# COMPSCI 4ML3, Introduction to Machine Learning
# Assignment 2, Fall 2024

Hassan Ashtiani, McMaster University

Due date: Monday, October 21, 11pm

**Notes.** This assignment has a programming component. You should upload only two files: a pdf file (call it MacID.pdf) for all the (typed) solutions including for question 1 and a Jupyter notebook file for your codes (call it MacID.ipynb). Do not compress these files. This assignment has bonus points (i.e., the points add up to more than 100). Bonus points can be used to compensate points lost in other assignments (but does not help with the points lost in the exams).

1. **[80 points]** Programming Component. Follow this link to open the Google colab environment and make a copy of the notebook. Include the answers/graphs/pictures/analyses of the four tasks in your final pdf report. Additionally, upload your modified Jupyter notebook that includes your code (as a separate ipynb file). (for clarifications, post your questions in the Team's Q/A channel).

2. **[30 points]** "Driving high" is prohibited in the city, and the police have started using a tester that shows whether a driver is high on cannabis. The tester is a binary classifier (1 for positive result, and 0 for negative result) which is not accurate all the time:

   - If the driver is truly high, then the test will be positive with probability $1-\beta_1$ and negative with probability $\beta_1$ (so the probability of wrong result is $\beta_1$ in this case)
   - If the driver is not high, then the test will be positive with probability $\beta_2$ and negative with probability $1-\beta_2$ (so the probability of wrong result is $\beta_2$ in this case)

   Assume the probability of (a randomly selected driver from the population) being "truly high" is $\alpha$.

   - **[7 points]** What is the probability that the tester shows a positive result for a (randomly selected) driver? (write your answer in terms of $\alpha, \beta_1, \beta_2$).
   - **[7 points]** The police have collected test results for $n$ randomly selected drivers (i.i.d. samples). What is the likelihood that there are exactly $n_+$ positive samples among the $n$ samples? Write your solution in terms of $\alpha, \beta_1, \beta_2, n_+$ and $n$.
   - **[10 points]** What is the maximum likelihood estimate of $\alpha$ given a set of $n$ random samples from which $n_+$ are positive results? In this part, you can assume that $\beta_1$ and $\beta_2$ are fixed and given. Simplify your final result in terms of $n, n_+, \beta_1, \beta_2$.
   - **[6 points]** What will be the maximum likelihood estimate of $\alpha$ for the special cases of
     - (i) $\beta_1 = \beta_2 = 0$
     - (i) $\beta_1 = \beta_2 = 0.5$
     - (i) $\beta_1 = 0.2$, $\beta_2 = 0.3$