Mathematics Performance in Secondary Education

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

June 12, 2025

Outline

Introduction

- Dataset: *Math-Students Performance Data* from Kaggle (Shamim, 2025).

Introduction

- Dataset: *Math-Students Performance Data* from Kaggle (Shamim, 2025).
- Variables `G1`, `G2`, `G3`, and `absences` were provided by the school.
- Remaining variables were collected via questionnaires and are mostly categorical.

Goal

Analyze `G3` to identify influential variables.

Data Grouping

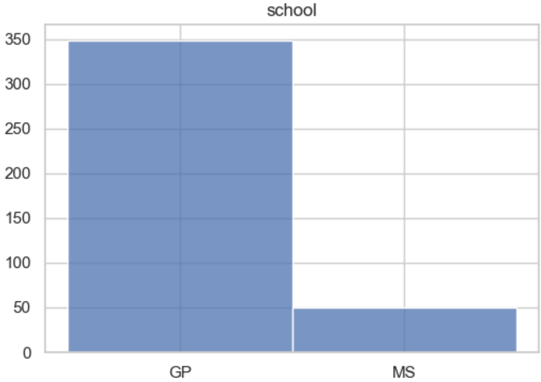
We grouped the data into the following:

Groups	**Variables**
support	schoolsup, famsup, paid
family	address, famsize, Pstatus, guardian, traveltime, famrel
parents	Medu, Fedu, Mjob, Fjob
performance	failures, studytime, absences
alcohol	Dalc, Walc, health
after_class	activities, freetime, goout
school_choice	reason, nursery, higher
score	G1, G2, G3

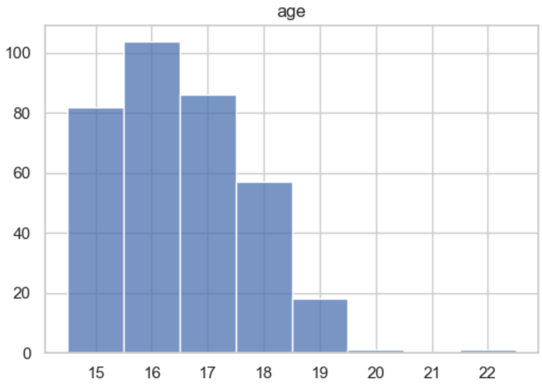
The variables not yet assigned to any group are:

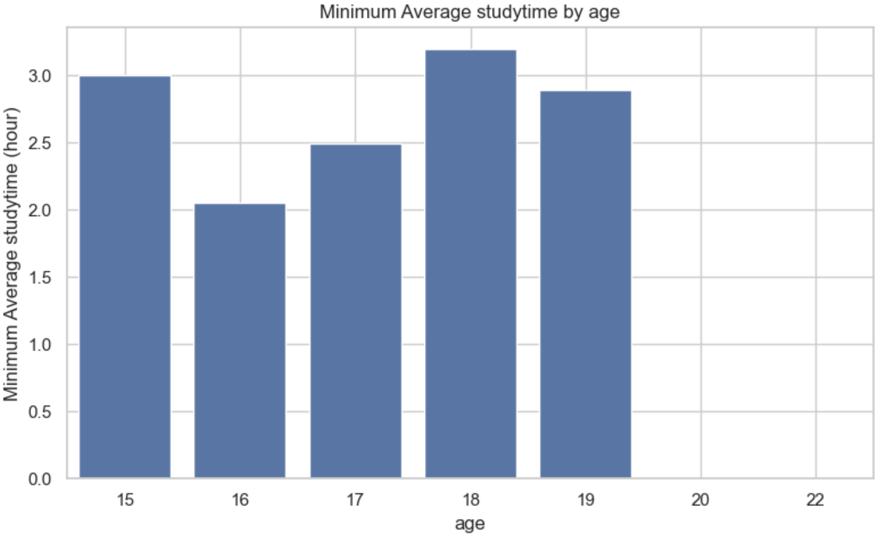
sex, age, internet, romantic.

