Predict Student Performance in Secondary Education Based on Machine Learning Models and Deep Learning Models

Yanchi Liu[1*]

Corresponding author(s). E-mail(s): yliu2488@wisc.edu;

1 Introduction

Academic performance is a complex construct influenced by a myriad of factors. Among these, family characteristics and personal attributes of students are recognized as pivotal determinants. Some literatures explore the interplay between these factors and their impact on educational outcomes.

Some family characteristics include socioeconomic status, parental education, family structure and stability, and parental involvement. Studies consistently show that students from higher SES backgrounds tend to perform better academically. This is often attributed to greater access to resources, such as quality educational materials and tutoring services (Entwisle Alexander, 1992). Parental education level is strongly correlated with student performance. Parents with higher education are more likely to engage in activities that foster cognitive development and academic support (Hill Tyson, 2009). A stable family environment with consistent routines and clear expectations can positively influence academic performance. Conversely, frequent disruptions and instability can lead to stress and reduced focus on studies (Amato, P. R., 2010.) Active parental involvement in a child's education, including homework assistance and school engagement, is linked to improved academic outcomes (Henderson Mapp, 2002).

Personal characteristics of students include cognitive abilities, motivation and engagement, learning strategies, emotional intelligence, and personal interests and passions. Innate cognitive abilities, including intelligence and memory, are foundational to academic success. However, the extent to which these abilities are nurtured and developed can vary greatly (Sternberg Grigorenko, 2002). Intrinsic motivation and engagement with learning material are critical. Students who are self-motivated and find learning personally relevant tend to perform better (Ryan Deci, 2000). Effective learning strategies, such as organization, time management, and critical thinking, are associated with higher academic performance (Pintrich, 2004). Emotional intelligence, including self-awareness and the ability to manage emotions, can influence academic performance by affecting social interactions and stress management (Mayer, Salovey, Caruso, 2008). Students who pursue subjects they are passionate about often exhibit higher levels of engagement and performance (Diemberger, L., 2021).

In order to explore the relationships between the student's academic performance and these potential characteristics and build a predictive model for predicting the student's performance based on these characteristics, some machine learning models and deep learning models are applied to the dataset of student performance published on the UCI machine learning repository.

2 Description of Dataset

The dataset used in this report is from the UCI machine learning repository, which collected the student academic performance data in secondary education of two Portuguese schools. The data were obtained by school reports and questionnaires including student grades, demographic, social, school-related features, and parental characteristics. There are two versions of datasets with respect to two subjects: Maths and Portuguese language. In this project, the latter one is chosen for analysis. In addition, there are three grades recorded in this dataset. For simplicity, only the final grade at 3rd period is focused. More details about the dataset can be found at the website, https://archive.ics.uci.edu/dataset/320/student+performance.

3 Exploratory Data Analysis

The dataset has 643 observations and 33 variables, which is glimpsed Figure 1. The descriptive statistics of some numerical variables are shown in Figure 2. Figure 3 displays the relationship between the outcome variable (final grade) and personal features. There are significant differences in final grades for students of two schools. The average grade of female students is slightly higher than that of male students. The grades of students who are greater than 18 years old are higher than those who are not. With the weekly study time increases, the average grades tend to be higher. The number of past class failures is a good indicator of grade, with a positive correlation between grades. As for the family features shown in Figure 4, mother's education seems to be positively correlated to students' performance. The parents' job is also a significant factor, where the students whose parents are teachers tend to have a higher grade.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
644	MS	F	19	R	GT3	T	2	3	services	other	...	5	4	2	1	2	5	4	10	11	10
645	MS	F	18	U	LE3	T	3	1	teacher	services	...	4	3	4	1	1	1	4	15	15	16
646	MS	F	18	U	GT3	T	1	1	other	other	...	1	1	1	1	1	5	6	11	12	9
647	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	6	10	10	10
648	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	4	10	11	11

