Instructions: This assignment covers topics of classification.

Submission Requirements:

In *w2*, this part, you will be asked to walk through methods and calculations **manually**. In *a2-R* or *a2-python*, you will use packages and libraries to implement classification models.

You must prepare your solution to this part as a document using LaTeX. I have provided a template for this assignment, where you can create your answers.

For this assignment, you are to work in your **groups**. I highly suggest using Overleaf to work on your submission together in your group.

Follow the submission template where work for each question must **start on a new page**. Do not put work for multiple problems on the same page.

Questions:

1. Naïve Bayes Classification
   Consider the fruit data set below. You will construct and evaluate a Naïve Bayes classifier to predict whether a fruit is an orange or apple using the remaining features (weight, height, and width).

Weight	Height	Width	Type
0	1	2	1
0	1	2	1
0	2	1	1
0	2	0	0
0	2	0	1
0	0	0	0
0	2	1	0
0	0	1	1
0	0	0	1
0	0	0	1
0	0	0	1
0	0	0	1
0	1	1	0
0	0	0	1
1	1	0	0
1	1	2	0
1	1	1	0
1	1	0	0
1	2	0	0
1	2	0	0
1	2	0	1
1	0	0	1
1	0	0	1
1	0	0	0
1	0	0	1

Columns: Type, Weight, Height, Width

Weight (g)
- 0 - if wt ≤ 179.42
- 1 - otherwise

Height (cm)
- 2 - if ht > 8
- 1 - if $6.8 \leq$ ht ≤ 8
- 0 - otherwise

Width (cm)
- 2 - if width > 7.8
- 1 - if $6.5 \leq$ width ≤ 7.8
- 0 - otherwise

Type
- 1 - apple
- 0 - orange

   (a) (18 points) Estimate the conditional probabilities needed for Naïve Bayes classification using Laplace smoothing, where $\alpha = 1$ and β is size of variable's domain. Estimate an unsmoothed prior probability.

$$P(Y = y_k) = \frac{\#(Y = y_k)}{n}$$

$$P(X_i = j \,|\, Y = y_k) = \frac{\#(X_i = j, Y = y_k) + 1}{\#(Y = y_k) + |domain(X_i)|}$$

Fill in the following tables (**report as fractions**):

Prior	Prob.
$P(apple)$	
$P(orange)$	

Cond.	Prob.
$P(Wt = 0 \mid apple)$	
$P(Wt = 1 \mid apple)$	
$P(Wt = 0 \mid orange)$	
$P(Wt = 1 \mid orange)$	

Cond.	Prob.
$P(Ht = 0 \mid apple)$	
$P(Ht = 1 \mid apple)$	
$P(Ht = 2 \mid apple)$	
$P(Ht = 0 \mid orange)$	
$P(Ht = 1 \mid orange)$	
$P(Ht = 2 \mid orange)$	

Cond.	Prob.
$P(Wid = 0 \mid apple)$	
$P(Wid = 1 \mid apple)$	
$P(Wid = 2 \mid apple)$	
$P(Wid = 0 \mid orange)$	
$P(Wid = 1 \mid orange)$	
$P(Wid = 2 \mid orange)$	

(b) (24 points) Report the predicted class on the following test samples using the estimated parameters from the tables above.

Sample Num.	Test Data [Weight, Height, Width]	Prediction
1	[1, 0, 0]	
2	[0, 0, 1]	
3	[1, 2, 0]	
4	[0, 1, 1]	

Show the calculations, writing out the probabilities that go into making the prediction.

2. Decision Trees

Suppose you are building a decision tree on a data set with three classes A, B, C. At the current position in the tree you have the following samples available:

$$N = \begin{pmatrix} A & 80 \\ B & 60 \\ C & 60 \end{pmatrix}.$$

You are examining two ways to split the data. The variable X_1 splits the data as,

$$N_{1,1} = \begin{pmatrix} A & 43 \\ B & 18 \\ C & 7 \end{pmatrix}, \qquad N_{1,2} = \begin{pmatrix} A & 37 \\ B & 42 \\ C & 53 \end{pmatrix}.$$

The variable X_2 splits the data as,

$$N_{2,1} = \begin{pmatrix} A & 46 \\ B & 26 \\ C & 13 \end{pmatrix}, \qquad N_{2,2} = \begin{pmatrix} A & 14 \\ B & 12 \\ C & 35 \end{pmatrix}, \qquad N_{2,3} = \begin{pmatrix} A & 20 \\ B & 22 \\ C & 12 \end{pmatrix}.$$

(a) (14 points) Compute the information gain (based on entropy) for the two possible attributes. Show the form of the calculations, not just the final numbers.

i. Entropy before the split

ii. Information gain for variable, X_1

iii. Information gain for variable, X_2

Note, when calculating Entropy use $\log_2$

(b) (2 points) Which variable would be preferred to be included next in the decision tree?

(c) (14 points) Compute the gain in GINI index for the two possible attributes. Show the form of the calculations, not just the final numbers.

i. GINI before the split

ii. gain in GINI for variable, X_1

iii. gain in GINI for variable, X_2

(d) (2 points) Which variable would be preferred to include next in the decision tree?