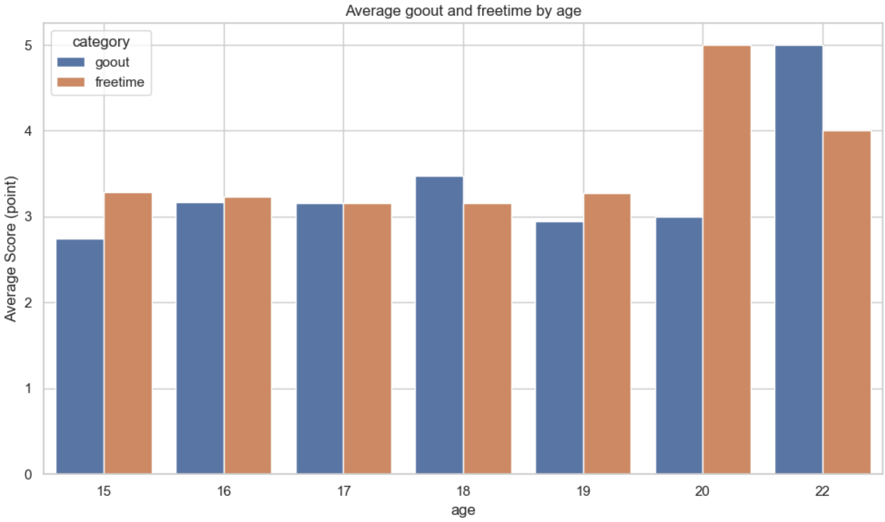
Data Visualization

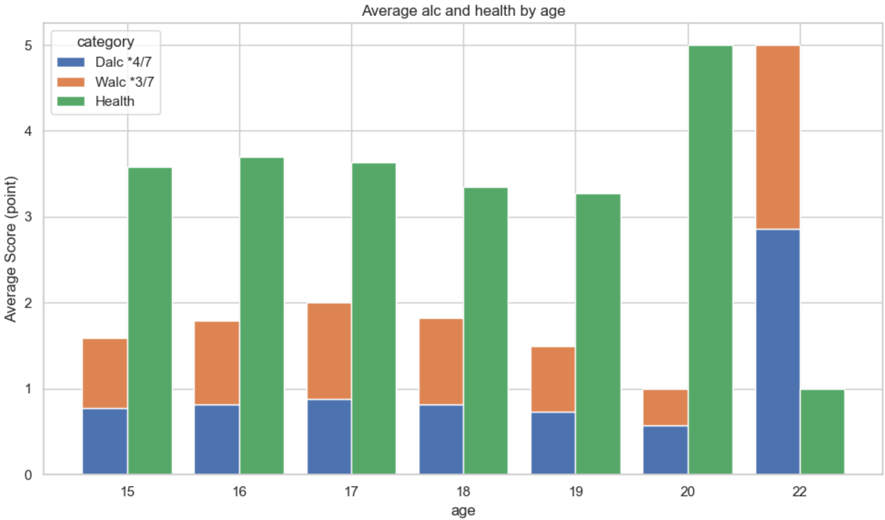


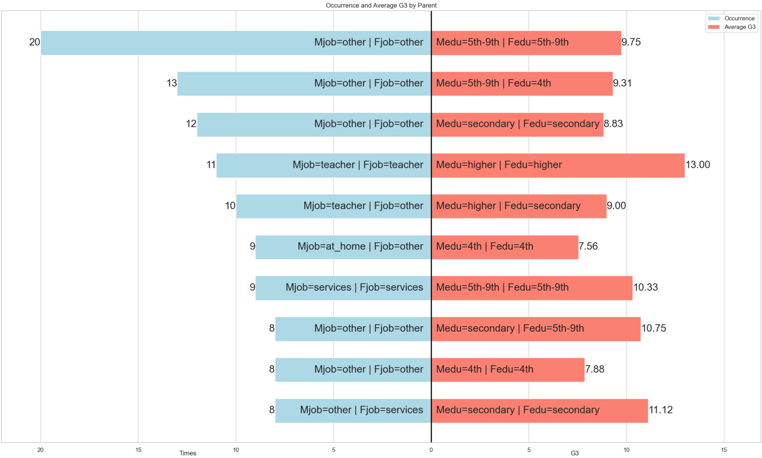
GP (Gabriel Pereira) **MS (Mousinho da Silveira)**

