**Title:** Predicting Cannabis Use in Youth

**Mentors:** Dr. Kit Elam, Dr. Valentin Dinu

**Background:** Cannabis is the most commonly used drug in the United States. Its use is widespread among adolescents and young adults. According to the National Institute on Drug Abuse, 7.8 percent of college-aged young adults are using cannabis, which is the highest it has been since the early 1980s. Previous studies have shown the influence of genetic and environmental factors on cannabis use. Given the genetic predisposition of individuals, which can be represented by polygenic risk scores, and their family dynamic, it may be possible to quantify the likelihood of an individual to engage in cannabis use. To what extent genetic and environmental factors can predict cannabis use in youth was explored using machine learning techniques: logistic regression and random forest. The performance of both methods was compared and evaluated in this project.

**Methods:** The data consists of 364 total observations. Variables related to genetic predisposition and family dynamic were used to predict the following: max cannabis use, cannabis use in the past year, cannabis use in the past 3 months, and general cannabis use (“yes” or “no”). The data was preprocessed prior to analysis. First, variables which had data for only 70 percent or more of the total observations were discarded. Variable types were explicitly declared to avoid confusion or error during analysis. All explanatory categorical variables were converted to dummy variables. Missing data imputation was carried out using the k-nearest neighbors method, where values were generated based on the three most similar cases (k=3). The data was standardized to have mean equal to zero and standard deviation equal to one. Standardized data was used only for logistic regression. Fifteen percent of the data was left out for testing. Variable selection was performed in a similar manner for both logistic regression and random forest. L1 regularization or LASSO method was applied to logistic regression in order to shrink the coefficients of less important variables to zero. Logistic regression with L1 penalty was fitted to the data repeatedly until maximum cross validation accuracy had been reached. Similarly, with random forest, variables with zero coefficients were continually removed until the set of variables providing maximum cross validation accuracy had been found. The selected predictors and model parameters were used to predict on the testing data.

**Results/Discussion:** Random forest had a tendency to skew the model in favor of the class containing the highest number of observations. Logistic regression was not as affected by imbalanced data. To address this, weights were added to the classes based on the number of observations in each class. In future studies, a more robust approach for selecting class weights can be explored. Overall, random forest afforded only a slight improvement in accuracy compared to logistic regression as shown in the table below. The reason for this along with interpretation of the models, their parameters, and selected variables will be discussed in the final paper.