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 report, we want to identify the factors (variables) that influence students' final scores `G3`.

1.2 Data Processing

Because the number of students in MS(50) is much smaller than that in GP(349), we decided to exclude the MS data and focus our analysis solely on the GP data.

Because `G1` and `G2` are highly correlated with `G3`, we decided to exclude `G1` and `G2` from the analysis.

1.3 Data Grouping

In Table 1, we group the data:

Table 1: Group all variables

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`.

2 Data Visualization

We want to find the top 10 most frequent combinations of `Medu`, `Fedu`, `Mjob`, `Fjob`, and compute the average `G3` for each combination.

In both plots below, the blue bars on the left show the number of times each combination appears, and the red bars on the right show the average `G3` for each combination.

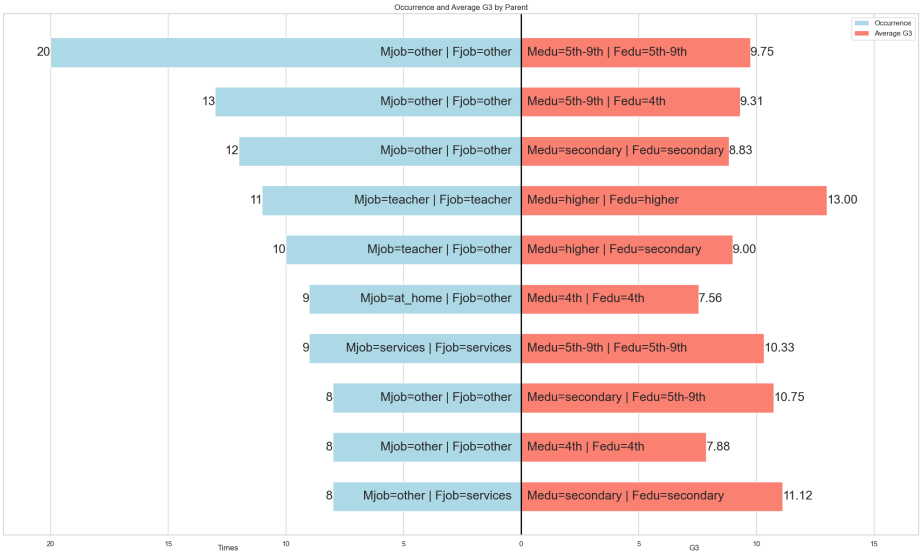

Figure 1: Top 10 combinations for parents education and job, ordered by frequents.

After we computed the average `G3` for the top 10 most frequent combinations, we want to find the top 10 combinations with the highest average `G3`.

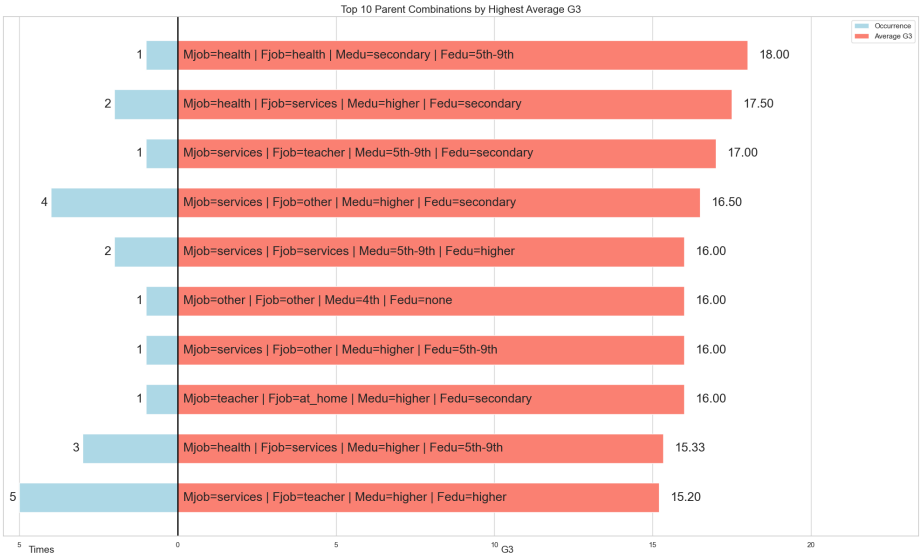

Figure 2: Top 10 combinations for parents education and job, ordered by `G3`.

We can see that when both `Medu` and `Fedu` are higher, and both `Mjob` and `Fjob` are teachers, the students' average `G3` is 13.00; when `Fjob` is other, the average `G3` is 9.00; and when `Mjob` is services, the average `G3` is 15.20.

Therefore, we can see that, at the same education levels, different `Mjob` and `Fjob` combinations also affect `G3`.