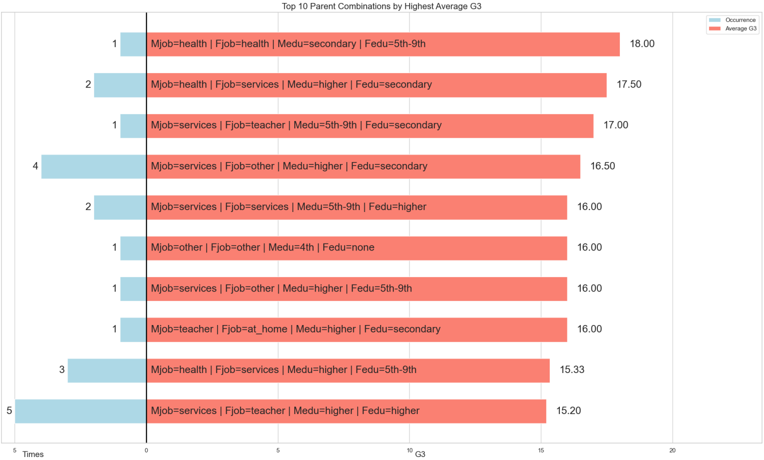

age
100
80
60
40
20
0
15
16
17
18
19
20
21
22


Minimum Average studytime by age
Minimum Average studytime (hour)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15
16
17
18
19
20
22
age

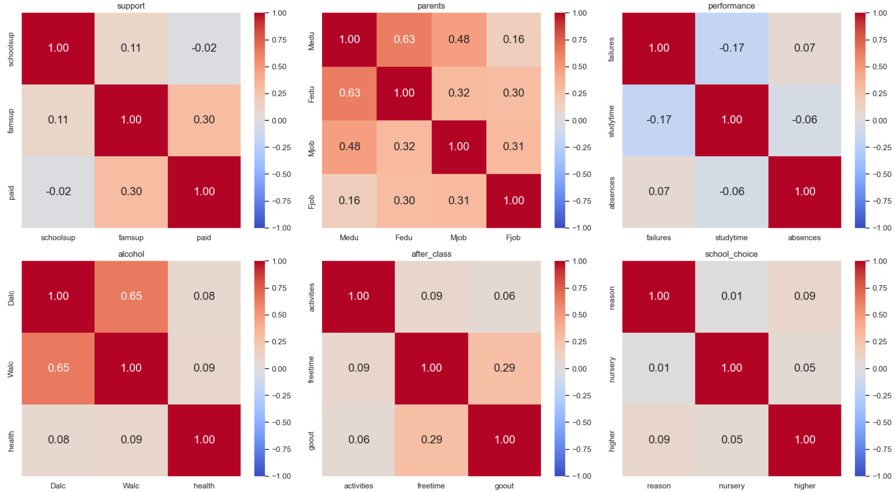

Average goout and freetime by age
category
goout
freetime
Average Score (point)
0
1
2
3
4
5
15
16
17
18
19
20
22
age


Average alc and health by age
category
Dalc *4/7
Walc *3/7
Health
Average Score (point)
0
1
2
3
4
5
15
16
17
18
19
20
22
age