649 rows × 33 columns

Fig. 1 The screenshot of the dataset in Jupyter notebook

	count	mean	std	min	25%	50%	75%	max
age	649.0	16.744222	1.218138	15.0	16.0	17.0	18.0	22.0
Medu	649.0	2.514638	1.134552	0.0	2.0	2.0	4.0	4.0
Fedu	649.0	2.306626	1.099931	0.0	1.0	2.0	3.0	4.0
traveltime	649.0	1.568567	0.748660	1.0	1.0	1.0	2.0	4.0
studytime	649.0	1.930663	0.829510	1.0	1.0	2.0	2.0	4.0
failures	649.0	0.221880	0.593235	0.0	0.0	0.0	0.0	3.0
famrel	649.0	3.930663	0.955717	1.0	4.0	4.0	5.0	5.0
freetime	649.0	3.180277	1.051093	1.0	3.0	3.0	4.0	5.0
goout	649.0	3.184900	1.175766	1.0	2.0	3.0	4.0	5.0
Dalc	649.0	1.502311	0.924834	1.0	1.0	1.0	2.0	5.0
Walc	649.0	2.280431	1.284380	1.0	1.0	2.0	3.0	5.0
health	649.0	3.536210	1.446259	1.0	2.0	4.0	5.0	5.0
absences	649.0	3.659476	4.640759	0.0	0.0	2.0	6.0	32.0
G1	649.0	11.399076	2.745265	0.0	10.0	11.0	13.0	19.0
G2	649.0	11.570108	2.913639	0.0	10.0	11.0	13.0	19.0
G3	649.0	11.906009	3.230656	0.0	10.0	12.0	14.0	19.0