3 Dimensionality Reduction

3.1 Multi-dimensional Scaling

The original dataset contains 33 variables, and the pairwise correlations among them are generally negligible (close to zero). Therefore, we first grouped together the variables that appeared to be related, then applied MDS separately to each group. By comparing the distances between samples within the same group, we were able to reduce each group's dimensionality from a high-dimensional space down to two dimensions, dramatically cutting the number of variables while enhancing each group's explanatory power for `G3`. After MDS, we performed PCA on the resulting low-dimensional representations. This PCA was conducted with respect to each group's variables and `G3`, with the aim of identifying which variables most strongly influence the `G3`.

3.2 Stepwise

We applied stepwise method to selected the variable that it is most important for regression analysis or prediction. The stepwise regression comes with three parts:

1. **Forward Stepwise**:

As the name suggest that in the initial step, there is no variable along the regression. Therefore, we later add the variables gradually then we combine those selected variables as predictors.

2. **Backward Elimination**:

We listed all those predictors in initial step. Then, we eliminate the variable one by one then, check whether the model is hold up.

3. **Bidirectional Elimination**:

It adds predictors like forward selection but also kicks out any that become irrelevant along the way.

In this project we applied Bidirectional elimination to select over variables for reduced model.

3.3 Principal Component Analysis

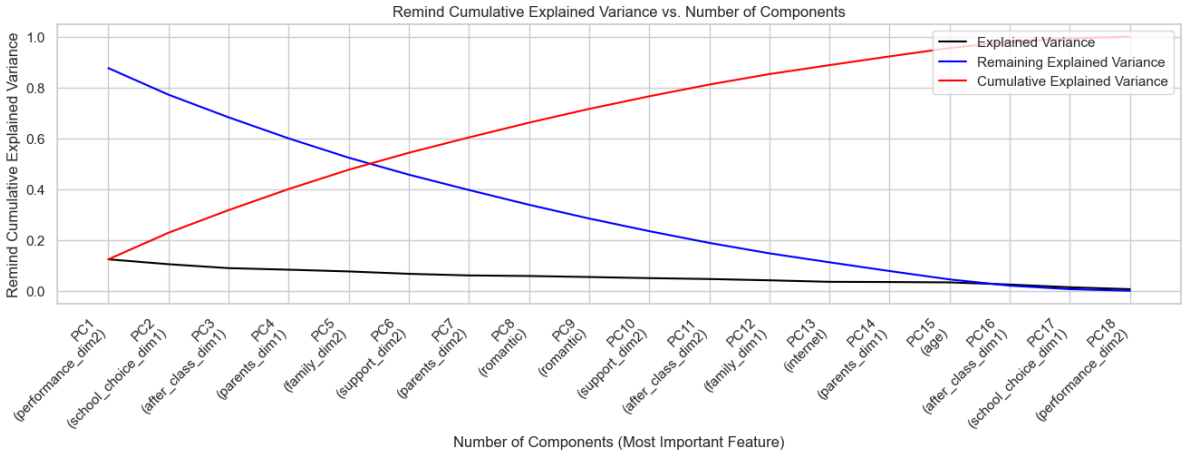

Figure 3: PCA with original data

Figure 3 is the result of performing MDS on the original data and applying PCA.

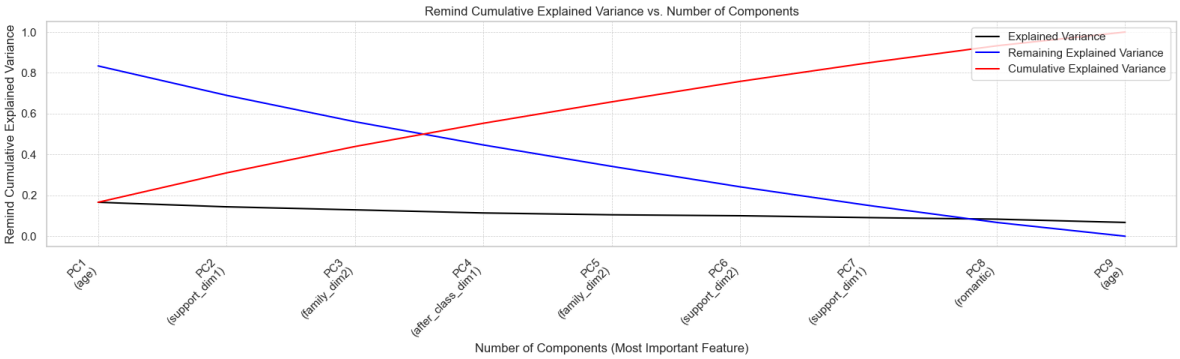

Figure 4: PCA with original data + MDS (Reduced)

Figure 4 is the result of performing MDS on the original data, then conducting stepwise variable selection, and finally applying PCA.

