Data Mining 2020 final

Q1) i) classification => is a process of finding a Model That describes and distinguishes Data classes

ii) clustering => analyze Data objects without class labels. The objects are grouped based on Similarity

iii) Dendrogram => Graph That show us How we Can Merge Two object Together and we Can Construct The resulted clusters after doing Agglomerative Algo

iv) Classification Vs cluster

• Dividing Data Set based on Values of Class labels	• Dividing Data set based on Similarity
• Predictive	• Descritive
• Supervised learning	• unSupervised Learning

v) Model Construction => is first Process of classification That use Training Data Set To Construct Model. This Model help us To Predict The Value of class Label of New object

vi) Model evaluation => is Second Process of classif That DeTeminer How The Model is good for Predict The cL Value ~~of Data on The Data set~~ Set of records of Testing Data set

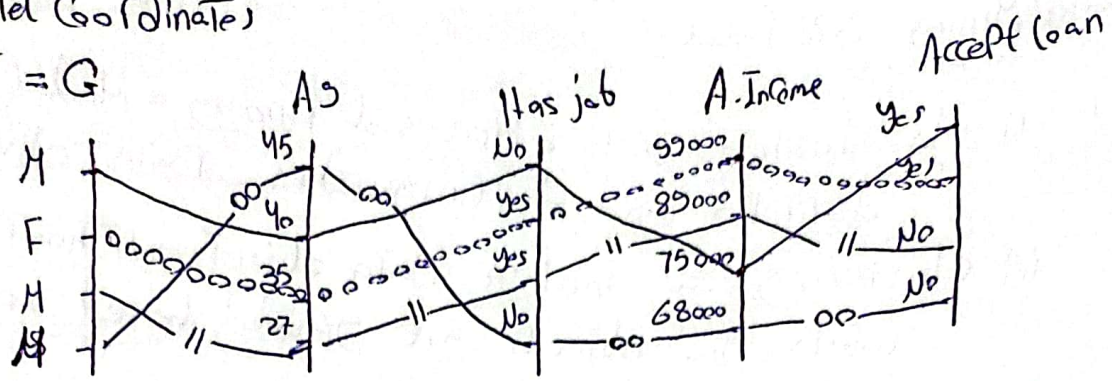
vii) Model usage => use Model To Predict a ~~New~~ cL Value of New object

viii) Evaluation Matrix => This Matrix help us To Compute accuracy & Precision & Recall of classifier To Determine How This classifier is Good

TP	FN
FP	TN

[1]

b) i) Parallel Coordinates



ii) obj1 & obj4

$obj_1 = (M, 40, No, 75000, yes)$

$obj_4 = (M, 45, No, 68000, No)$

$\delta_G = 1, \delta_{Ag} = 1, \delta_{Has\ job} = 1, \delta_{Ainc} = 1, \delta_{Accept\ Loan} = 1$

$Dis_G = 0 \quad (M = M)$

$Dis_{Ag} = \frac{|40-45|}{45-27} = 0.27$

$Dis_{Has\ job} = 0 \quad (No = No)$

$Dis_{Ainc} = \frac{|75000-68000|}{99000-68000} = 0.22$

$Dis_{Accept\ Loan} = 1 \quad (yes \neq No)$

$Dis(obj1, obj2) = \frac{1*0 + 1*0.27 + 1*0 + 1*0.22 + 1*1}{1+1+1+1+1} = 0.298$

$Sim(obj1, obj2) = 1 - Dis = 1 - 0.298 = 0.702$

obj2 & obj3

$obj_2 = (F, 35, yes, 99000, yes)$

$obj_3 = (M, 27, yes, 89000, No)$

$Dis_G = 1 \quad (F \neq M)$

$Dis_{Ag} = \frac{|35-27|}{45-27} = 4/9$

$Dis_{Has\ job} = 0$

$Dis_{Ainc} = \frac{|99000-89000|}{99000-68000} = 10/31$

$Dis_{Accept\ Loan} = 1$

$Dis(obj2, obj3) = \frac{1*1 + 1*4/9 + 1*0 + 1*10/31 + 1*1}{1+1+1+1+1}$

$= 0.553$

$Sim(obj2, obj3) = 1 - Dis = 1 - 0.553 = 0.447$

(2)

Q2

a) Mean لـ Age في الـ Missing
= (25+27+32+45+43+28+29+17+21+31+36+42) / 12 = 31.3 ≈ 31

Credit Rating في الـ Missing

Missing ⇒ fair

Student في الـ Missing

Missing ⇒ yes

b) New object (20, low, yes, fair, ?)

