**1. Classification vs Regression**

Your goal is to identify students who might need early intervention - which type of supervised machine learning problem is this, classification or regression? Why?

Ans: This is a classification machine learning problem because desired output for each student is: whether she is going to pass or fail.

**2. Exploring the Data**

Can you find out the following facts about the dataset?

* Total number of students
* Number of students who passed
* Number of students who failed
* Graduation rate of the class (%)
* Number of features

Use the code block provided in the template to compute these values.

**3. Preparing the Data**

Execute the following steps to prepare the data for modeling, training and testing:

* Identify feature and target columns
* Preprocess feature columns
* Split data into training and test sets

Starter code snippets for these steps have been provided in the template.

**4. Training and Evaluating Models**

Choose 3 supervised learning models that are available in scikit-learn, and appropriate for this problem. For each model:

* What is both the theoretical space complexity to represent the model and the time for the algorithm to make a prediction? You can either provide the big-O notation, or list several the of major features that may affect the algorithm and state if the largest driving factor is constant, linear, logrithmic, polynomical, etc in nature.
* What are the general applications of this model? What are its strengths and weaknesses?
* Given what you know about the data so far, why did you choose this model to apply?
* Fit this model to the training data, try to predict labels (for both training and test sets), and measure the F1 score. Repeat this process with different training set sizes (100, 200, 300), keeping test set constant.
* Produce a [table](https://docs.google.com/document/d/1Goxw-6M0umOCokCFqTr-g7SU_5f6MIm_wqXmh9Cnjhw/pub) showing training time, prediction time, F1 score on training set and F1 score on test set, for each training set size.

**Note**: You need to produce 3 such tables - one for each model.

**5. Choosing the Best Model**

Based on the experiments you performed earlier, in 1-2 paragraphs explain to the board of supervisors what single model you chose as the best model. Which model is generally the most appropriate based on the available data, limited resources, cost, and performance?

In 1-2 paragraphs explain to the board of supervisors in layman’s terms how the final model chosen is supposed to work (for example if you chose a decision tree or support vector machine, how does it make a prediction).

Fine-tune the model. Use gridsearch with at least one important parameter tuned and with at least 3 settings. Use the entire training set for this.

What is the model’s final F1 score?