Assignment 4 (Group 3)

Income Range	Magazine Promo	Watch Promo	Life Insurance Promo	Credit Card Insurance	Sex	Age
40-50,000	Yes	No	No	No	Male	45
30-40,000	Yes	Yes	Yes	No	Female	40
40-50,000	No	No	No	No	Male	42
30-40,000	Yes	Yes	Yes	Yes	Male	43
50-60,000	Yes	No	Yes	No	Female	38
20-30,000	No	No	No	No	Female	55
30-40,000	Yes	No	Yes	Yes	Male	35
20-30,000	No	Yes	No	No	Male	27
30-40,000	Yes	No	No	No	Male	43
30-40,000	Yes	Yes	Yes	No	Female	41
40-50,000	No	Yes	Yes	No	Female	43
20-30,000	No	Yes	Yes	No	Male	29
50-60,000	Yes	Yes	Yes	No	Female	39
40-50,000	No	Yes	No	No	Male	55
20-30,000	No	No	Yes	Yes	Female	19

Continuous valued attributes such as income range & age in this assignment has been generalized to transform to discrete valued attributes as shown below. This is done to facilitate information gain calculation as discussed during class sessions.

Income Range – {50-60,000} = High, {40-50,000, 30-40,000} = Medium, {20-30,000} = Low
Age – {45, 55} = Senior, {43, 42, 41, 40, 39, 38, 35} = Middle aged, {29, 27, 19} = Youth

Income Range	Magazine Promo	Watch Promo	Life Insurance Promo	Credit Card Insurance	Sex	Age
Medium	Yes	No	No	No	Male	Senior
Medium	Yes	Yes	Yes	No	Female	Middle aged
Medium	No	No	No	No	Male	Middle aged
Medium	Yes	Yes	Yes	Yes	Male	Middle aged
High	Yes	No	Yes	No	Female	Middle aged
Low	No	No	No	No	Female	Senior
Medium	Yes	No	Yes	Yes	Male	Middle aged
Low	No	Yes	No	No	Male	Youth
Medium	Yes	No	No	No	Male	Middle aged
Medium	Yes	Yes	Yes	No	Female	Middle aged
Medium	No	Yes	Yes	No	Female	Middle aged

Low	No	Yes	Yes	No	Male	Youth
High	Yes	Yes	Yes	No	Female	Middle aged
Medium	No	Yes	No	No	Male	Senior
Low	No	No	Yes	Yes	Female	Youth

Class label attribute distinct values = {Yes, No}, therefore,

I = - 9/15 log2 (9/15) – 6/15 log2 (6/15) = 0.97 bits.

I (Income Range) = 9/15 x (-5/9 log2 5/9 – 4/9 log2 4/9) + 2/15 x (-2/2 log2 2/2) + 4/15 x (- 2/4 log2 2/4 – 2/4 log2 2/4) = .86 bits.

Gain (Income Range) = .97 - .86 = .11 bits.

I (Magazine Promo) = 8/15 x (-6/8 log2 6/8 – 2/8 log2 2/8) + 7/15 x (-3/7 log2 3/7 – 4/7 log2 4/7)
= .89 bits.

Gain (Magazine Promo) = .97 - .89 = .08 bits.

I (Watch Promo) = 8/15 x (-6/8 log2 6/8 – 2/8 log2 2/8) + 7/15 x (-3/7 log2 3/7 – 4/7 log2 4/7)
= .89 bits.

Gain (Watch Promo) = .97 - .89 = .08 bits.

I (Credit Card Insurance) = 3/15 x (-3/3 log2 3/3) + 12/15 x (-6/12 log2 6/12 – 6/12 log2 6/12) = .80 bits

Gain (Credit Card Insurance) = .97 - .80 = .17 bits.

I (Sex) = 8/15 x (-3/8 log2 3/8 - 5/8 log2 5/8) + 7/15 x (-6/7 log2 6/7 - 1/7 log2 1/7) = .78 bits

Gain (Sex) = .97 - .78 = .19 bits

I (Age) = 3/15 x (-3/3 log2 3/3) + 9/15 x (-7/9 log2 7/9 - 2/9 log2 2/9) + 3/15 x (-2/3 log2 2/3 - 1/3 log2 1/3) = .64 bits

Gain (Age) = .97 - .64 = .33

From the above it is clear that **'Age'** has the highest information gain among the other attributes, so it is selected as the first splitting attribute and then further branched as per the attribute with next highest gain.
Also, when Age is first splitting criteria then we can see for **Age > 43** there are no Life Insurance policy takers and for the final model decision tree we can remove records for **age > 43** as below.

2’nd Iteration (Removing Age > 43 Records)

Income Range	Magazine Promo	Watch Promo	Life Insurance Promo	Credit Card Insurance	Sex	Age
30-40,000	Yes	Yes	Yes	No	Female	40
40-50,000	No	No	No	No	Male	42
30-40,000	Yes	Yes	Yes	Yes	Male	43
50-60,000	Yes	No	Yes	No	Female	38
30-40,000	Yes	No	Yes	Yes	Male	35
20-30,000	No	Yes	No	No	Male	27
30-40,000	Yes	No	No	No	Male	43
30-40,000	Yes	Yes	Yes	No	Female	41
40-50,000	No	Yes	Yes	No	Female	43
20-30,000	No	Yes	Yes	No	Male	29
50-60,000	Yes	Yes	Yes	No	Female	39
20-30,000	No	No	Yes	Yes	Female	19