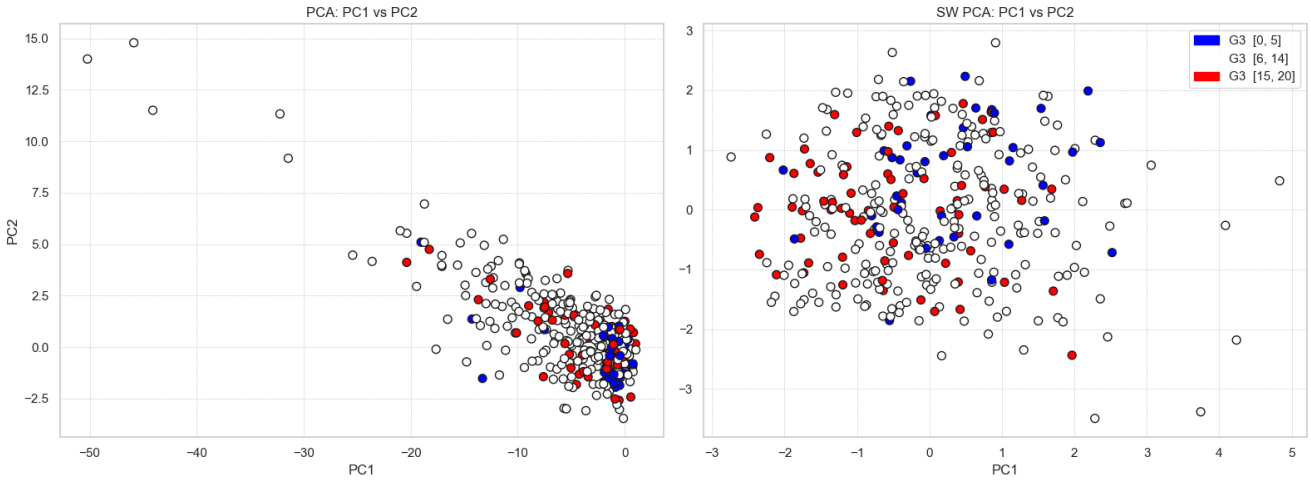

Figure 5: PC1 and PC2 with `G3`

Figure 5 is a PCA score scatter plots showing samples in the PC1 and PC2 plane, with points colored by their `G3`, where high-score range (15 to 20) in red and low-score range (0 to 5) in blue.

PCA: PC1 vs. PC2

The high-score, mid-score and low-score observations are all mixed together, with no clear separation along either PC1 or PC2. This shows that when PCA is applied to all original variables, even though the first two components capture the largest share of total variance, they are not necessarily the ones most related to `G3`.

SW PCA: PC1 vs. PC2

Compared to the left plot, the first two principal components from the stepwise-selected PCA exhibit an initial separation among the different `G3` ranges. This indicates that by first using stepwise regression to pick variables highly correlated with `G3` and then performing PCA, PC1 and PC2 gain greater discriminative power over the score distributions:

- The high-score group falls mostly in the positive PC1 region.

- The low-score group clusters in the negative PC1 region.

4 Regression

We compared six models:

Table 2: Selected Features by Different Methods

Methods	**Process**
MLR for original data	Perform multiple linear regression using all original variables.
MDS + MLR	Group related variables and apply MDS to each group to reduce to two dimensions, then fit MLR on the combined MDS coordinates.
MDS + PCA + MLR	After MDS reduction, concatenate all two dimensions coordinates, perform PCA on that matrix, and fit MLR using all principal components.
MDS + PCA with 80% variance + MLR	After MDS reduction, perform PCA retaining components that explain 80% of variance, then fit MLR on those selected components.
Original data + MDS + Stepwise + MLR	After MDS reduction, apply stepwise regression to select features before fitting MLR.
MDS + PCA + Stepwise + MLR	After MDS and PCA reduction, apply stepwise regression to the principal components to select significant features, then fit MLR.

5 Conclusion

The following tables are our conclusion for the six models, we also compared the recommend variables in the paper(2008):

Table 3: Comparison of model performance metrics

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

The variables selected by the paper: absences, schoolsup, higher, failures, Mjob.

Table 4: Most important feature to each model

Models	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**

In Table 4, we have:

- MLR for original data has best performance in R^2, Adj. R^2, MSE.
- Original data + Stepwise + MLR has best performance in P-value, AIC .

In order to select the variables that have the greatest influence on G3, we selected the model with the best AIC performance. Based on the **Original data + MDS + Stepwise + MLR** model, the variables identified as potentially affecting G3 are Mjob and failures.

6 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