CS 418 FINAL PROJECT

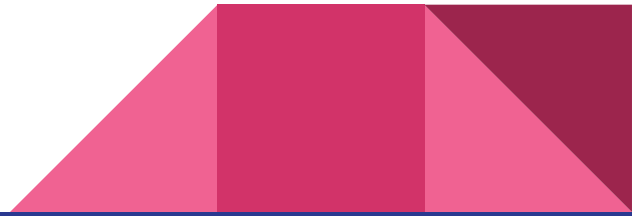
By: Saikrishna Yadavalli, Saivikas Nethi, & Kaushal Mamgain

I. Problem Statement

- Predict a secondary school student's final grade given the student's information from two Portuguese secondary schools.

II. Data Sources

- Two multivariate datasets for both Portuguese and math courses: (1) **student-por.csv** and (2) **student-mat.csv**.
- Both datasets each have 649 observations w/ 33 attributes: age, sex, address, parents' education, parents' employment, final grades, etc.



III. Brief Description of Data Science Solution

- Prepared data by merging both datasets to include info about each of 382 students in each row.
- Applied dimensionality reduction to remove irrelevant or redundant attributes.
- Built linear regression and classification models.
- Used K-Nearest neighbors, Naive Bayes, a linear support vector machine (SVM), and a kernel SVM as classification techniques.

IV. Results

After doing an exploratory and regression analysis on the data, it results that using attributes like sex, internet, activity makes the prediction better for the grades that the students get. For this linear regression model we got a score of 0.8667.

The classifier that works best for our data is Kernel based Support Vector Machine which gives us a F-1 Score of 0.94972 for one class and 0.30769 for another

V. Conclusions

- From the results and from exploring the dataset, we came to the conclusion that predicting the final grades can be most accurately predicted based on:
  - The gender of the student
  - The amount of time they study outside of class
  - Whether they have access to internet at home or not