C-M-004: Machine Learning I '21-22

HW 4 (Given November 20, 2021; Due November 24, 2021)

Your answers must be entered in LMS by midnight of the day it is due. If the question requires a textual response, you can create a PDF and upload that. The PDF might be generated from MS-WORD, LaTeX, the image of a handwritten response, or using any other mechanism. Code must be uploaded and may require demonstration to the TA. Numbers in the bold indicate points allocated to the question and *make sure that you explain your choice in each question below*

---

1. *Consider the following dataset (x_1 is a categorical input, x_2 is a numerical input and y is a categorical output),*

x_1	x_2	y
F	12	F
F	14	Y
T	13	Y
T	16	F

   Now answer the following questions (indicate your answer by unambiguously filling the bubble next to your choice),

   (a) *The best discretization for x_2 from an information gain perpective is based on which threshold (if a threshold is θ, then $x_2 \leq \theta$ is 0, else 1): (i)* 12, *(ii)* 13, *(iii)* 14, *(iv)* 16 **(10 points)**

   (b) *Based on entropy, the first split will be based on: (i) x_1, (ii) x_2, (iii) Doesn't matter* **(10 points)**

2. *Is a random forest of random forests a good idea? (i) No, Absolutely not, (ii) Yes, of course, (iii) Maybe, varies from case to case, (iv) None of the above* **(20 points)**