**Intro to Data Science**

**Project Proposal**

**March 29, 2017  
Andrea Chamorro, Mariam Khan, Janani Shankaran**

**(1) Hypothesis and research questions**

The question we are addressing is how risky behaviors such as gang membership, drug use, alcohol consumption, and sexual activity, among others, impact grades for students at Somerville High School in Massachusetts. Specifically, what is the impact of certain behaviors and activities on the student GPA?

Our hypothesis is that an increase in risky behavior is associated with a decrease in GPA.

**(2) Data**   
We will be using City of Somerville’s Youth Risk Behavior Survey, an annual student survey administered in Somerville High School. The dataset includes over 6,374 survey responses between 2002 to 2014. Variables include race, drug use, sexual behavior, dietary behavior, physical activity and self-image. The dataset can be accessed here: <http://bit.ly/2nRvJYa>.

**(3) Data experiment design**

A model will be constructed to predict each respondent’s GPA based on risky behaviors, such as whether the respondent has used cocaine or other types of drugs. While the GPAs of respondents have been provided, they are self-reported, which may be a concern. In addition, GPA is not a numeric variable; instead, it is a categorical variable. More specifically, it provides whether respondents had received “mostly A’s,” “mostly B’s,” “mostly C’s,” “mostly D’s,” or “mostly F’s.”

**(4) Proposed processing methods**

In the dataset, they are several categorical variables, however, many of them need to be recoded as binary variables. For example, the survey asks respondents how many times they have used cocaine in the past 30 days. Respondents had the option to choose “0 times”, “1 or 2 times”, “3 to 9 times”, and “10 to 19 times.” In this case, this variable would be recoded as a binary variable, in that respondents who have tried cocaine at least once, would be coded as a value of one. Similarly, those who have not tried cocaine, would be coded as a value of zero.

**(5) Analytical methods**

The data will be divided into a 70-15-15 partition. The following techniques which will be considered for this project includes: Decision Trees, Random Forest, and Ordered Probit. These methods will help determine which predicting model is optimal. It is acknowledged that Random Forest may overestimate some of our values, however, for the sake of comparison, it will be utilized. When deciding which technique to implement, the AUC value and the Mean-F1 score would be compared.

**(6) Ethical considerations**  
Risk behavioral questions are sensitive in nature. It is difficult to ask students to provide serious responses regarding drugs, alcohol, sex and crime. In addition, the self-reported nature of these questions are cause for concerns regarding any measurement error. However, these type of questions provide insight into potential school and/or education policy changes. To address teen pregnancies, for example, increased funding towards sex education programs can be an option. Any school policy changes at the student-level might infringe upon the anonymity and confidentiality of a risk behavior survey. Maintaining the anonymity and confidentiality are key to getting accurate data in the future. Since this project will focus on one high school in Massachusetts, Somerville High School, we have to take into account the demographic characteristics of the City of Somerville. Being close to the City of Cambridge and the City of Boston might have some influence on its demographic characteristics.