$P(yes) = 9/14$

$P(No) = 5/14$

$P(20|yes) = \frac{1}{7.58 * \sqrt{2\pi}} * e^{-\frac{(20-30.3)^2}{2*(7.58)^2}} = 0.020$

$P(low|yes) = \frac{3}{9} = 1/3$ $P(fair|yes) = \frac{6}{9} = 2/3$

$P(yes|yes) = \frac{7}{9}$

$P(yes|New) = P(20|yes) * P(low|yes) * P(yes|yes) * P(fair|yes) * P(yes)$

$= 0.020 * 1/3 * 7/9 * 2/3 * 9/14 = \boxed{2.32 * 10^{-3}}$

هكذا الـ Probab تاعت الـ New ع yes

نشوف تاعت الـ New ع No

<u>$P(20|No)$</u>

$\mu_{(age)} = \frac{25+27+43+28+42}{5} = 33$

$\sigma_{(age)} = \sqrt{\frac{(25-33)^2 + (27-33)^2 + \cdots + (42-33)^2}{5}} = 7.823$

$P(20|No) = \frac{1}{7.82 * \sqrt{2\pi}} e^{-\frac{(20-33)^2}{2 * (7.82)^2}} = 0.012$

$P(low|No) = \frac{1}{5}$ $P(fair|No) = \frac{2}{5}$

$P(yes|No) = \frac{1}{5}$

$P(No|New) = P(20|No) * P(low|No) * P(yes|No) * P(fair|No) * P(No)$

$= 0.012 * 1/5 * 1/5 * 2/5 * 5/14 = \boxed{6.85 * 10^{-5}}$

يبقى الـ New object دا بينتمي للـ class label ⇐ yes

كان الـ Prob بتاعه هي الاعلى

c) هحل الـ Discretize للـ Age في 2 Values

A , B

طول الفترة $= \frac{45 - 17}{2} = 14$

17, 21, 25, 27, 28, 29, 31, 31, 31, 32, 36, 42, 43, 45

[17 , 31) , [31 , 45]

A: 17, 21, 25, 27, 28, 29

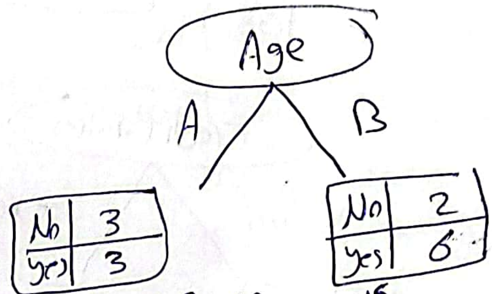
B: 31, 31, 31, 32, 36, 42, 43, 45

تعالى نشوف الـ Decision Tree لو جابت هنشوف

الـ Best Attr من الـ 4 عشان نحدد الـ Root

دي الـ Data Set الكبيرة

No	5
yes	9



$Gini(Parent) = 1 - (5/14)^2 - (9/14)^2 = 45/98$

$Gini(left) = 1 - (3/6)^2 - (3/6)^2 = 1/2$

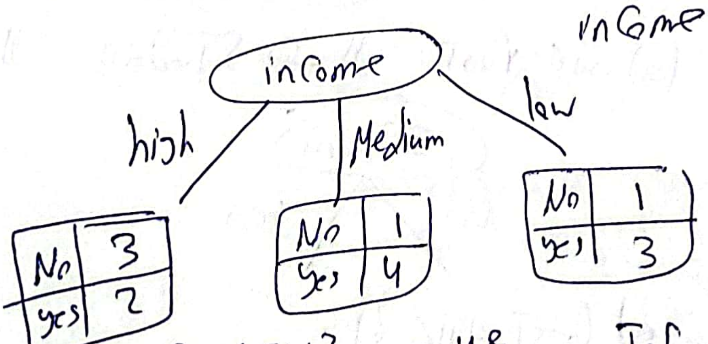
$Gini(right) = 1 - (2/8)^2 - (6/8)^2 = 3/8$

