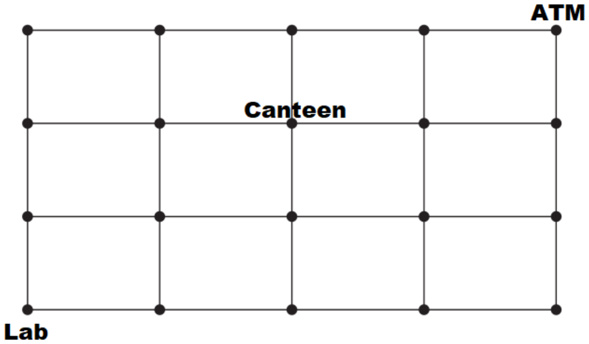
MANIPAL SCHOOL OF INFORMATION SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)

AML5103 | Applied Probability and Statistics | Sessional-1

1. [10 points] [CO 1, BT 3] Consider the grid of locations shown here. Starting from the lab, you want to go to the ATM. Suppose that you can go one step up (↑) or one step to the right (→) at each move.



   (a) How many different paths from the lab to the ATM are possible?

   (b) How many different paths from the lab to the ATM are possible if you want to avoid the crowded canteen?

2. [10 points] [CO 1, BT 5] Mr. Brown needs to take 1 tablet of type A and 1 tablet of type B together on a regular basis. One tablet of type A corresponds to a 1 mg dosage, and so does 1 tablet of type B. He keeps these two types of tablets in two separately labeled bottles as they cannot be differentiated easily. One day, on a business trip, Mr. Brown brought 10 tablets of type A and 10 tablets of type B.. Unfortunately, he drops the bottles and breaks them resulting in all 20 tablets getting mixed up. He does not have the time to go to a pharmacy to buy a new set of tablets but he needs to take his required dosage of both tablets A and B. The safe dosage that he needs for both tablets A and B is given by

$$0.9\,\text{mg} \leq \text{safe dosage} \leq 1.1\,\text{mg}.$$

Taking either an excess or a shortage of the required intake will result in serious health issues. Suppose Mr. Brown decides to break each tablet in the pile into 10 smaller pieces having exactly the same size (0.1 mg each), and then randomly pick 20 of those smaller pieces. Justify whether he has a more than 50% chance of developing a serious health issue following this strategy.

Hint: Let the 10 type-A tablets $A_1, \ldots, A_{10}$ and the 10 type-B tablets $B_1, \ldots, B_{10}$ after breaking into smaller pieces be represented as:

$$\underbrace{A_{1,1}, A_{1,2}, \ldots, A_{1,10}}_{\text{pieces from } A_1}, \underbrace{A_{2,1}, A_{2,2}, \ldots, A_{2,10}}_{\text{pieces from } A_2}, \ldots, \underbrace{A_{10,1}, A_{10,2}, \ldots, A_{10,10}}_{\text{pieces from } A_{10}},$$
$$\underbrace{B_{1,1}, B_{1,2}, \ldots, B_{1,10}}_{\text{pieces from } B_1}, \underbrace{B_{2,1}, B_{2,2}, \ldots, B_{2,10}}_{\text{pieces from } B_2}, \ldots, \underbrace{B_{10,1}, B_{10,2}, \ldots, B_{10,10}}_{\text{pieces from } B_{10}}$$

3. [10 points] [CO 2, BT 5] Suppose that an insurance company classifies people who buy two-wheeler

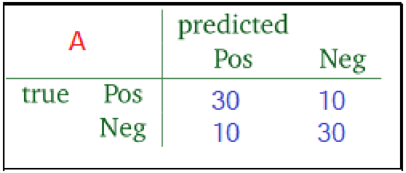
insurance from them into one of four classes: good, fair, average, and bad risks. As a data scientist for the company, you have access to the following customer data for the calendar year 2023-24:

Class	**Total no. of riders**	**No. involved in accident**
Bad risk	5000	2000
Average risk	4000	1200
Fair risk	3000	600
Good risk	3000	300

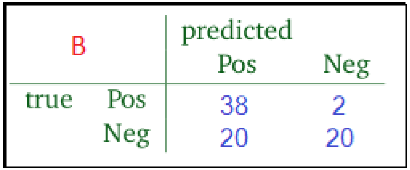
(a) What is the probability that a new customer will meet with an accident during 2024-25?

(b) If a customer had not met with an accident during 2023-24, what is the probability that they are a good rider? Based on the answer, justify whether not having met with an accident is a reasonable indication of good driving behavior.

4. [10 points] [CO 2, BT 5] Consider the performance of two algorithms A and B shown below for a binary classification task:



A		predicted	
		Pos	Neg
true	Pos	30	10
	Neg	10	30



B		predicted	
		Pos	Neg
true	Pos	38	2
	Neg	20	20

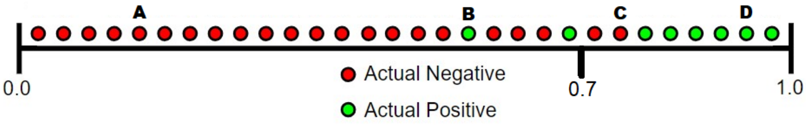
(a) For both algorithms, fill the entries of the table below:

	Accuracy	Recall	Precision	TNR
A	?	?	?	?
B	?	?	?	?

(b) In each one of the following scenarios, justify which algorithm you would use:

- spam email detection for busy bank staff;
- airport security screening for explosives.

5. [10 points] [CO 2, BT 5] Suppose we train a probabilistic classification algorithm for predicting the presence of a life-altering disease in 30 patients about whom we know their actual status (presence/absence of the disease). The result is presented below where four patients **A**, **B**, **C**, and **D** are specifically highlighted. Suppose we use a threshold of 0.7 to determine the predicted classes:



(a) Match **A**, **B**, **C**, and **D** with one each of TP, TN, FP, and FN along with an English explanation of what it means for the four patients.

(b) Calculate the area under the RoC curve

(c) What is the FPR? Justify if the resulting value is a good indication of the model performance.

(d) Decide on which one among the *area under the RoC curve* and the *area under the precision-recall curve* will be a good measure of the algorithm's performance with a brief justification.