Data Exploration Findings:  
a. Checking first year persistence class (specially checking the imabalnces)

**About 79.2% of the students in the dataset did persist (represented by the value 1).**

**About 20.8% of the students did not persist (represented by the value 0).**

So, roughly 4 out of 5 students in this dataset continued their studies after the first year. This is useful to know because it shows the target variable ('First Year Persistence') isn't perfectly balanced; there are many more students who persisted than those who didn't.  
  
Possible Solution: SMOTE the goat  
  
b. Checking the number of missing values in each column.  
This is hard, the high school average mark has the most number of missing values, it is the interesting and usefull one, sooooo….. let’s using another solution to preserve it.  
  
Possible solution: Median Imputation + Missing Indicator Column, this is big dawg --

Start with calculating the median of the non-missing 'High School Average Mark' values, create that new binary indicator column, and then fill the NaNs in the original column with the calculated median. This gives you a complete dataset to move forward with while acknowledging the original missingness.

c. Checking the class balance for each columns (This is super fun)  
so this number 6 is only the values for school, this dataset only has the values for engineering students.  
  
Overall, heavy imbalance for most of the columns  
  
The target class for each of the problems:

1. Problem: Persistence (1st year persistence) - Classification

Target Column: First Year Persistence

(Required Model): Neural Network (Classification)

(Other Common Models):

1. Logistic Regression (Good baseline)
2. Support Vector Machines (SVM)
3. Random Forest Classifier
4. Gradient Boosting Machines (XGBoost, LightGBM, CatBoost) - Often perform very well on tabular data.  
     
   2. Problem: Academic performance (Predicting Second Term GPA) - Regression

Target Column: Second Term Gpa

(Required Model): Neural Network (Regression)

(Other Common Models):

1. Linear Regression (Good baseline)
2. Support Vector Regression (SVR)
3. Random Forest Regressor
4. Gradient Boosting Machines (XGBoost, LightGBM, CatBoost) - Also strong for regression on tabular data.