$weighted\ AVG = 6/14 * 1/2 + 8/14 * 3/8 = 3/7$

$Inform = \frac{45}{98} - 3/7 = \boxed{0.0306}$ — ①

(5)

نحل income



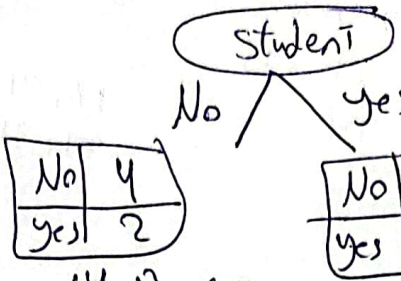
$Gini(left) = 1 - (3/5)^2 - (2/5)^2 = 0.48$

$Info = \frac{45}{98} - \frac{11}{28} = \boxed{0.0663}$ — ②

$\sim (Middle) = 1 - (1/5)^2 - (4/5)^2 = 0.32$

$\sim (low) = 1 - (1/4)^2 - (3/4)^2 = 0.375$

$weighted\ AVG = 5/14 * 0.48 + 5/14 * 0.32 + 4/14 * 0.375 = 0.392 = 11/28$



No	4
yes	2

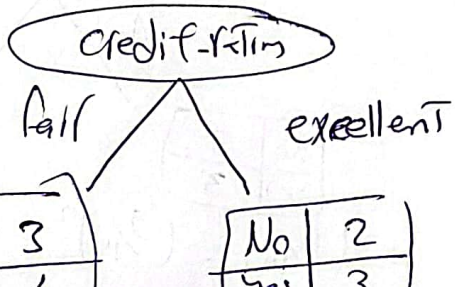
No	1
yes	7

Gini (left) $= 1 - (4/6)^2 - (2/6)^2 = 4/9$

Gini (right) $= 1 - (1/8)^2 - (7/8)^2 = 1/32$

Weighted AVG $= 6/14 * 4/9 + 8/14 * 1/32 = 0.3154 = 53/168$

Info_ $= 45/98 - 53/168 = \boxed{0.1437}$ — ③



No	3
yes	6

No	2
yes	3

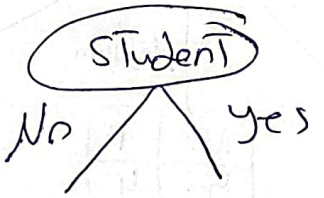
Gini (left) $= 1 - (3/9)^2 - (6/9)^2 = 4/9$

right $= 1 - (2/5)^2 - (3/5)^2 = 12/25$

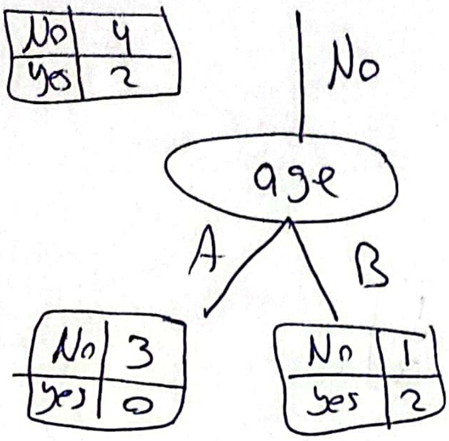
AVG $= 9/14 * 4/9 + 5/14 * 12/25 = 16/35$

