**Paper topic : Comprehensively analyze and compare various machine learning algorithms (e.g., RF, SVM, LR, etc.) on predicting student math achievement.**

**Main research questions:**

1. Which method has the best classification performance based on model-performance measures?

2) Which factors are found significant on *student math achievement*; what is their order of importance?

**Data analysis plan**

Step 1: Preparing the data and selecting factors (dependent and independent).

* The dependent variable will be a binary variable (yes, no). I will recode the math score to binary.

Step 2: Creating the training and testing data.

* Classification is performed based on the dependent variable. The dataset is split into two sets: training (75% of the data) and testing (25%).

Step 3: Designing and evaluating classification algorithms.

* The classifiers that will be used include: Decision Tree, Support Vector Machine, Naive Bayes/Gaussian, Random Forest, logistic regression, neural networks, Ridge regression/lasso/Elastic Net regression, Boost (Adaboost, gradient descent boost, XGboost).

* After reading papers, I found that decision trees, Support Vector Machine, Naive Bayes/Gaussian, Random Forest, logistic regression, and neural networks are most commonly used.
* The algorithms’ classification results will be assessed based on the different performance measures: TP rate, FP rate, precision, and MCC; ROC Area and PRC Area; F-1, Kappa statistics (κ).
* The best performing algorithm will be decided according to the quantities of performance measures

Step 4: Testing the differences in performance measures.

* Friedman’s two-way analysis of variance will be used to examine whether these performance measures statistically differ among the algorithms.

Step 5: Find the most effective factor on outcome variable.

* The methods of correlation-ranking filter, gain ratios, and info gain will be used to find the factors’ order of importance.