Mathematics Performance in Secondary Education

Applied Multivariate Statistical Analysis

National Dong Hwa University, Spring 2025

Yao-Chih Hsu Xuan-Chun Wang Wen-Lee Sin

1 Datasets

1.1 Introduction

Our dataset was obtained from the *Math-Students Performance Data* (Shamim, 2025), a Kaggle dataset provided by Adil Shamim. It originates from the UCI Machine Learning Repository (Cortez, 2008) and was initially introduced by P. Cortez and A. Silva in their study titled *Using Data Mining to Predict Secondary School Student Performance* (2008).

In this dataset, the variables `G1`, `G2`, `G3`, and `absences` were directly provided by the school, while the remaining variables were collected through questionnaires. As a result, the dataset contains a large number of categorical variables.

1.2 Variables

The following table lists the attributes in our dataset along with their descriptions.

Attribute	**Description**	**Attribute**	**Description**
sex	student's sex	famsup	family educational support
age	student's age	activities	extra-curricular activities
school	student's school	paidclass	extra paid classes
address	student's home address type	internet	Internet access at home
Pstatus	parent's cohabitation status	nursery	attended nursery school
Medu	mother's education	higher	wants to take higher education
Mjob	mother's job	romantic	with a romantic relationship
Fedu	father's education	freetime	free time after school
Fjob	father's job	goout	going out with friends
guardian	student's guardian	Walc	weekend alcohol consumption
famsize	family size	Dalc	workday alcohol consumption
famrel	quality of family relationships	health	current health status
reason	reason to choose this school	absences	number of school absences
traveltime	home to school travel time	G1	first period grade
studytime	weekly study time	G2	second period grade
failures	number of past class failures	G3	final grade
schoolsup	extra educational school support		

Table 1: Student Performance Dataset Attribute Description

1.3 Questions

This dataset contains numerous categorical variables, which could impact subsequent analyses, such as the construction of a correlation matrix.

To address this, we grouped the variables in a way that preserves the information while integrating and reducing dimensionality.

Our goal is to analyze `G3` based on the other variables in the dataset, in order to identify which ones have an impact on it.

2 References

[1] Cortez, P. (2008). *Student Performance* [Dataset]. UCI Machine Learning Repository. `https://doi.org/10.24432/C5TG7T`

[2] Cortez, P., & Silva, A. M. (2008). Using data mining to predict secondary school student performance.

[3] Shamim, A. (2025). *Math students performance data*. Kaggle. `https://www.kaggle.com/datasets/adilshamim8/math-students`

3 Appendix

This is a part of the dataset we used, sourced from Kaggle (Shamim, 2025).

Index	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob
0	GP	F	18	U	GT3	A	4	4	at_home	teacher
1	GP	F	17	U	GT3	T	1	1	at_home	other
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
397	MS	M	19	U	GT3	T	4	4	teacher	other
398	MS	M	18	U	GT3	T	4	4	teacher	at_home

Index	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	4	3	4	1	1	3	6	5	6	6
1	5	3	3	1	1	3	4	5	5	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
397	5	3	2	1	2	4	0	8	7	7
398	5	3	2	1	2	4	0	8	7	7

Table 2: Student Performance Dataset