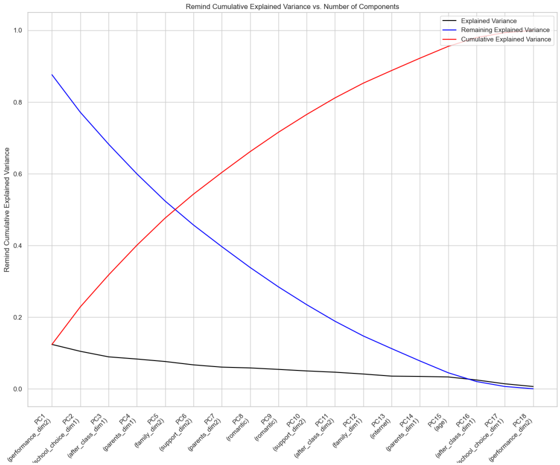

Occurrence and Average G3 by Parent
Occurrence
Average G3
20 Mjob=other | Fjob=other
Medu=5th-9th | Fedu=5th-9th 9.75
13 Mjob=other | Fjob=other
Medu=5th-9th | Fedu=4th 9.31
12 Mjob=other | Fjob=other
Medu=secondary | Fedu=secondary 8.83
11 Mjob=teacher | Fjob=teacher
Medu=higher | Fedu=higher 13.00
10 Mjob=teacher | Fjob=other
Medu=higher | Fedu=secondary 9.00
9 Mjob=at_home | Fjob=other
Medu=4th | Fedu=4th 7.56
9 Mjob=services | Fjob=services
Medu=5th-9th | Fedu=5th-9th 10.33
8 Mjob=other | Fjob=other
Medu=secondary | Fedu=5th-9th 10.75
8 Mjob=other | Fjob=other
Medu=4th | Fedu=4th 7.88
8 Mjob=other | Fjob=services
Medu=secondary | Fedu=secondary 11.12
20 15 10 5 0 5 10 15
Times
G3


Top 10 Parent Combinations by Highest Average G3
Occurrence
Average G3
1
Mjob=health | Fjob=health | Medu=secondary | Fedu=5th-9th
18.00
2
Mjob=health | Fjob=services | Medu=higher | Fedu=secondary
17.50
1
Mjob=services | Fjob=teacher | Medu=5th-9th | Fedu=secondary
17.00
4
Mjob=services | Fjob=other | Medu=higher | Fedu=secondary
16.50
2
Mjob=services | Fjob=services | Medu=5th-9th | Fedu=higher
16.00
1
Mjob=other | Fjob=other | Medu=4th | Fedu=none
16.00
1
Mjob=services | Fjob=other | Medu=higher | Fedu=5th-9th
16.00
1
Mjob=teacher | Fjob=at_home | Medu=higher | Fedu=secondary
16.00
3
Mjob=health | Fjob=services | Medu=higher | Fedu=5th-9th
15.33
5
Mjob=services | Fjob=teacher | Medu=higher | Fedu=higher
15.20
5 Times
0
5
10 G3
15
20

Multidimensional Scaling



PCA


Remind Cumulative Explained Variance vs. Number of Components
Explained Variance
Remaining Explained Variance
Cumulative Explained Variance
Remind Cumulative Explained Variance
0.0
0.2
0.4
0.6
0.8
1.0
PC1 (performance_dim2)
PC2 (school_choice_dim1)
PC3 (after_class_dim1)
PC4 (parents_dim1)
PC5 (family_dim2)
PC6 (support_dim2)
PC7 (parents_dim2)
PC8 (romantic)
PC9 (romantic)
PC10 (support_dim2)
PC11 (after_class_dim2)
PC12 (family_dim1)
PC13 (internet)
PC14 (parents_dim1)
PC15 (age)
PC16 (after_class_dim1)
PC17 (school_choice_dim1)
PC18 (performance_dim2)

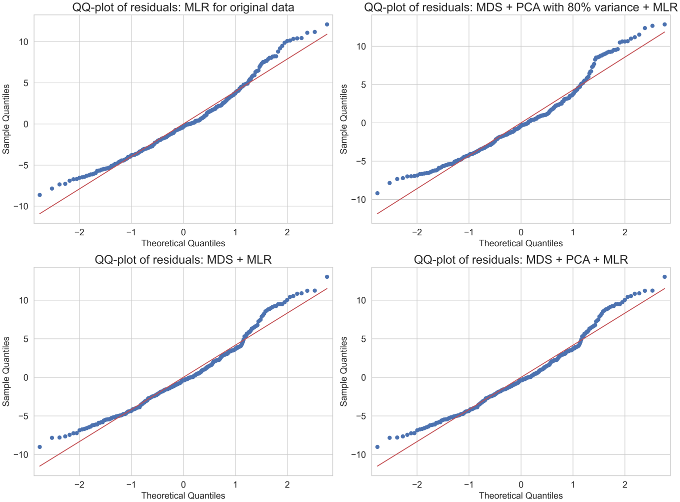
Variables of MLR

We compared four models:

- Original + MLR
- MDS + MLR
- MDS + PCA + MLR
- MDS + PCA 80% + MLR

The variables in the dataset transformed by MDS are as follows:

