MIDTERM **NAME:**

Problem 1. For the information system given below, find the set of rules describing C in terms of E,F,G. Apply CART algorithm (Gini Index). Also, assume that: Dom(E)={e1,e2}, Dom(F)={f1,f2,f3}, Dom(G)={g1,g2}, Dom(C)={c1,c2}.

X	E	F	G	C
x1	e2	f2	g2	c2
x2	e1	f3	g1	c1
x3	e2	f3	g2	c2
x4	e1	f1	g1	c2
x5	e2	f2	g1	c2
x6	e1	f2	g2	c1

Solution:
G(C)= 4/6 * 2/6 = 8/36 = 2/9
G(E/C)= 3/6[3/3*0/3] + 3/6[2/3*1/3]= 2/18 ; Gain(E/C)=4/18-2/18 = 2/18
G(F/C)= 3/6[2/3*1/3]+2/6[1/2*1/2]+1/6*0 ; Gain(E/C) is larger
G(G/C)= 3/6[2/3*1/3]+3/6[2/3*1/3] ; Gain(E/C) is larger
So, E is taken as the splitting attribute.

e2 -> c2 is the first rule [sup=3, conf=1].

New sub-table (obtained by following edge e1].

X	F	G	C
x2	f3	g1	c1
x4	f1	g1	c2
x6	f2	g2	c1

It is obvious that F is the next splitting attribute. So, the rules are:

e1*f3 -> c1 sup=1, conf=1
e1*f1 -> c2 sup=1, conf=1
e1*f2 -> c1 sup=1, conf=1

Problem 2. Find the set of all representative rules RR(3,75%) for the set of transactions: (A,B,C,E,D,F), (A,B,C,D,E,F), (A,C,D,E,H,I), (B,C,D,E,F), (A,C,F,H).

Solution:
Frequent sets generation (red color – not representative)
First Loop: A-4, B-3, C-5, D-4, E-4, F-3, H-2, I-1.
Second Loop: AB-2, AC-4, AD-3, AE-3, AF-3 , BC-3, BD-3, BE-3, BF-3 , CD-4, CE-4, CF-4 , DE-4 , DF-3 , EF -3

Third Loop: ACD-3 , ACE-3 , ACF-3 , ADE-3 , ADF-2 , AEF-2 , BCD-3 , BCE-3 , BCF-3 , BDE-3, BDF-3 , BEF-3 , CDE-4 , CDF-3, CEF-3 , DEF-3
Fourth Loop: ACDE-3 , ACDF-2 , ACEF-2 , BCDE-3, BCDF-3, BCEF-3 , BDEF-3 , CDEF-3
Fifth Loop: ACDEF-2 , BCDEF-3

So, we have 5 representative sets: BCDEF-3, ACDE-3, CDE-4, ACF-3, AC-4.

Representative rules for BCDEF-3 (in black color):

B -> CDEF, conf=3/3
C -> BDEF, conf=3/5
D -> BCEF, conf= 3/4
E -> BCDF, conf= 3/4
F -> BCDE, conf= 3/3

Representative rules for ACDE-3 (in black color):

A -> CDE, conf=3/4
C -> ADE, conf=3/5
D -> ACE, conf= 3/4
E -> ACD, conf= 3/4

We use similar strategy to find representative rules for the remaining representative sets.

Problem 3.
Use tolerance relation based strategy to find all rules describing D in terms of A, B, C in the table below.

X	A	B	C	D(NEW)	D
x1	1		2	{1,2}	1
x2		0		{1,2}	2
x3	2	2	1	{1,2}	1
x4	2			{1,2}	2
x5	1	2	1	{1}	1

Solution:
Creating new decision attribute:
d(x1)= {x1,x2}, d(x2)={x2,x1,x4},d(x3)={x3,x4},d(x4)={x4,x2,x3},d(x5)={x5}

Discernibility matrix:

	X1	X2	X3	X4
X2	-			
X3	-	-		
X4	-	-	-	
X5	C	B	A	A

F(A,B,C)= A*B*C which means {A,B,C} is the reduct
Rules generated:

For x1: C=2 -> D=1 -> D=2 or C=2 -> [D=1 or D=2] (useless)
For x2: B=0 -> D=1 -> D=2 or C=2 -> [D=1 or D=2] (useless)
For x3 & x4: A=2 -> D=1 -> D=2 or C=2 -> [D=1 or D=2] (useless)
For x5: [A=1 & B=2 & C=1] -> D=1

So, only one rule is obtained (the last one).

Problem 4.
Find optimal reduct in the table below by following RSH (RS Heuristic) strategy. Attribute h is the decision attribute.

	a	b	c	d	e	f	g	h
x1	1	0	1	0	2	1	2	1
x2	0	0	1	1	2	1	2	0
x3	1	1	0	1	2	1	1	1
x4	1	1	1	2	0	0	1	1
x5	0	2	0	2	0	0	1	0
x6	2	2	0	0	1	1	0	0
x7	2	0	1	1	1	0	0	0

Solution:

Core – Empty Set
Looking for reducts:

{a} – generates 3 rules of sup [3, 2, 2]
{b} - generates two rules of sup [2, 2]
{c} - no rules generated
{d} – no rules generated
{e} – generates one rule of sup [2]
{f} - no rules generated
{g} - generates one rule of sup [2]

So, {a} is the winner and the reduct (no tuples are left in the table).