Info_ $= 45/98 - 16/35 = \boxed{2.04 * 10^{-3}}$ — ④

Info_ (b) ويه root هو الـ student يعني الـ



الـ best attr. اول فرع ونتوقف عنده



نكمل الـ No الأول

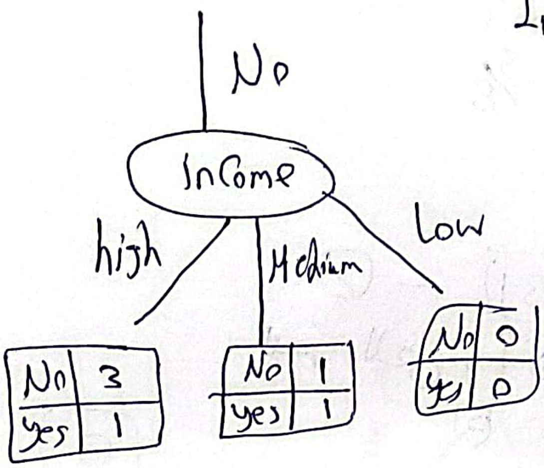
$Gini(left) = 1 - (3/3)^2 - (0/3)^2 = 0$

$Gini(right) = 1 - (1/3)^2 - (2/3)^2 = 4/9$

$AVG = 3/6 \times 0 + 3/6 \times 4/9 = 2/9$

$Gini(Parent) = 1 - (4/6)^2 - (2/6)^2 = 4/9$

① $Im\_ = 4/9 - 2/9 = 2/9 = 0.222$



$Gini(left) = 1 - (3/4)^2 - (1/4)^2 = 3/8$

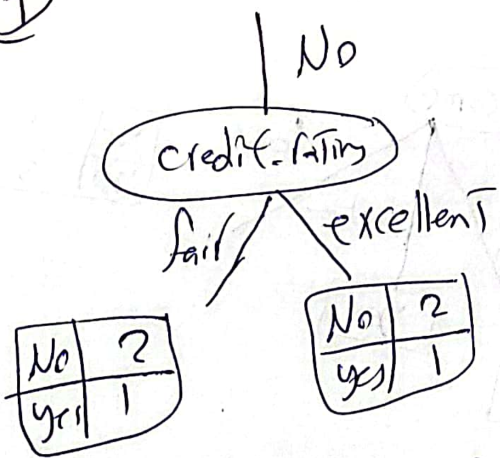
$Gini(Meddle) = 1 - (1/2)^2 - (1/2)^2 = 1/2$

$right = 0$

$AVG = 4/6 \times 3/8 + 2/6 \times 1/2 + 0 = 5/12$

② $Imf\_ = 4/9 - 5/12 = 0.027$

⑦



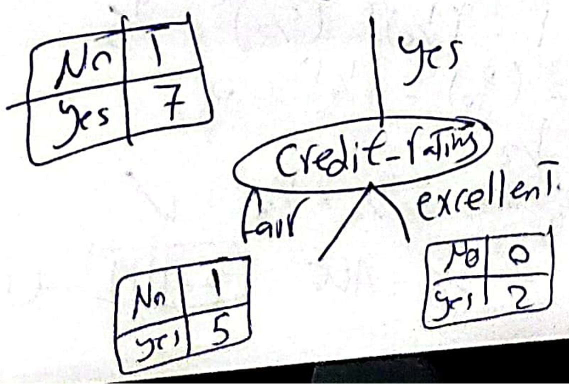
$Gini(left \& right) = 1 - (2/3)^2 - (1/3)^2 = 4/9$

