points. Due to that separated model will be trained for each subject in grade 10. To evaluate the performance of the models, it could only work to calculate the RMSE and MAE for each subject in grade 10. Then calculate the average MAE and RMSE across all subjects.

Models

Decision Tree and Random Forest

A decision tree is a supervised learning algorithm that can be used for both classification and regression tasks. It creates a tree-like model of decisions and their possible consequences. Each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a numerical value. Decision trees are easy to interpret and can handle categorical and numerical data, such as scores, easily. However, decision trees can suffer from overfitting, leading to low precision in many cases.

Random forest is a model that creates multiple decision trees and combines them to provide a more accurate and stable prediction. Each tree is built independently using a random subset of the data, and the algorithm combines the predictions of all the trees to make a final decision. This process, known as bootstrapped aggregation (or bagging), reduces overfitting and generally leads to better predictive performance compared to individual decision trees.

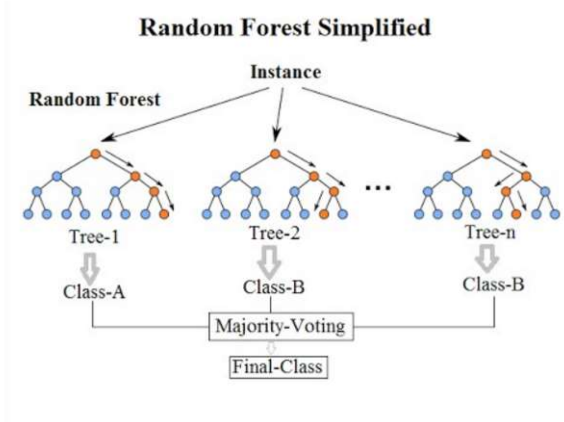

Figure 1, Random Forest Regressor

For 1D data like ours, decision tree and random forest can be easily implemented by using the “scikit-learn” library using python.

K-Nearest Neighbours

KNN is a supervised learning algorithm that can be used for both classification and regression tasks. It works by finding the K most similar data points to a given query point, based on some distance metric, and assigns it the label or value that is most common among those neighbours. KNN is also easy to implement and can handle numerical data effectively.
Again, KNN can be easily implemented using the “scikit-learn” library using python.

Deep Learning Approaches

Multilayer Perceptron (MLP)

A multilayer perceptron is a type of feedforward network used in deep learning. A feedforward network is a unidirectional neural network, which means that the information in the model flows in only one direction. MLP consists of an input layer, one or more hidden layers, and an output layer, each of which contains one or more neurons. In this study, a 7-layer MLP with 5 hidden layers and a 5-layer MLP with 3 hidden layers will be trained and tested.

Long Short-Term Memory (LSTM)

LSTM is a type of recurrent neural network (RNNs). Unlike feedforward networks, RNNs contain cyclic connections, this enables RNNs to process sequences of variable length and make predictions based on the order and context of inputs. While LSTM is a type of RNN that have the capability to learn from previously results from previous layers. Which give LSTM to predict data base on time series.

In this study,1d array grades will be input into a LSTM shape by setting the timestamp, which is the second dimension as 1. This will make LSTM not perform as well as having multiple timestamps, but due to data limitation, this is all that could be done.

Two LSTM models, one single layer LSTM and one multilayer LSTM with 3 layers in the model will be trained and tested.

Convolution Neural Networks (1D Convolution Layers)

In this study, 1D convolution layers are utilized as part of the Convolutional Neural Network (CNN) approach for predicting future grades of high school students. Convolutional Neural Networks are commonly used in image analysis tasks, but they can also be applied to 1D data, such as the grades dataset in this study.

The main advantage of using 1D convolutions in this study is that they can capture local dependencies and patterns in the sequence of grades. By sliding the filters across the input data, the model can detect important patterns, such as upward or downward trends, sudden changes, or recurring patterns in the grades.

In this study, LSTM models will be utilized by adding 1d convolution layers in front them. Which will theoretically increase the model's performance.

Results and Analysis

Model	MAE	RMSE
Decision Tree	6.213585267	9.271595492
Random Forest	5.33707971	8.179813685
KNN	5.691962555	8.552175896
5 Layer MLP	10.72331882	13.5741314
7 Layer MLP	5.748673787	8.486028022
Single Layer LSTM	6.2636655	9.049063984
Multi-Layer LSTM	6.136094972	8.942161071
Conv1D+LSTM*3	6.263691651	9.049071594
Conv1D+LSTM	6.270312046	9.038555939
Conv1D*2+LSTM	5.896450494	8.924071343
Conv1D*2+LSTM*3	6.263652676	9.049025002

Table 3, Results of Each type of Model

Random Forest model demonstrated the best predictive performance, as evidenced by the lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values amongst all models tested. This outperformance was observed in comparison to both traditional machine learning models and deep learning models.

