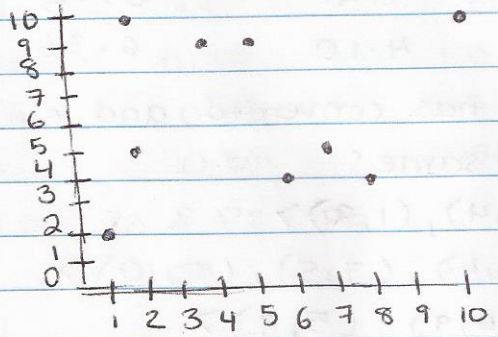
Big Data Analytics & Management Assignment 4. → TTS150030

Part 1:- Clustering

A) Points: (2,10), (2,5), (8,4), (5,9), (7,5), (6,4), (1,2), (4,9), (10,10)

i) plotting all the points :-



ii) points to begin with : (2,5), (5,8) & (4,9)

points	$C_1(2,5)$	$C_2(5,8)$	$C_3(4,9)$	cluster
(2,10)	5	3.6	2.23	C3
(8,4)	6.08	5	6.40	C2
(5,9)	3	1	1	C3
(7,5)	5	3.6	5	C2
(6,4)	4.12	4.12	5.36	C1
(1,2)	3.16	7.21	7.61	C1
(10,10)	9.43	5.38	6.08	C2

In order to get final clusters :

$$C_1 = \frac{2+6+1}{3}, \frac{5+4+2}{3} = (3, 3.67)$$

$$C_2 = \frac{5+8+7+10}{4}, \frac{8+4+5+10}{4} = (7.5, 6.75)$$

$$C_3 = \frac{4+2+5}{3}, \frac{9+10+9}{3} = (3.67, 9.33)$$

iii)

points	$C_1(3,3.67)$	$C_2(7.5,6.75)$	$C_3(3.67,9.33)$	clust
(2,10)	6.40	6.39	1.79	C3
(2,5)	1.66	5.77	4.67	C1
(8,4)	5.01	2.795	6.86	C2
(5,9)	5.69	3.36	1.37	C3
(7,5)	4.21	1.82	5.46	C2

Points	$C_1(3, 3.67)$	$C_2(7.5, 6.75)$	$C_3(3.67, 9.3)$	Clust
(6,4)	3.01	3.13	5.81	C_1
(1,2)	2.6	8.05	7.8	C_1
(4,9)	5.42	4.16	0.45	C_3
(10,10)	9.42	4.10	6.36	C_2

It is clear that k-means has converged and cluster assignments remain the same:

C_1 holds $<(2,5), (6,4), (1,2)>$

C_2 holds $<(5,8), (8,4), (7,5), (10,10)>$

C_3 holds $<(2,10), (4,9), (5,9)>$

B) Single Linkage for similarity Matrix / Min Criterion :-

- Link P2 & P5 :-

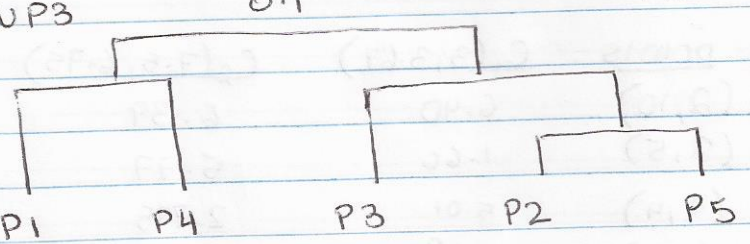
	P1	P2 ∪ P5	P3	P4
P1	1.0	0.1	0.41	0.55
P2 ∪ P5	0.1	1.0	0.64	0.47
P3	0.41	0.64	1.0	0.44
P4	0.55	0.47	0.44	1.0

- Link P2 ∪ P5 & P3

	P1	P2 ∪ P5 ∪ P3	P4
P1	1.0	0.1	0.55
P2 ∪ P5 ∪ P3	0.1	1.0	0.44
P4	0.55	0.44	1.0

- Link P1 ∪ P4

	P1 ∪ P4	P2 ∪ P5 ∪ P3
P1 ∪ P4	1.0	0.1
P2 ∪ P5 ∪ P3	0.1	1.0



Complete Linkage for similarity Matrix / Max Criterion :-

- Link P2 & P5

	P1	P2 ∪ P5	P3	P4
P1	1.0	0.35	0.41	0.55
P2 ∪ P5	0.35	1.0	0.85	0.76
P3	0.41	0.85	1.0	0.44
P4	0.55	0.76	0.44	1.0

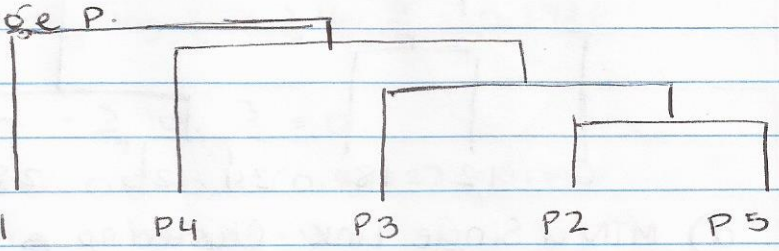
- Link P3 & P2 ∪ P5

	P1	P2 ∪ P5 ∪ P3	P4
P1	1.0	0.41	0.55
P2 ∪ P5 ∪ P3	0.44	1.0	0.76
P4	0.55	0.76	1.0

- Link P4 & P2 ∪ P5 ∪ P3

	P1	P2 ∪ P5 ∪ P3 ∪ P4
P1	1.0	0.55
P2 ∪ P5 ∪ P3 ∪ P4	0.55	1.0

- Now merge P.



c) Points :- {6, 12, 18, 24, 25, 28, 30, 42, 48}

a) 1) {5, 7.5}

clusters would be as follows :-

C_1 = {6}

C_2 = {12, 18, 24, 25, 28, 30, 42, 48}.