$AVG = (3/6 \times 4/9) \times 2 = 4/9$

③ $Imf\_ = 4/9 - 4/9 = 0$

يبقى الـ age هو الـ [illegible] Inf + للفرع ١

نكمل للفرع بتاع الـ Yes



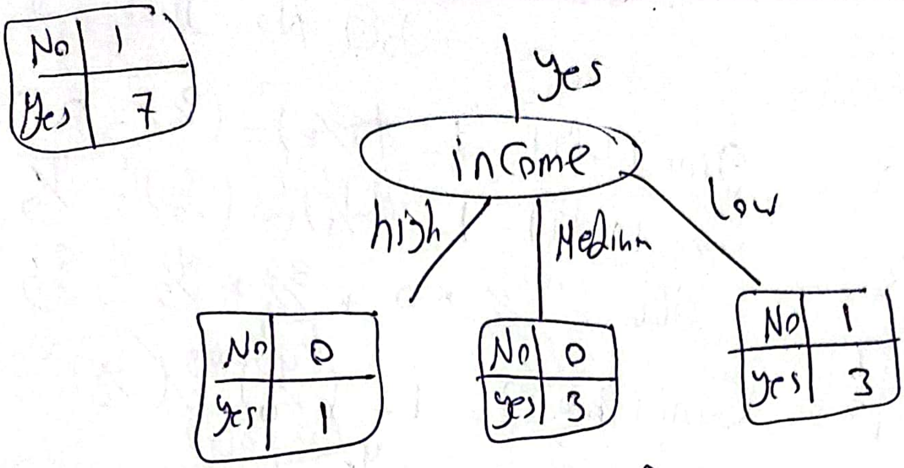
$Left = 1 - (1/6)^2 - (5/6)^2 = 5/18$

$right = 1 - (0/2)^2 - (2/2)^2 = 0$

$AVG = 6/8 \times 5/18 + 0 = 5/24$

$Gini(Parent) = 1 - (1/8)^2 - (7/8)^2 = 7/32$

① $Inf\_ = 7/32 - 5/24 = 0.0104$



$Gini(left) = 1 - (0/1)^2 - (1/1)^2 = 0$

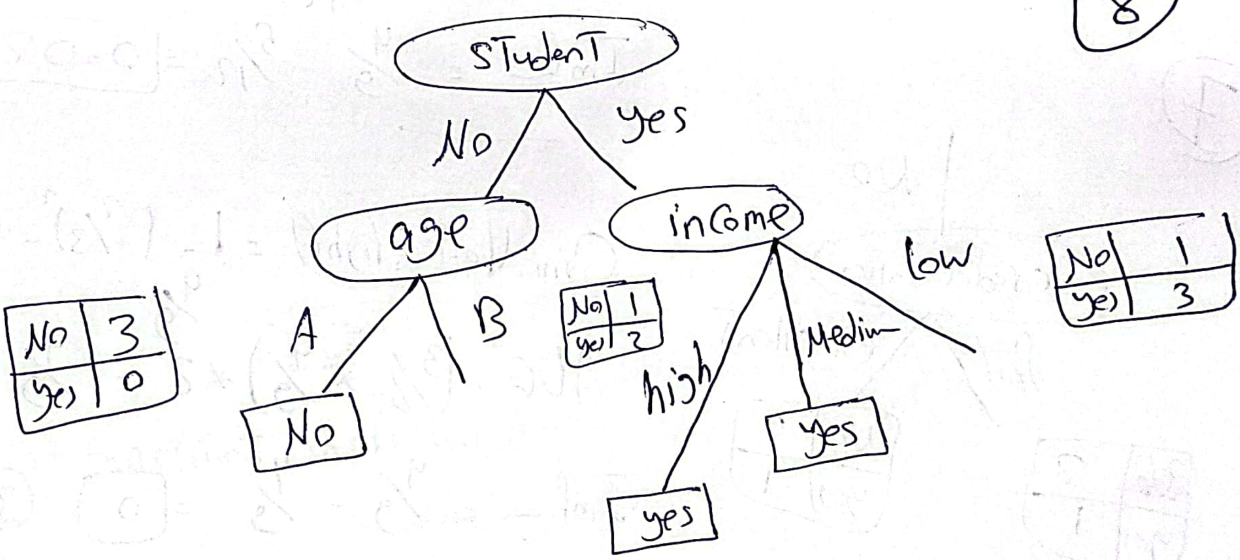
$Gini(Middle) = 1 - (0/3)^2 - (3/3)^2 = 0$

$Gini(right) = 1 - (1/4)^2 - (3/4)^2 = 3/8$

$AVG = 3/8 * 4/8 = 3/16$

$Inf = 7/32 - 3/16 = \boxed{0.031}$ — ②

لحد الفرع دا هيبقى ال ATTR الاعلى income ده
اعلى Inf



كدا هحط ال Credit_rating يا اما نحاول نشوف

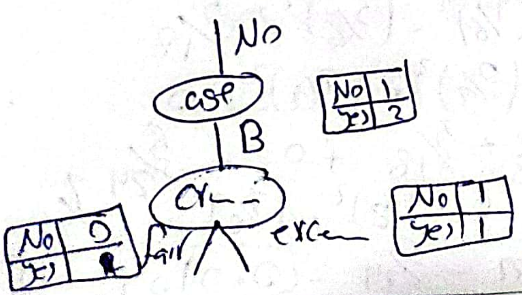
$Gini(Parent) = 1 - (1/3)^2 - (2/3)^2 = 4/9$

