23/4/18

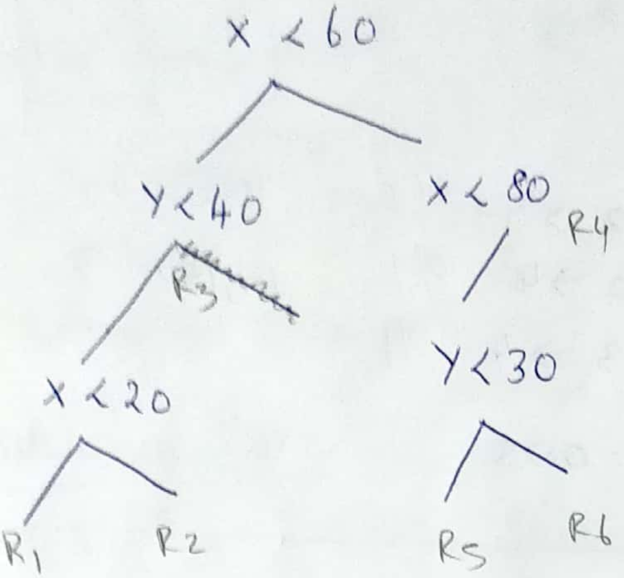
Data preparation and Analysis

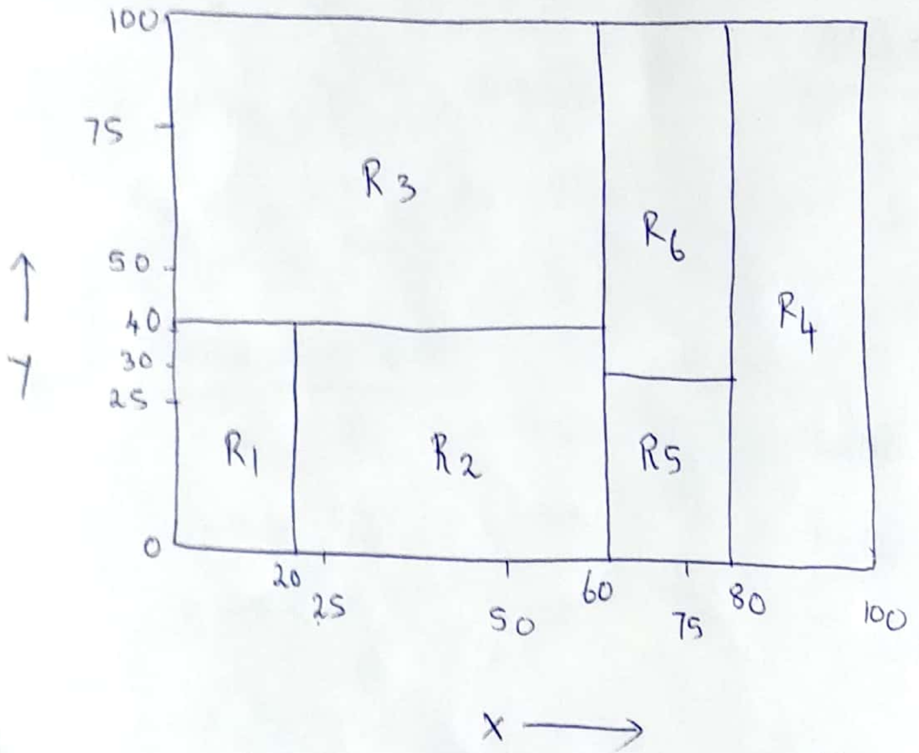
Assignment 4

A20406657
Avinash Villuri

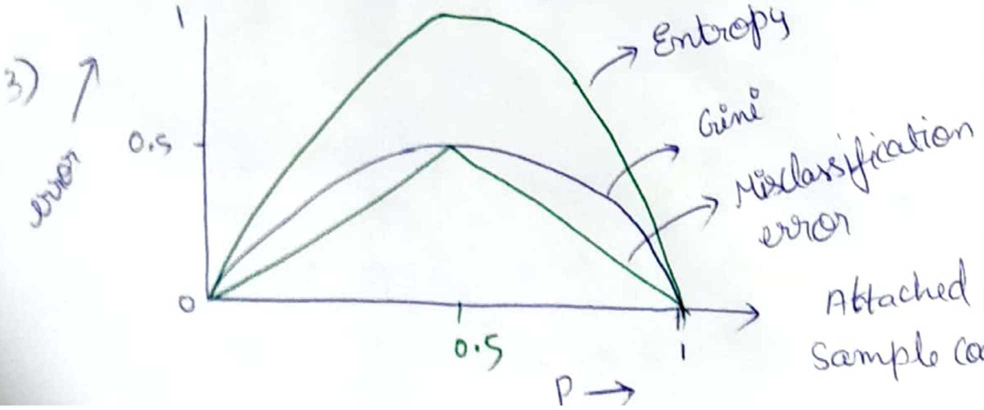
Problem 1.1:-

1)