Fig. 2 The descriptive statistics of some variables in the dataset

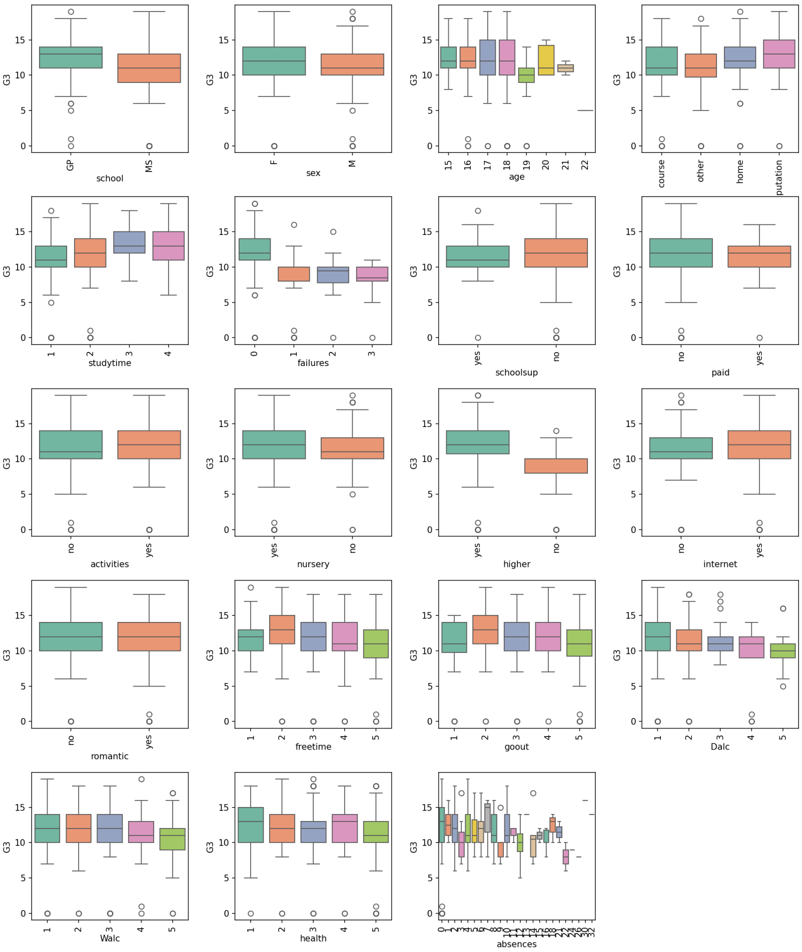

Fig. 3 The boxplots of final grade versus personal features

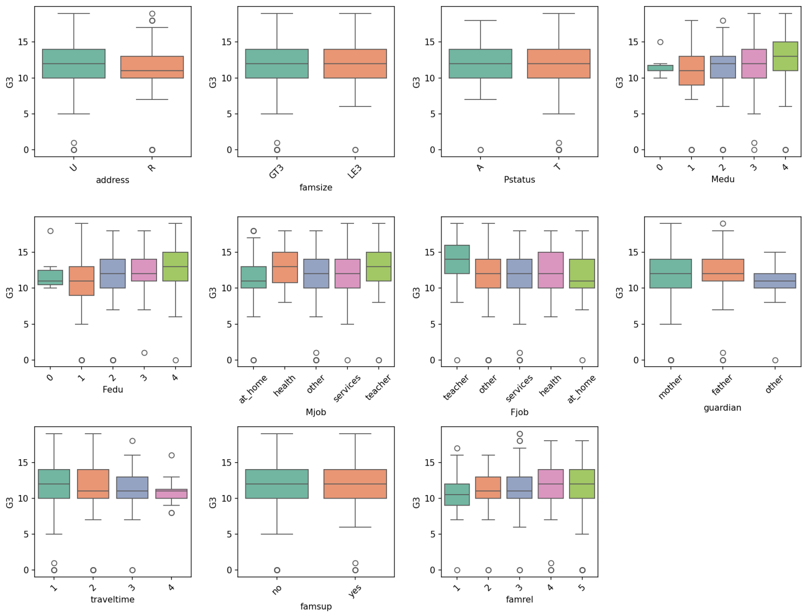

Fig. 4 The boxplots of final grade versus family features

4 Models

4.1 Regression Tree

Regression trees (Breiman et al, 1984, James et al (2013)) are a non-parametric supervised learning method used for both classification and regression. They create a model in the form of a tree structure, where each internal node denotes a test on an attribute, each branch represents the outcome of the test, and each leaf node holds a numerical value (in regression) or a class label (in classification). The goal is to split the data based on feature values that best separate the target variable, minimizing the variance within each node.

In this report, a regression tree with the maximal depth of 5, and minimal samples in leaves node of 10 is trained. The model performance of this model is summarized in the following table.

Table 1 The performance of the regression tree

Index	Training	Test
MSE	2.45555	2.96739
MAE	1.88155	2.28019
R^2	0.42715	0.09704

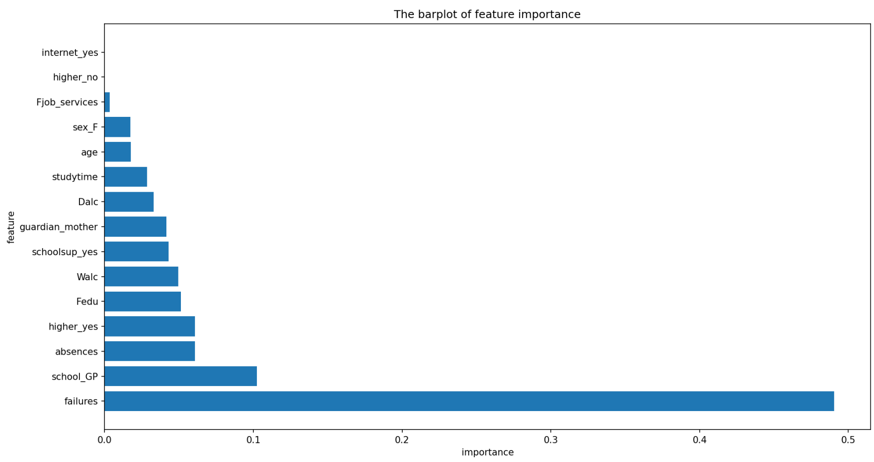

Fig. 5 The bar plot of feature importance obtained by regression tree

The bar plot of feature importance demonstrates that the number of past class failures, the school, and the number of school absences are the top three important features. It shows that personal features are more influential on students' academic performance. As for parental features, the guardian of the student and the father's job are significantot factors. The performance of XGBOOST is greater than that of the regression tree, showing the power of ensemble learning. The feature importance results are similar with that of the regression tree.

4.2 XGBOOST

XGBoost (Chen et al, 2016) is an optimized distributed gradient boosting system designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the gradient boosting framework. XGBoost provides a parallel tree boosting that solve many data-science problems in a fast and accurate way. It is widely used for machine learning competitions and real-world applications due to its speed and performance.