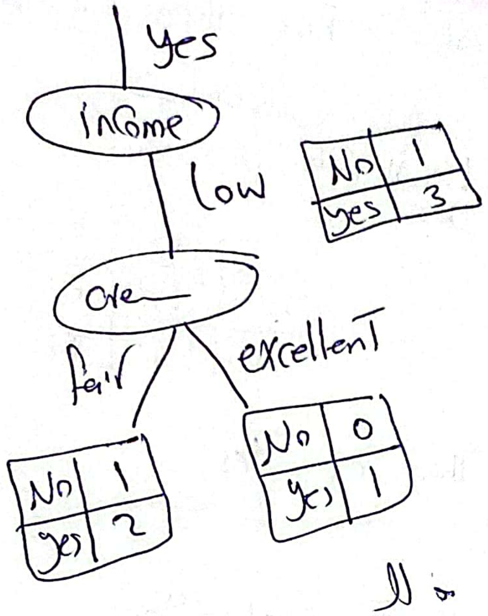
$Gini(left) = 1 - (0/1)^2 - (1/1)^2 = 0$

$Gini(right) = 1/2$

$AVG = 1/3 * 0 + 2/3 * 1/2 =$ ✓

$Inf = 4/9 - AVG = \boxed{0.1111}$ — ①





$$Gini(Parent) = 1 - (1/4)^2 - (3/4)^2 = 3/8$$

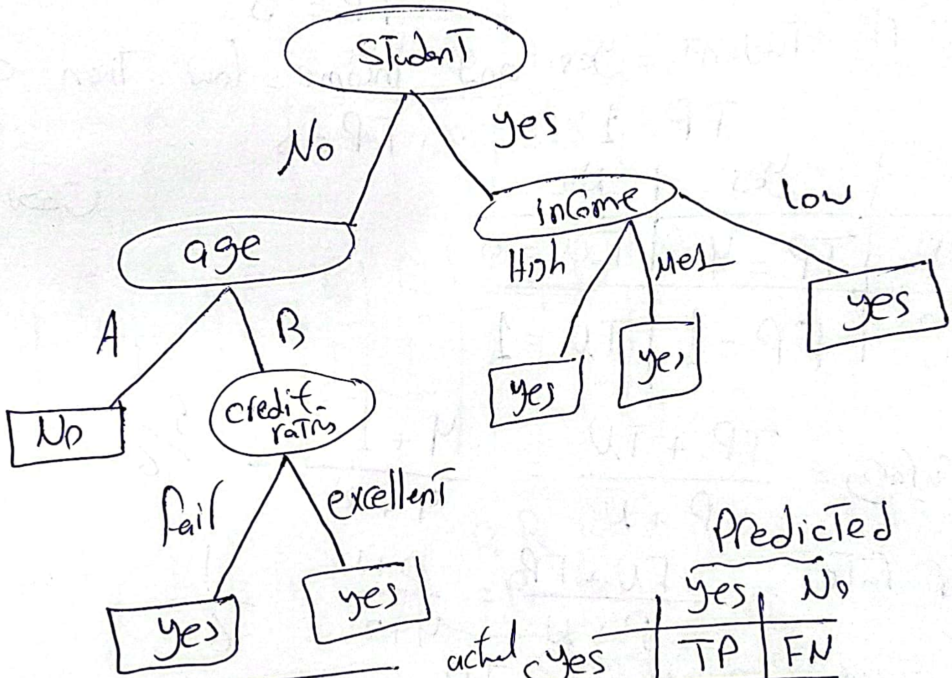
$$Gini(left) = 1 - (1/3)^2 - (2/3)^2 = 4/9$$

$$Gini(right) = 1 - (0/1)^2 - (1/1)^2 = 0$$

$$AVG = 3/4 \times 4/9 = 1/3$$

$$Inf \Rightarrow 3/8 - 1/3 = \boxed{0.041}$$

بعد أحط ال Income على الفرع التاني ف ال
Decision Tree هتبقا كدا



		Predicted Yes	Predicted No
actual	Yes	TP	FN
	No	FP	TN

d) ii) Construct Confusion Matrix

هنجيب Rule Rule من ال Tree ونشوف

Rule1: if Student = No and age = A Then CL = No

انا دلوقتي دايت Predict ال CL ب No ف كدا هشوف
هل ف ال Test Set ال rule دي هتطلع No و Yes
و انا كدا بشوف الاتنين دول