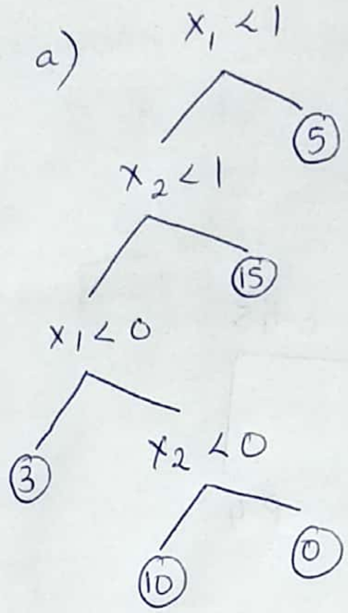


3)

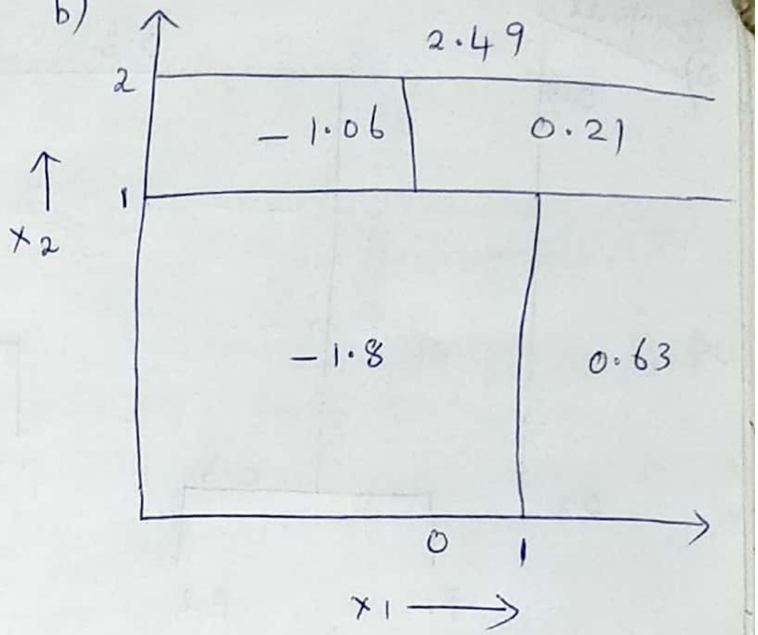


Attached pdf with sample code

4) a)



b)



5)