Q-Q plots of MLR

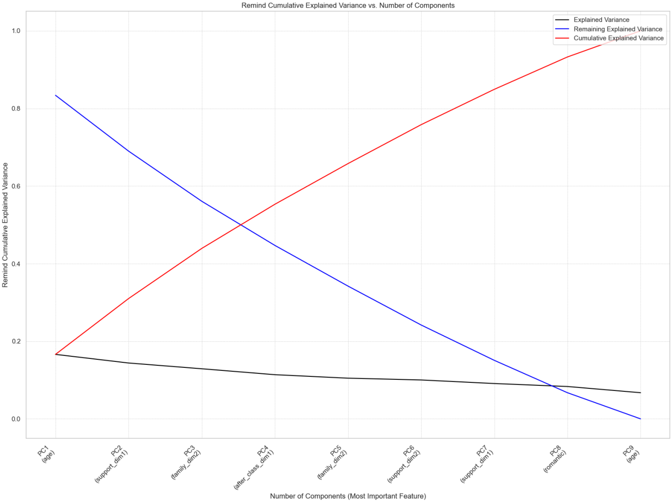


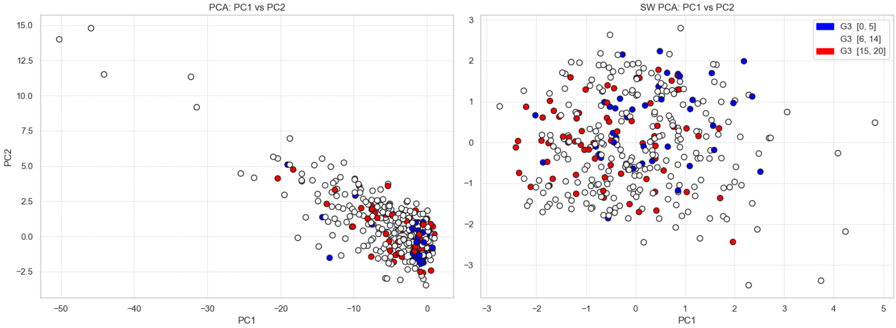
Variables of Reduced MLR

Models	Selected Features
Original + MDS + Stepwise + MLR	failures, goout, parents_dim1, after_class_dim1 family_dim2, romantic support_dim2, after_class_dim2 support_dim1, age performance_dim1
MDS + PCA + Stepwise + MLR	parents_dim1, after_class_dim1 family_dim2, romantic support_dim2, after_class_dim2 support_dim1, age performance_dim1

Table 1: Selected Features by Different Methods

Original data + MDS + Stepwise + PCA





PC1 (`school_choice_dim1`)
PC2 (`after_class_dim1`)

PC1 (`parent_dim2`)
PC2 (`romantic`)

Conclusion

Models	R^2	Adj. R^2	MSE	P-value	AIC
MLR for original data	**0.27**	**0.19**	**62.58**	0.9157	2015.10
MDS + MLR	0.19	0.14	77.58	0.7494	2023.96
MDS + PCA + MLR	0.19	0.14	77.58	0.6143	2023.96
MDS + PCA with 80% variance + MLR	0.14	0.11	92.69	0.9442	2031.05
Original data + MDS + Stepwise + MLR	0.20	**0.19**	293.55	**0.0367**	**1993.82**
MDS + PCA + Stepwise + MLR	0.17	0.15	140.12	0.5113	2013.68
Paper Proposed	0.17	0.16	216.03	0.4624	2005.069

Table 2: Comparison of model performance metrics

The variables selected by the paper:

`absences`, `schoolsup`, `higher`, `failures`, `Mjob`

Conclusion

Model	TOP1	TOP2	TOP3
MLR for original data	**failures**	**schoolsup**	**paid**
MDS + MLR	romantic	support_dim2	support_dim1
MDS + PCA + MLR	romantic	support_dim2	age
MDS + PCA with 80% variance + MLR	romantic	support_dim2	parents_dim1
Original data + MDS + Stepwise + MLR	**Mjob_at_home**	**failures**	**Mjob_other**
MDS + PCA + Stepwise + MLR	age	support_dim1	family_dim2
Paper Proposed	**higher**	**schoolsup**	**failures**

Table 3: Most important feature to each model

Conclusion

The variables we selected are:

`Mjob`, `failures`

References

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

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

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

Thanks for listening!