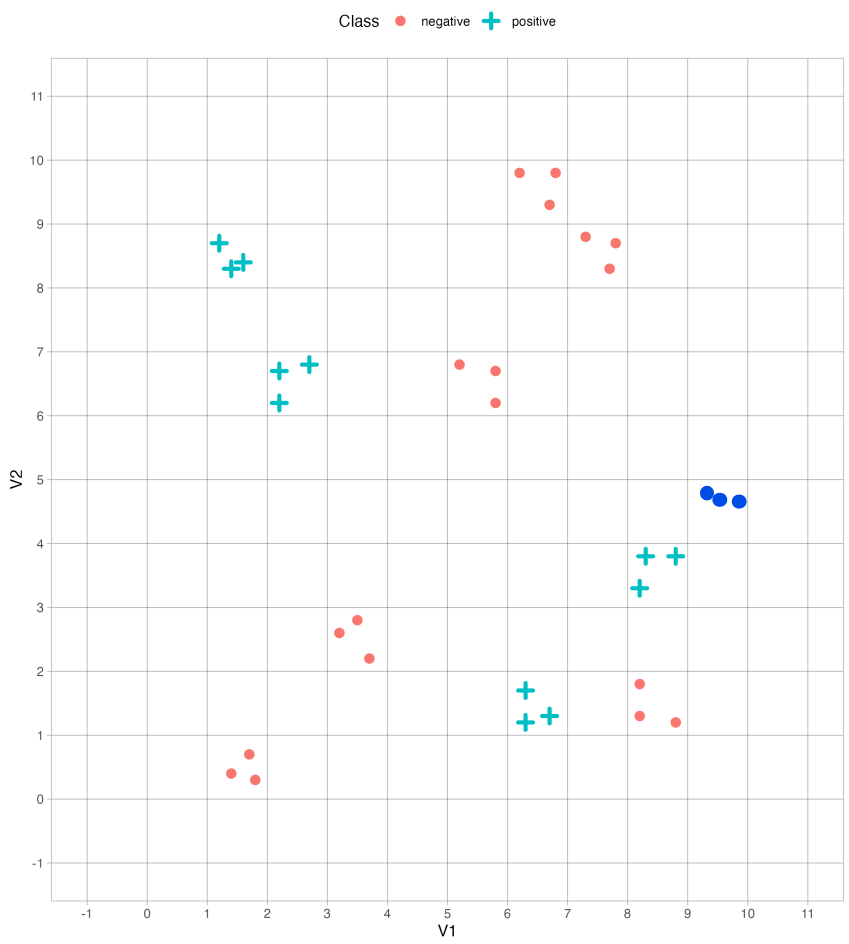
Show your working in all calculations.

Important: The following questions will require you to use parts of your **six-digit candidate ID** (the same one you are using to identify yourself in this assessment). For instance, if your six-digit candidate ID was 987654, then the digits would be referred to as follows:

9	8	7	6	5	4
↑	↑	↑	↑	↑	↑
A	B	C	D	E	F

Note: your values will be different, it's the **position** that's important!

1. (This question is about **Classification**) Assume that the data set illustrated below (Fig 1) relates to two measurements (columns **V1**, **V2**) coming from different volunteers (column **ID**), each of which was measured three times, related to a given characteristic of interest (column **Class**).



ID	V1	V2	Class	Test?
A	1.7	0.7	negative	
A	1.4	0.4	negative	
A	1.8	0.3	negative	
B	1.2	8.7	positive	
B	1.4	8.3	positive	
B	1.6	8.4	positive	
C	3.7	2.2	negative	
C	3.2	2.6	negative	
C	3.5	2.8	negative	
D	6.3	1.2	positive	
D	6.3	1.7	positive	
D	6.7	1.3	positive	
E	5.8	6.7	negative	
E	5.8	6.2	negative	
E	5.2	6.8	negative	
F	8.2	3.3	positive	
F	8.3	3.8	positive	
F	8.8	3.8	positive	
G	2.7	6.8	positive	
G	2.2	6.7	positive	
G	2.2	6.2	positive	
H	7.3	8.8	negative	
H	7.7	8.3	negative	
H	7.8	8.7	negative	
I	6.2	9.8	negative	
I	6.7	9.3	negative	
I	6.8	9.8	negative	
J	8.2	1.8	negative	
J	8.2	1.3	negative	
J	8.8	1.2	negative	
K			positive	
K			positive	
K			positive	

Figure 1: Data set and figure for Question 1

: 9 8 7 6 5 4
: A B C D E F
1 V1=9.2; V2=4.73
2 V1=9.5; V2=4.6
3 V1=9.8; V2=4.60

a) (2 marks) Add the following three points to the data set, related to *volunteer K*. **Include them in both the table and the plot in Figure 1**, using only 2 decimal places (replace A, B, C and D by the corresponding digits from your candidate ID). *Note: This figure and table are also available in an editable, separate document in the Special Resources folder*.