Majority approach:-

we have 6 out of 10 samples which has probability greater than 0.5

So predicted class = Red

Average approach:-

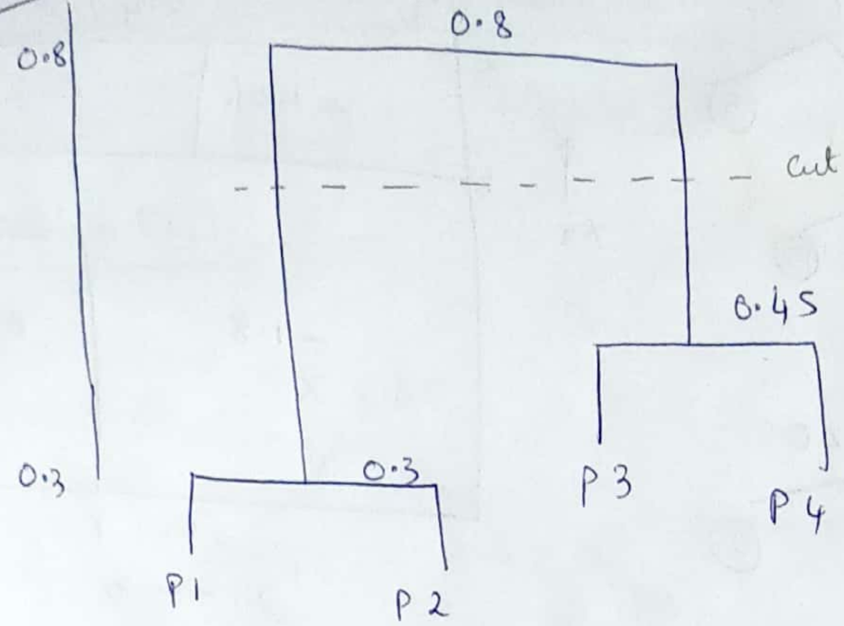
mean (observed probabilities) = 0.45

So the predicted Class = Green since the average probability is less than 0.5

Problem 1.2

2)
$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

a) Complete



	P_1	P_2	P_3	P_4
P_1	0			
P_2	$\boxed{0.3}$ Min	0		
P_3	0.4	0.5	0	
P_4	0.7	0.8	0.45	0

Max operation

	P_1-P_2	P_3	P_4
P_1-P_2	0		
P_3	0.5	0	
P_4	0.8	$\boxed{0.45}$ Min	0

⇒

Max operation

	P_1-P_2	P_3-P_4
P_1-P_2	0	
P_3-P_4	$\boxed{0.8}$ Min	0

b) Single

	P_1	P_2	P_3	P_4
P_1	0			
P_2	$\boxed{0.3}$ Min	0		
P_3	0.4	0.5	0	
P_4	0.7	0.8	0.45	0

Minoperation

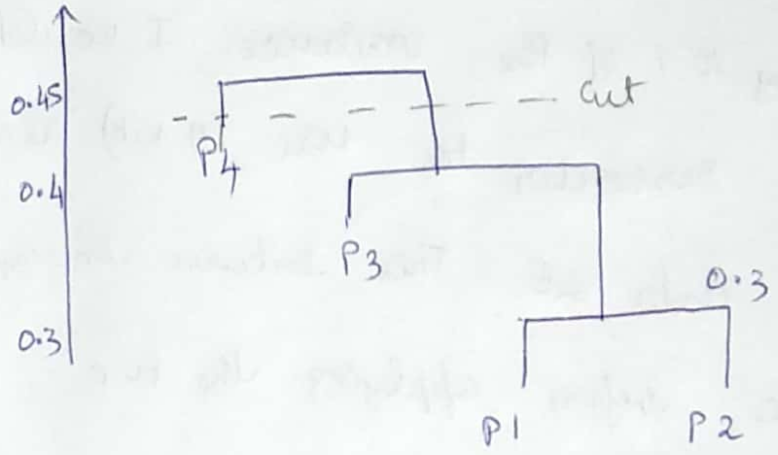
	P1-P2	P3	P4
P1-P2	0	min	
P3	$\boxed{0.4}$	0	
P4	0.7	0.45	0

⇒

Min operation

	P1-P2-P3	P4
P1-P2-P3	0	
P4	min $\boxed{0.45}$	0

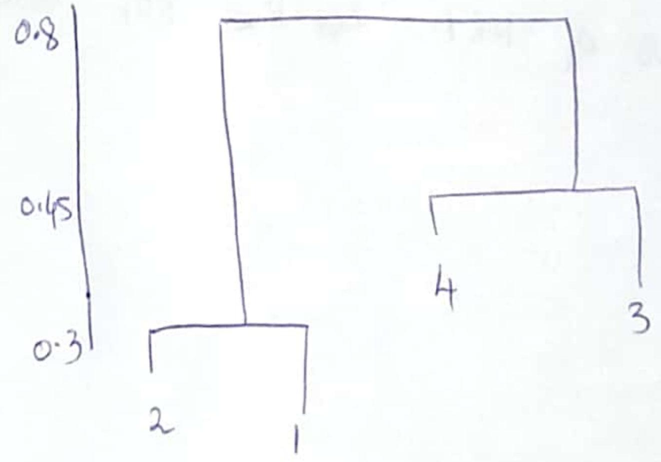
⇒ P1-P2-P3-P4



c) (1,2) (3,4)

d) (4) (1,2,3)

e)



3) Attached Assignment4_Theory.Rmd file.

6) a) "explains 10% of the Variation" means the first principle component

has captured only 10% of the total variance. and 90% of information is lost by Projecting tissue sample and gene data set.

b) Since the first principle component explained only 10% of the variance. I would suggest the researcher to use (A VsB) as a feature of the data set. This inturn increases the variance before applying the two-Sample t-test.

c) Attached R-code in Assigment4_Theory.Rmd

Analysis:- After performing A VsB there is an Improvement of 1.6% in the PC$_1$ Variance.