2) {15, 25}

Clusters would be as follows :-

C_1 = {6, 12, 18}

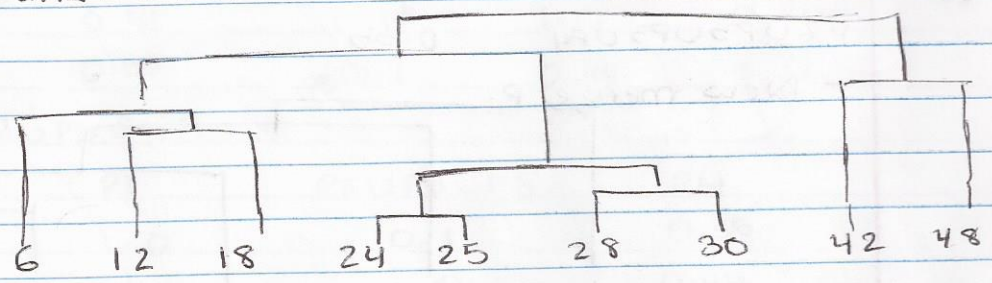
C_2 = {24, 25, 28, 30, 42, 48}

b) 1) The first centroid does not represent a stable state since there would be a change when the centroid gets run again.

2) The second centroid does represent a stable state since there would not be a change when the centroid gets run again.

c) Single link / Min criterio run on two clusters :-

- Link 24 & 25
- Link 28 & 30
- Link 24 U 25 & 28 U 30
- Link 12 & 18
- Link 6 & 12 U 18
- Link 42 & 48
- Link 6 U 12 U 18 & 24 U 25 U 28 U 30
- Link 6 U 12 U 18 U 24 U 25 U 28 U 30 & 42 U 48



d) MIN or Single Link Clustering is the most natural clustering this is because of the uniform Density.

e) What explains the previous behavior is that the distance between 30 & 42 is much more than any other existing point, hence these would be the most natural points. K-Means would not be able to capture this, however MIN can.

Part 2 :- Classification

Entropy (y) $= \frac{-6}{11}\log_2\frac{6}{11} - \frac{5}{11}\log_2\frac{5}{11} = 0.2992$

Splitting criterion

$x_1 = a = \frac{-3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.2922$

$x_1 = b = \frac{-3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 0.301$

Gain $= 0.2992 - (0.2922 + 0.301) = -0.294$

$x_2 = c = \frac{-2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.2922$

$x_2 = b = \frac{-2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 0.30$

Gain $= 0.2992 - (0.2922 + 0.301) = 0.293$

$x_3 = k = \frac{-3}{3}\log_2\frac{3}{3} = 0$

$x_3 = v = \frac{-2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.2922$

$x_3 = r = \frac{-3}{3}\log_2\frac{3}{3} = 0.$

Gain $= 0.2992 - (0.2922) = 0.007$

- Split on x_3 when $x_3 = v$

x_1	x_2	y	
b	w	+1	
a	c	+1	Entropy = 0.2922.
a	u	-1	
b	c	-1	
b	d	-1	

$x_1 = a = \frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 0.301$

$x_1 = b = \frac{-1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.2764$

Gain $= 0.2922 - (0.30 + 0.2764) = 0.2852$

$x_2 = c = \frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 0.301$

$x_2 = m$, $x_2 = g$, $x_2 = w$

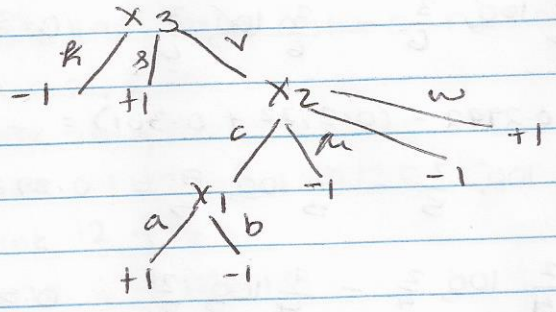
Gain = 0.2922 − 0.301 = −0.088

- Now split on x_2

When $x_2 = c$ choose $x_1 = a = +1$

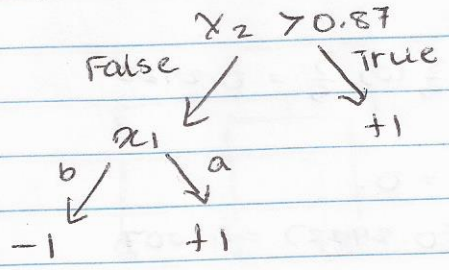
& $x_1 = b = -1$.

So :-



$(x_2)_{val} = \frac{val - min}{max - min} = \frac{val - 140}{45}$

Tree reduces to become :-



Part 3 :- Programming

E) source code included with submission.

MNIST_PYTORCH.ipynb

MNIST_TENSORFLOW.ipynb.

F) source code included with submission.

MSE of model :- 86.8%