Team name: The P-Hackers
Team members: Oscar Fawcett, Yesh Munagala, Ian Morton, Weston Murdock, and Emily Przykucki
Date: 04/07/2022

Team roles for this report (write down name):

Facilitator(s): Weston Murdock

Recorder(s): Oscar Fawcett

Deliverer(s): Ian Morton

Planner(s): Yesh Munagala

Team Contact(s): Emily Przykucki

0. **Describe briefly what the main goal of your team is (so the peer reviewer has some context). E.g. we are working on image classification for blah de blah. Our goal is blah de blah etc. In the initial part of the semester before your proposal it is ok to put down "we are still coming up with ideas on team project".**

The purpose of our project is to predict on drug time period usage based on personality and demographic data. Specifically, we have response variables such as cannabis which has values CL0, CL1, CL2, CL3, CL4, CL5, and CL6 each corresponding to the following:

CL0 – Never Used

CL1 – Used over a Decade Ago

CL2 – Used in Last Decade

CL3 – Used in Last Year

CL4 – Used in Last Month

CL5 – Used in Last Week

CL6 – Used in Last Day

We have variables in this form for drugs from heroin to chocolate. Our personality data is of the Big Five personality test for each respondent for our dataset.

I. **What was done during the report period regarding the project:** If you want to include code include this in the Appendix. Describe what the group did (including contributions of individual team members) with regards to the group project during this report period. Give enough details so I understand what you folks have been doing over the week. Include dates of your meeting(s) and who met on these days.

During the reporting period we created nine new response variables: five which indicate the time period in which the respondent last consumed a drug from the US scheduling categorizations (i.e., sched_1 for schedule 1 drugs, sched_2 for schedule 2 drugs, etc), and four which indicate the time period in which the respondent last consumed stimulant, depressant, hallucinogen, or stimulant not including caffeine. Then, we applied various models to these response variables.

Oscar: Created the new response variables. Implemented KNN models to the response variables using the personality data. Each model received a test accuracy ~ 0.45 which, though better than guessing randomly, is need of improvement.

Weston: Worked on creating random forests for each response variable.

Ian: Worked on creating LDA/QDA models for each response variable.

Yesh: Created logistic regression models for each response variable. Specifically, ordinal logistic regression with a combination of personality and demographic predictors. This resulted in test accuracy ~ 0.24 which needs improvement, even though it is better than a random guess.

Emily: Created a naïve Bayes model for each response variable by converting the responses from multinomial to binomial. These models yielded accuracies ~0.85, but these are artificially inflated since the conversion to binary variables binned most observations into the same class.

A tweaked version of this, however, could be used as a good baseline to judge the other models due to its simplicity.

II. **What were obstacles faced if any in working on the project?** This could be technical (like not being able to implement or understand particular techniques) or time issues (midterms for other courses etc).

The main obstacles we faced this reporting period were failures of our group to properly coordinate work with one another, which led to us missing our self-set deadlines for our project. Though we accomplished what we aimed to do, we didn't on the right schedule.

III. **What is the plan for the next reporting period including what each team member is planning to work on. Describe goals and potential timelines (" I plan to finish understanding x to see if it can be implemented for our project by Wednesday etc".)**

Our plans for this week include the following:

- Saturday 9th – Begin to start fine-tuning our models, i.e., tuning parameters for LDA/QDA, choosing the right predictors for KNN, initialize the random forests method, etc. This is all for the purpose of improving test accuracy.
- Monday 11th – Begin drafting the presentation.
- Thursday 14th – Finish all modeling and statistical analysis, then begin drafting the final report/results.
- Sunday 17th – Finish preparation for the presentation/writing the final report.