- Point 1: $V1 = A + 0.2; \quad V2 = 4.2 + B/15$
- Point 2: $V1 = A + 0.5; \quad V2 = 4.2 + C/15$
- Point 3: $V1 = A + 0.8; \quad V2 = 4.2 + D/15$

b) (8 marks) Divide this data into a *Training Set* containing *approximately* 2/3 of the observations, and a *Test Set* with *approximately* 1/3 of the observations. **Explain your data splitting rationale**. Mark the observations of the *Test Set* with an "X" in the corresponding column of the table in Figure 1 (note: you may also want to indicate them in the figure to prevent confusion in item 1c).

c) (15 marks) Build a Decision Tree classifier based on your *Training Set* to predict *Class*, using the following rules:

- Only integer values can be used for splitting (i.e., no decimal places in the thresholds - use integer values such as 1, 2, etc.).
- Use the plot in Figure 1 to decide the attribute to be used for splitting at each node and the threshold to use. Your splits should be reasonable in terms of decreasing node impurity, but they don't need to be optimal - meaning that you don't need to calculate the information gain of all possible splits. At each node, pick an attribute and a threshold that look reasonable in terms of decreasing impurity and use it for that node. **Explain your rationale** for choosing the attribute/thresholds at each node.
- Maximum tree depth of 3, i.e., at most 3 *splits* from root to leaf.

Draw your decision tree highlighting (i) the **split criteria** used at each non-leaf node, (ii) the **Information Gain** of each split, and (iii) the **proportion of training examples** of each class in the leaf nodes.

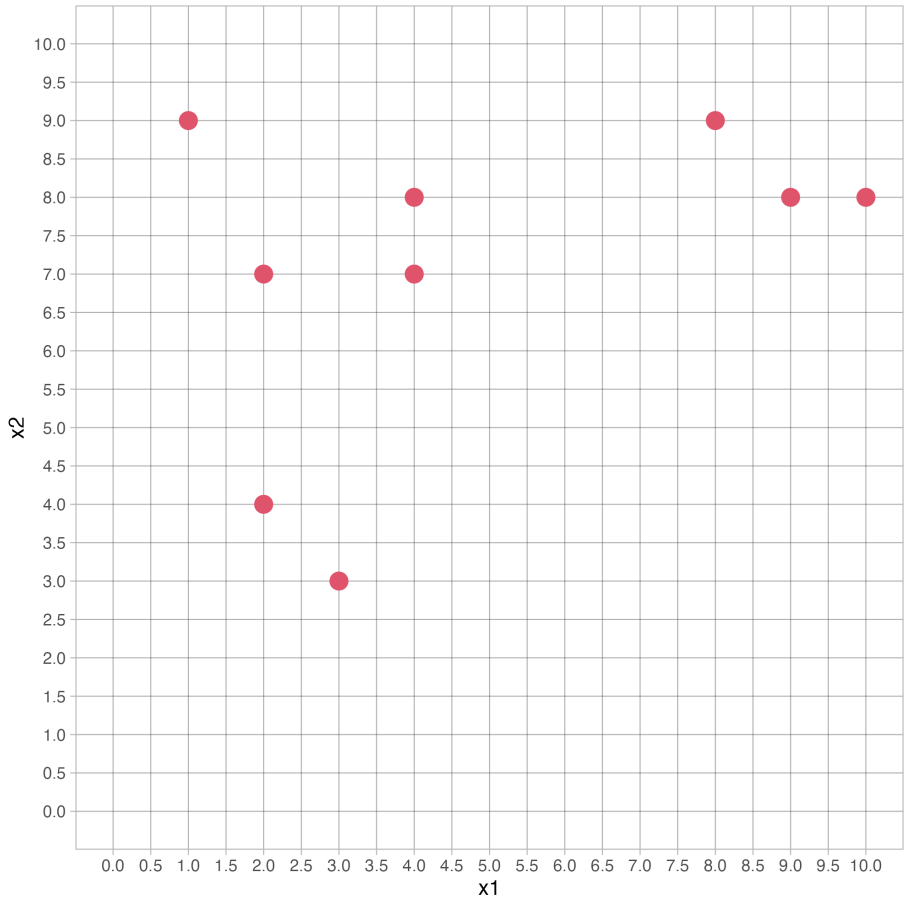
d) (5 marks) Use your *Test Set* to estimate the performance of your model. **Draw the *confusion matrix*** for the test set and **calculate the *F1 score*** obtained by your model.

e) (5 marks) Discuss the **definition** and **interpretation** of two other performance indicators that could be used to assess the quality of a classification model (not including the F1 score).