Table 2 The performance of XGBOOST

Index	Training	Test
MSE	0.00265	2.75889
MAE	0.00136	2.15162
R^2	1	0.21947

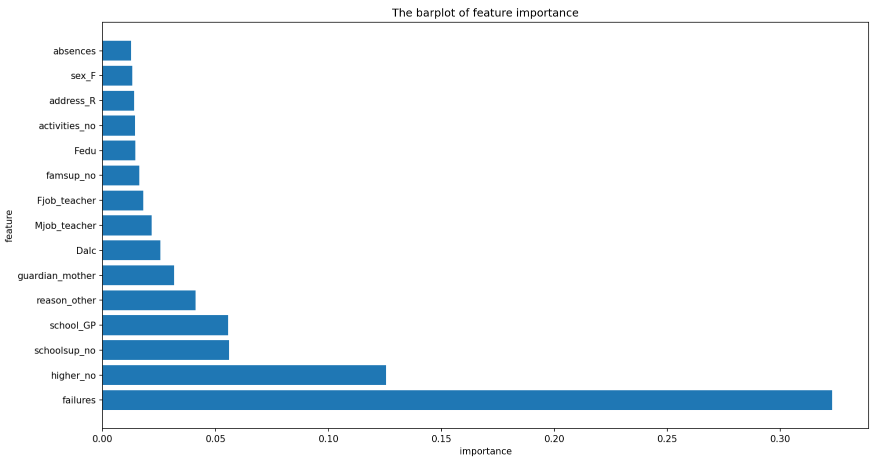

Fig. 6 The bar plot of feature importance obtained by regression tree

4.3 MLP

Multilayer Perceptrons (MLPs) are a class of feedforward artificial neural networks that consist of an input layer, one or more hidden layers, and an output layer (Goodfellow, 2016). They use a supervised learning technique called backpropagation for training, where the error is propagated backward to update the weights. MLPs can model complex relationships between inputs and outputs and are widely used for classification and regression tasks. Based on the following table, it can be found that the performance of MLP is not better than XGBOOST.

Table 3 The performance of MLP

Index	Training	Test
MSE	1.87993	2.92744
MAE	1.35962	2.19092
R^2	0.66424	0.12119

4.4 DNN

Deep Neural Networks (DNNs) (Goodfellow, 2016) are a class of artificial neural networks with multiple layers of interconnected nodes, beyond the simple perceptron structure. They leverage the power of deep learning to learn hierarchical feature representations from data. DNNs have revolutionized fields like computer vision and natural language processing due to their ability to capture complex patterns and relationships within large datasets.

The loss function used in this project is MSE. The training loss is 0.0032, while the test loss is 12.6576. The trace plot of training loss for DNN model is given in the following figure.

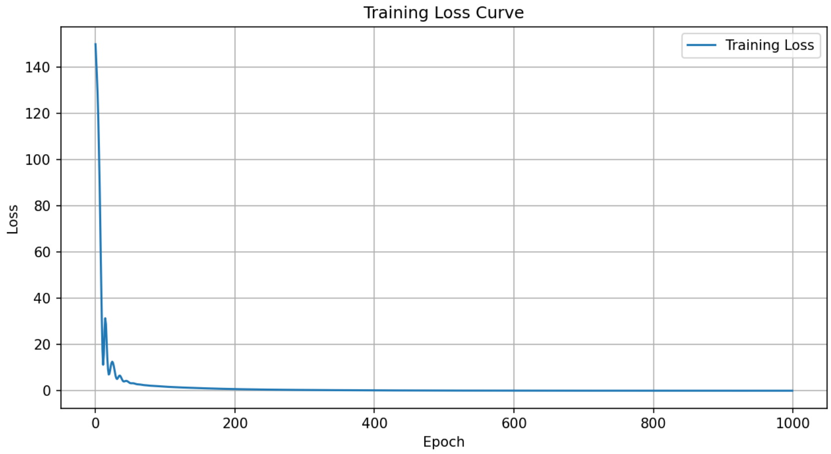

Fig. 7 The training loss plot of DNN

4.5 CNN

Convolutional Neural Networks (CNNs) are a type of deep learning model designed for processing data with a grid-like topology, such as images (Krizhevsky, 2012; Simonyan, 2014). They utilize convolutional layers to automatically and adaptively learn spatial hierarchies of features from input data. CNNs have achieved state-of-the-art performance in image recognition, video analysis, and other visual computing tasks.

