## Capstone Project  1 - Proposal

# Factors Influencing ACT Scores

**1. Problem Statement**

The ACT test is designed to assess high school students' general educational development and their ability to complete college-level work. The ACT test covers four subject areas: English, mathematics, reading, and science. Each subject area test receives a score ranging from 1 to 36. The data is obtained from the California Department of Education in order to address these questions. (1) Which public high school's students received the highest overall  score? (2) Which county has the highest performing schools? (3) At which subject tests are the student performing better? (4) Is there a linear increase at  the success level of the  schools with time? (5) Possible questions to add later on: Is there a correlation between the success and the student demographics? (6)  Does income have an effect on the success level?  ( Linking the information from https://www.greatschools.org/california)

**2. Target Client**

High school principles, as  instructional  leaders  and  decision makers, and educators are the major  clients  who can benefit from the outcomes of the study. High  schools can use  the study to identify subject areas which need further  improvement  and plan  changes in  improvements in  curriculum, provide  guidance and assistance  for students accordingly. Colleges can use the study to evaluate the selectivity of ACT scores to admit high performing students in college.

**3. Dataset**

The data to be used is obtained from the California Department of Education.
( https://www.cde.ca.gov/ds/sp/ai/ )  2017-18 ACT (XLS), 2016-17 ACT(XLS), 2015–16 ACT (XLS), 2014–15 ACT (XLS), 2013–14 ACT (XLS))  Each dataset consists of a row for every accredited high school in California with its code number, school name, county name, enrollment based on the number of students in grade twelve, number of students, the average english, math and reading scores, the number of students who scored 21 points or more, the percent of students who scored 21 points or more.  The data covers a  period  from 2013 up through 2018.

**4. Approach to solve the  problem**

The first step will be preparing the data for analysis which includes identifying the common data problems such as missing data, inconsistent column names, duplicate data and  outliers.
After cleaning the data, the next step includes performing exploratory data analysis and building some visualizations using mathplotlib or seaborn.
Since there are five different datasets to combine the data concatenating techniques will be applied.

Once this is completed, statistical technique can be done across the variables included in the data set to build  a predictive  model.

**Project Outcomes**

Project outcomes will include a jupyter notebook file containing  the associated code, a report and a presentation.