The 7-Layer MLP model also shows good predictive accuracy, with relatively low MAE and RMSE values. However, the 5 Layer MLP model has the highest MAE and RMSE values among all models, this could be the result of underfitting.

The LSTM models (Single Layer LSTM and Multi-Layer LSTM) and the Conv1D+LSTM models exhibit similar performance, with comparable MAE and RMSE values. In contrast, the Conv1D*2+LSTM model performs well, with relatively low MAE and RMSE values compared to other deep learning models.

One possible reason for the Random Forest model's superior performance could be the relatively small amount of data available for training. Deep learning models generally require large amounts of data to achieve optimal performance.

Furthermore, the LSTM models' performance may have been affected by the limited depth in the time dimension of the dataset. LSTM models are designed to capture long-term dependencies by maintaining a memory of past inputs and selectively forgetting or updating this memory based on new inputs. However, in this study, the dataset only covers two years of data, and the model may not have enough information to effectively learn and generalize these dependencies.

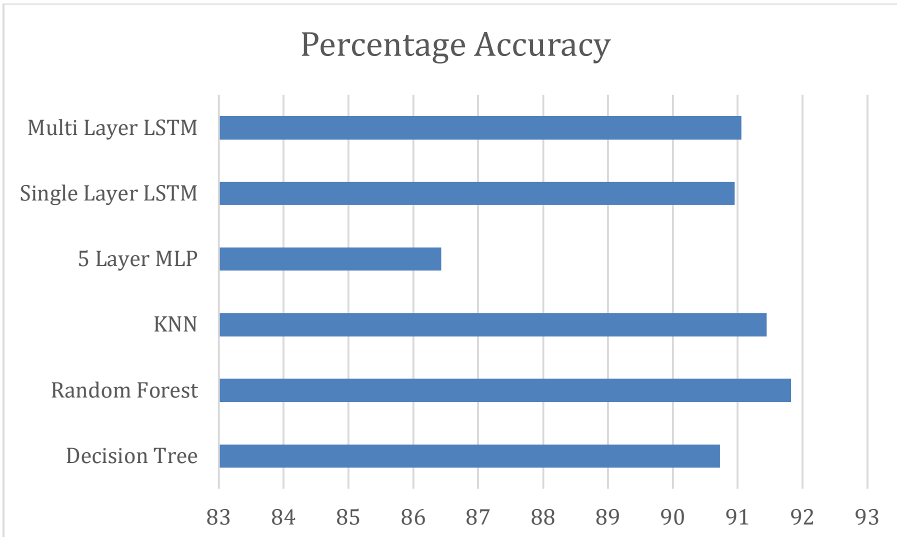

Figure 2, Comparison of Models' Accuracy

Conclusion and Future Study

In conclusion, the results of this study demonstrated the potential of machine learning and deep learning techniques in predicting high school students' future grades. The Random Forest model showed the best performance in terms of both MAE and RMSE, indicating its strong predictive accuracy. However, deep learning models, specifically the 7 Layer MLP and Conv1D*2+LSTM, also demonstrated promising results.

The study's findings could be beneficial for educators and school administrators, providing them with a tool to identify students who may be at risk of falling behind in their studies. Early identification of such students could allow for the implementation of targeted interventions to support their academic progress.

However, there are a lot of limitations. The dataset used in this study was relatively small and covered only two years of data. Deep learning models, such as LSTM, typically require large amounts of data and may perform better with a more extensive dataset. Additionally, the LSTM models' performance might have been influenced by the limited depth in the time dimension of the data.

For future studies, it would be beneficial to gather a larger dataset that covers a longer period. This would allow the models to capture more complex patterns and dependencies in the data, potentially improving their predictive accuracy. It would also be interesting to explore other deep learning techniques and model architectures, such as transformer models or graph neural networks, and assess their effectiveness in this task.

Furthermore, it would be worth investigating the integration of additional features into the models, such as demographic information, attendance records, or behavioral data, which could potentially enhance the models' predictive capabilities.

References

1. Xu, J., Moon, K. H., & van der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, *11*(5), 742–753. https://doi.org/10.1109/jstsp.2017.2692560

2. Patil, A. P., Ganesan, K., & Kanavalli, A. (2017). Effective deep learning model to predict student grade point averages. *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. https://doi.org/10.1109/iccic.2017.8524317

3. Poudyal, S., Mohammadi-Aragh, M. J., & Ball, J. E. (2022). Prediction of student academic performance using a hybrid 2D CNN model. *Electronics*, *11*(7), 1005. https://doi.org/10.3390/electronics11071005