FN = 0 TN = 1

Predicted No من ال Tree — actual No من ال Test Set

* P8)

rule2: if Student = No and Age = B and cred_ = fair Then cL = yes

TP = 0 FP = 1

هنا هيبص ع
لو غيّر ال rule ال cL = yes

rule3: if Student = No and Age = B and credit_ = exellent Then cL = yes

TP = 1 FP = 0

	Yes	No
Yes	TP	FN
No	FP	TN

rule4: if Student = Yes and income = high Then cL = Yes

TP = 1 FP = 0

rule5: if Student = Yes and income = Medium Then cl = Yes

TP = 1 FP = 0

rule6: if Student = Yes and income = low Then cl = Yes

TP = 1 FP = 0

جمع مع بعض

	Yes	No
Yes	TP = 4	FN = 0
No	FP = 1	TN = 1

$$accuracy = \frac{TP + TN}{P + N} = \frac{4+1}{4+2} = 5/6$$

$$error\ rate = \frac{FN + FP}{P + N} = \frac{0+1}{4+2} = \frac{1}{6}$$

$$Sensitivity = Recall = \frac{TP}{P} = \frac{4}{4} = 1$$

$$Specificity = \frac{TN}{N} = \frac{1}{2}$$

$$Precision = \frac{TP}{TP + FP} = \frac{4}{4+1} = 4/5$$

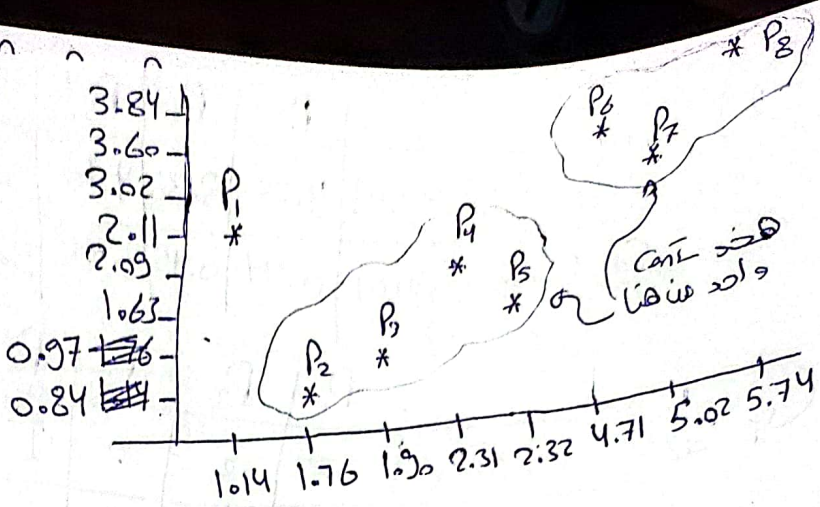
10

e) from rule 6 New object (20, low, Yes, fair, ?) is belong to class Label Yes

f) Similar

Q3

a) Scatter Plot



b) Appl K-means Algo for K=2

$C_1 = (5.02, 3.02)$, $C_2 = (2.31, 2.09)$

$D_0 =$	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
$C_1 = (5.02, 3.02)$	3.98	3.03	3.73	0	1.09	0.65	2.86	3.92
$C_2 = (2.31, 2.09)$	1.17	0.46	1.19	2.86	3.85	2.83	0	1.36

G_0

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
cluster1				1	1	1		
cluster2	1	1	1				1	1