The loss function used in this project is MSE. The training loss is 5.4281, while the test loss is 8.4673. The trace plot of training loss for CNN model is given in the following figure.

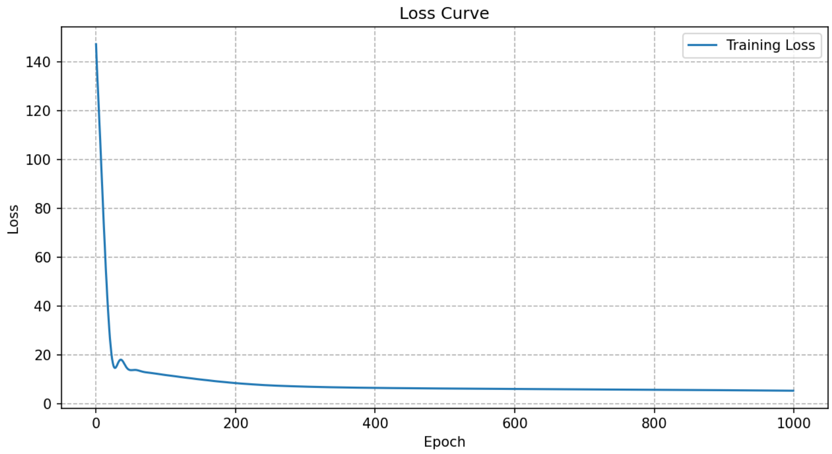

Fig. 8 The training loss plot of CNN

5 Discussion and Conclusion

In this project, some machine learning methods and deep learning methods are employed to explore the dataset for students' academic performance. Compared to deep learning models, two tree-based machine learning models are more interpretable, where the model results show that personal characteristics are more significant and influential for students' exam grades. As for parental features, the guardian and father's jobs are significant factors for students' performance. These deep learning models, MLP, DNN and CNN, are used for predicting the students' grades based on the characteristics. But these deep learning models do not outperform machine learning models. It may be due to reasons such as limited data availability, data sizes, overfitting when not enough regularization is applied, and the effectiveness of well-crafted features in traditional models.

Supplementary Material. The data, python code in the Jupyter notbook and Latex source code for this project can be found at my Github repository https://github.com/Yc-L722/psych755_project.git.

References

https://archive.ics.uci.edu/dataset/320/student+performance

Entwisle, D. R., Alexander, K. L. (1992). Summer setback: Race, poverty, school composition, and mathematics achievement in the first two years of school. American Sociological Review, 72-84.
Hill, N. E., Tyson, D. F. (2009). Parental involvement in middle school: a meta-analytic assessment of the strategies that promote achievement. Developmental psychology, 45(3), 740.

Amato, P. R. (2010). The marriage-go-round: The state of marriage and the family in America Today.

Henderson, A. T., Mapp, K. L. (2002). A New Wave of Evidence: The Impact of School, Family, and Community Connections on Student Achievement. Annual Synthesis, 2002.

Sternberg, R. J., Grigorenko, E. L. (2002). Dynamic testing: The nature and measurement of learning potential. Cambridge university press.

Ryan, R. M., Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American psychologist, 55(1), 68.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. Educational psychology review, 16, 385-407.

Mayer, J. D., Salovey, P., Caruso, D. R. (2008). Emotional intelligence: New ability or eclectic traits?. American psychologist, 63(6), 503.

Diemberger, L. (2021). Mindset Psychology of Success. Lorenz Diemberger.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and Regression Trees. Wadsworth Brooks/Cole Advanced Books Software.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

Goodfellow, I., Bengio, Y., Courville, A. (2016). "Deep Learning." MIT Press.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks." Advances in Neural Information Processing Systems, 25, 1097-1105.

Simonyan, K., Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.