(Total: 35 Marks)

Show your working in all calculations.

2. (This question is about **Clustering**) Consider the data shown in Figure 2.



ID	x1	x2
A	2	4
B	3	3
C	1	9
D	2	7
E	4	8
F	4	7
G	8	9
H	10	8
I	9	8
J		
K		
L		

Figure 2: Data set and figure for Questions 2a, 2b and 2c.

	A	B	C	D	E	F	G	H	I	J	K	L
A	0	X	X	X	X	X	X	X	X	X	X	X
B	2	0	X	X	X	X	X	X	X	X	X	X
C	6	8	0	X	X	X	X	X	X	X	X	X
D	3	5	3	0	X	X	X	X	X	X	X	X
E	6	6	4	3	0	X	X	X	X	X	X	X
F	5	5	5	2	1	0	X	X	X	X	X	X
G	11	11	7	8	5	6	0	X	X	X	X	X
H	12	12	10	9	6	7	3	0	X	X	X	X
I	11	11	9	8	5	6	2	1	0	X	X	X
J										0	X	X
K											0	X
L												0

Table 1: Dissimilarity matrix for question 2a.

a) (2 marks) Add the following three points to the data set. **Include them in both the table and the plot in Figure 2** (use the corresponding digits from your candidate ID). *Note: This figure and table are also available in an editable, separate document in the Special Resources folder*.

- Point J: $x_1 = A + 0.5; \quad x_2 = B + 0.5$
- Point K: $x_1 = C + 0.5; \quad x_2 = D + 0.5$
- Point L: $x_1 = E + 0.5; \quad x_2 = F + 0.5$

b) (15 marks) **Build an agglomerative hierarchical clustering of this data**, following the specifications below:

- Use the *L1 distance* to measure the dissimilarities between points.
- Use *complete linkage* for merging your clusters.
- In case of ties when deciding merges, simply choose one option arbitrarily.

A pre-populated dissimilarity table with the Manhattan distances between the original points in the dataset has been pre-populated for you (Table 1). **Complete this table with the dissimilarities calculated for your new points**, and use the table to support your clustering. You can also use the plot in Figure 2 to support your work. **Detail your merging criteria/rationale at each step**.

Note: as an example, the explanation of merging steps could look like this:

- *Merge 1: smallest distance (dist = 3) is between points F, K. Merged into cluster F,K.*
- *Merge 2: smallest distances (dist = 4) are between point G and cluster F,K and between points B and C. Selected the first to merge into cluster F,K,G.*
- *Merge 3: ...*

... until all points are joined into a single cluster.

c) (5 marks) **Draw the resulting hierarchical clustering diagram detailing the points under each cluster and the dissimilarity levels of the merges**. Figure 3 provides a simplified example of how this should look.

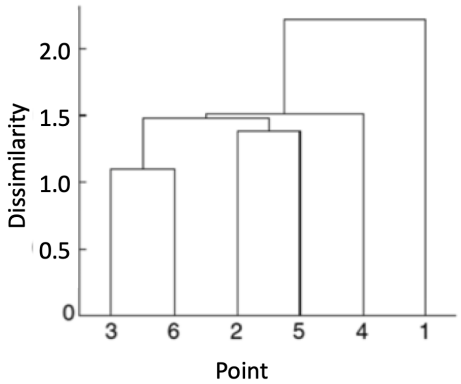


Figure 3: Example of hierarchical clustering diagram (on a completely different data set). Adapted from Tan at al. (2019).

d) (3 marks) Assume a dissimilarity threshold of 3.5. **Indicate this threshold on the clustering diagram** you produced for item (*2c*), and **circle the resulting clusters in the plot of Figure 2**.

e) (6 marks) Assume a data set of one-dimensional points defined as $\{A, B, C, D, 12, 23, 29, 37\}$ (replace A, B, C, D by the corresponding digits of your candidate ID). Perform **two iterations of the basic k-means** for these points ,using initial centroids given by $(E + 4)$ and $(F + 20)$. **Detail your iterations and clearly indicate the final clusters obtained**. Perform all calculations rounding to 1 decimal place.

f) (4 marks) **Discuss** two advantages and two disadvantages/limitations of k-means for clustering.

END OF ASSESSMENT