P_4 P_5 P_6 — cluster1

P_1 P_2 P_3 P_7 P_8 — cluster2

iteration = 2

$$C_1 = \left(\frac{5.02+5.74+4.71}{3}, \frac{3.02+3.84+3.60}{3}\right)$$

$$= (5.15, 3.48)$$

$$C_2 = \left(\frac{1.14+2.32+1.90+2.31+1.76}{5}, \frac{2.11+1.63+0.97+2.09+0.84}{5}\right)$$

$$= (1.88, 1.528)$$

D1)

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
$C_1 = (5.15, 3.48)$	4.23	3.38	4.10	0.47	0.69	0.46	3.16	4.79
$C_2 = (1.88, 1.578)$	0.94	0.45	0.55	3.47	4.31	3.31	0.71	0.69

$G_1 =$

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Cluster1	0	0	0	1	1	1	0	0
Cluster2	1	1	1	0	0	0	1	1

نفس الـ G بتاعت اول iteration

وقفت STOP

P_4 P_5 P_6 — Cluster1

P_1 P_2 P_3 P_7 P_8 — Cluster2

i) Single link = أقرب المسافات بين نقطتين في الـ 2 clusters

وهما P_6 في cluster1 و P_7 في cluster2

$$Dis(P_6, P_7) = \sqrt{(4.71 - 2.31)^2 + (3.60 - 2.09)^2} = 2.83$$

ii) Complete link المسافة أبعد نقطتين

وهم P_6 , P_1

$$Dis(P_1, P_6) = \sqrt{(1.14 - 4.71)^2 + (2.11 - 3.60)^2} = 3.86$$

iii) Centeroid link أخد الـ centroid هنا وهنا واحسب المسافة بينهم

$C_1 = (5.15, 3.48)$

$C_2 = (1.88, 1.528)$

$$Dis(C_1, C_2) = \sqrt{(5.15 - 1.88)^2 + (3.48 - 1.528)} = 3.81$$

d) we choose $P_1\ P_2\ P_3\ P_4$

	P_1	P_2	P_3	P_4
P_1	0			
P_2	1.36	0		
P_3	1.37	0.78	0	
P_4	3.98	3.03	3.73	0

اصغر قيمة هي 0.78
فهعمل Merge
لـ P_2 و P_3

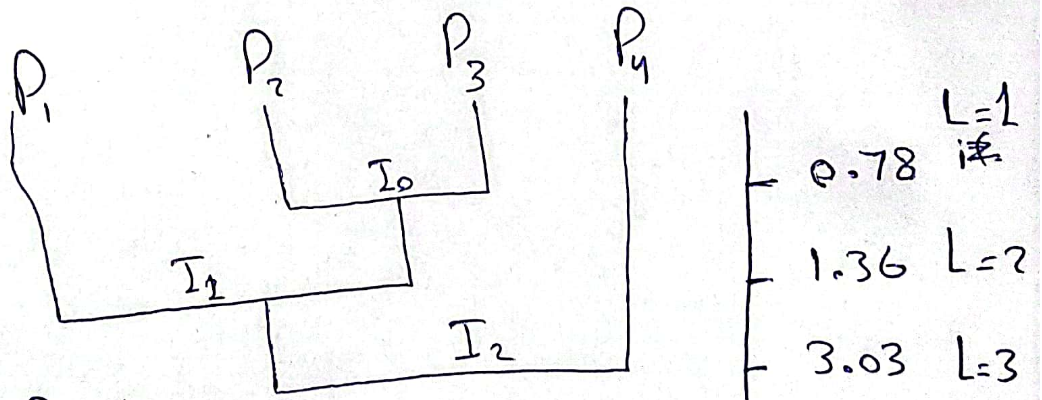
	P_1	(P_2, P_3)	P_4
P_1	0		
(P_2, P_3)	1.36	0	
P_4	3.98	3.03	0

اصغر قيمة هي 1.36
فهعمل Merge لـ
P_1 و (P_2, P_3)

	$(P_1, (P_2, P_3))$	P_4
$(P_1, (P_2, P_3))$	0	
P_4	3.03	0

بس كدا

e)



هعمل cut هنا ⇐

هيبقى result لـ

$P_1\ P_2\ P_3\ P_4$

cluster

#