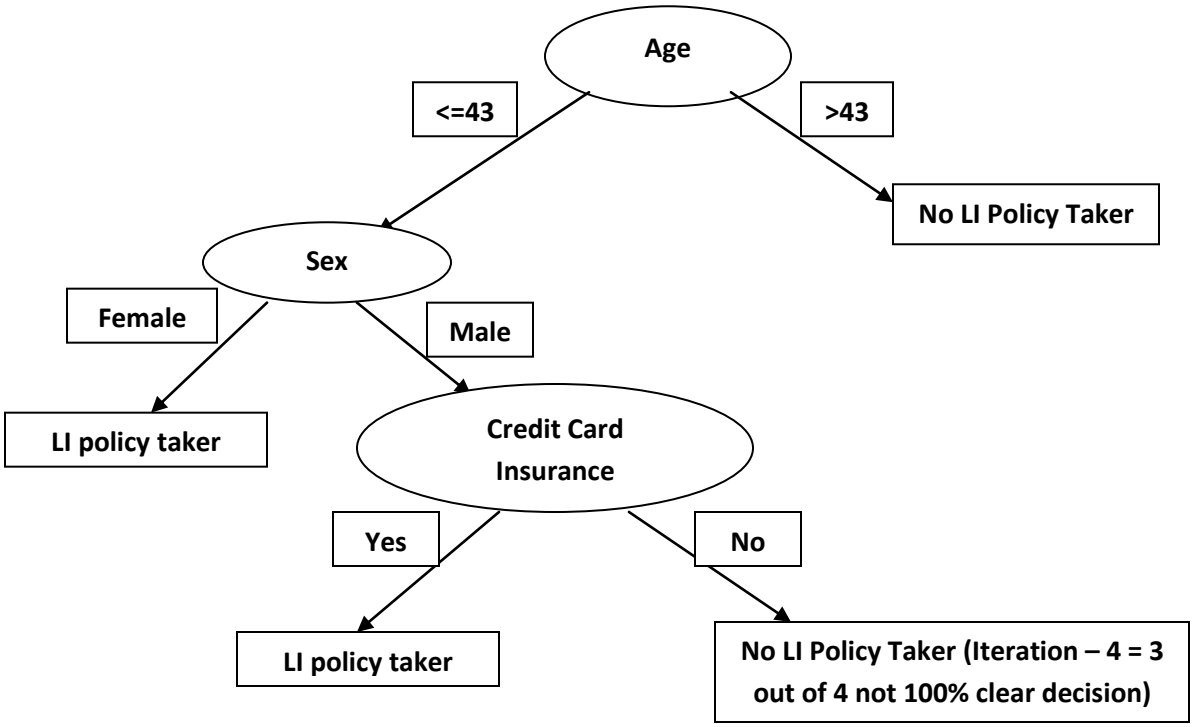
3’rd Iteration (Removing Female Records)

Income Range	Magazine Promo	Watch Promo	Life Insurance Promo	Credit Card Insurance	Sex	Age
40-50,000	No	No	No	No	Male	42
30-40,000	Yes	Yes	Yes	Yes	Male	43
30-40,000	Yes	No	Yes	Yes	Male	35
20-30,000	No	Yes	No	No	Male	27
30-40,000	Yes	No	No	No	Male	43
20-30,000	No	Yes	Yes	No	Male	29

4’th Iteration (Removing Credit Card Insurance Records)

Income Range	Magazine Promo	Watch Promo	Life Insurance Promo	Credit Card Insurance	Sex	Age
40-50,000	No	No	No	No	Male	42
20-30,000	No	Yes	No	No	Male	27
30-40,000	Yes	No	No	No	Male	43
20-30,000	No	Yes	Yes	No	Male	29

Final Decision Tree



LI = Life Insurance

Rough Work

watch promo - no of 'yes' samples = 8 > 6 li yes, 2 li no

watch promo - no of 'no' samples = 7 > 3 li yes, 4 li no

8/15 x (-6/8 log2 6/8 – 2/8 log2 2/8) + 7/15 x (-3/7 log2 3/7 – 4/7 log2 4/7)

=.53 x (.311 + .5) + .47 x (.524 + .462) = .43 + .46 = .89

==========

Credit card insurance - no of 'yes' samples = 3 > 3 li yes, 0 li no

Credit card insurance - no of 'no' samples = 12 > 6 li yes, 6 li no

3/15 x (-3/3 log2 3/3) + 12/15 x (-6/12 log2 6/12 – 6/12 log2 6/12) = .2 x (0) + .8 x (.5 + .5) = .80

==========

Sex - no of 'male' samples = 8 > 3 li yes, 5 li no

Sex - no of 'female' samples = 7 > 6 li yes, 1 li no

8/15 x (-3/8 log2 3/8 - 5/8 log2 5/8) + 7/15 x (-6/7 log2 6/7 - 1/7 log2 1/7) = .53 x (.53 + .42) + .47 x (.19 + .40) = .50 + .28 = .78

==========

Age - no of 'Senior' samples = 3 > li yes 0, li no 3

Age - no of 'Middle aged' samples = 9 > li yes 7, li no 2

Age - no of 'Youth' samples = 3 > li yes 2, li no 1

3/15 x (-3/3 log2 3/3) + 9/15 x (-7/9 log2 7/9 - 2/9 log2 2/9) + 3/15 x (-2/3 log2 2/3 - 1/3 log2 1/3) = .6 x (.28 + .48) + .2 x (.39 + .53) = .46 + .18 = .64