Final Project – Predicting whether a student will under-perform

- Due Date: 5/19/2020 23:59 pm
- Goal is to apply as much as what you had learned to solve a classification problem
- Dataset is based on a dataset of high-school performance on Math but with some modification https://www.kaggle.com/impapan/student-performance-data-set
- Use the modified version, student_performance_modified.csv, that I uploaded to Blackboard.
- Use the Final Project notebook as your starting point.
- Use any one of the models (Logistic, SVM, Decision Tree and Random Forecast, Naïve Bayes) that we have covered in class. You only need to pick one.
- Do not use any models that we have not covered in class.
- The target variable is the “underperformance”
- Math_G1 is the Grade after first grading period. It should be highly correlated with the final grade (ie under performance or not), but it won’t be available at the beginning of the school term. So it may not be available yet.

Final Project Grading Criteria

- You need to create 3 sets of models
  - Without using either the Math_G1 and Language_G1
  - Use both Math_G1 and Language_G1
  - Use only Math_G1 but without Language_G1
- You do NOT need to use ALL the variables in the dataset, but you should pick or have tried or investigated a good number of the variables.
- Comment on your final model, explain whether the Language Grade will be helpful to catch underperformance in the final Math grade
- Grading will be based on whether you have done it “right”, not the actual performance of the final model
- Make sure you have
  - checked missing values, removed outliers
  - performed basic exploration of relationship
  - plots and graphs
  - separated data set into training and testing
  - Setup dummy variables for categorical variables,
  - normalize numerical features if needed
  - tried at least two models and checked their model performance
  - performed cross-validations

Final Project

OK to discuss among yourselves, but do not share code

Good Luck and Have Fun!!!