Name - Shubham Kumar
Roll No - IIT2018146

ML Assignment

Problem 1

Sol^n :- Given

Age	Lable
37	0
41	0
44	0
48	1
49	0
52	0
53	0
54	1
56	0
56	0
56	1
56	0
37	0
57	0
57	0
62	1
63	1
63	0
67	1
67	1

0 → No heart disease
1 → Heart disease.

Here, due to 1 feature, only 1 level decision tree will be presence.

To find the threshold value, we will select the one with the least gini Impurity

Gini Impurity Formula.

$$GI = 1 - (P_{yes})^2 - (P_{No})^2$$

P = Probablity.

For non leaf node GI is the weighted average of GI of lead node.

Let N_L = # sample in left node (age <= t)
N_R = " " right " (age > t)
$(GI)_L$ = Gini Imp on left "
$(GI)_R$ = " " right node
GI = Total Gini Imp.

t	N_L	N_R	$(GI)_L$	$(GI)_R$	$(GI)_{Tot}$
37	1	19	0	0.46	0.44
41	2	18	0	0.47	0.42
44	3	17	0	0.48	0.41
48	5	15	0.32	0.48	0.44
49	6	14	0.27	0.48	0.42
52	7	13	0.24	0.49	0.40
53	8	12	0.37	0.48	0.44
54	9	11	0 34	0.49	0.44
56	12	8	0.37	0.5	0.44
57	15	5	0 32	0.32	0.42
62	16	4	0.37	0.37	0.32
63	18	2	0.40	0	0.375

let $t = 37$

with age $\leq 37 = 1$

$P_{yes} = 0$

$P_{no} = 1$

$(GI)_L = 1 - 0^2 - 1^2 = 0$

with age $> 37 = 19$

$P_{yes} = \frac{7}{19}$

$P_{no} = \frac{12}{19}$

$(GI)_R = 1 - (0.36)^2 - (0.63)^2$

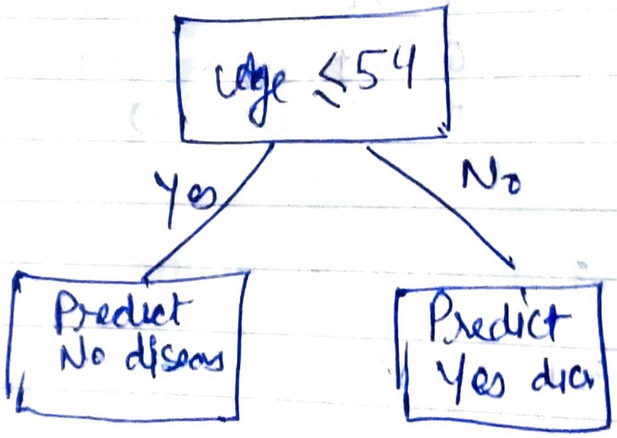
$= 0.465$

Similarly calculate for all GI .

So It can be observed that this has
11 min GI at Age = 54

12 The GI is 0.43649 . .

So decision tree



Problem 2

Solⁿ

Slope	Label
1	0
1	0
3	0
1	1
1	0
1	0
3	0
1	1
2	0
2	0
1	1
2	0
1	0
1	0
3	0
2	1
3	1
2	0
2	1
	1

Only one feature name slope is available so one level decision tree will be there

$$\text{Entropy}(S) = -P_{yes}\log(P_{yes}) - P_{No}\log(P_{No})$$

Information gain
(IG) = Entropy (S) – [Average entropy of children]

Calculate E(S)

Total entropy

$E(S) = -P_{yes} \log(P_{yes}) - P_{NO}(\log(P_{NO}))$

$= \left(\frac{-7}{20}\right)\log\left(\frac{7}{20}\right) - \left(\frac{13}{20}\right)\log\left(\frac{13}{20}\right)$

$= 0.934$

(i) Threshold = 1
# of sample, with slope <= 1 : 9

$P_{yes} = 0$
$P_{NO} = 1$
$E(\text{Left}) = 0$

# sample with slope > 1 : 11
$P_{yes} = 7/11$
$P_{NO} = 4/11$
$E(\text{Right}) = 0.946$

$IG = E(S) -$ weight av entropy

$= 0.934 - \frac{1}{20} \times 0.946$

$= 0.4137$

(ii) Threshold = 2

sample with slope <= 2 = 15

$P_{yes} = 3/15 = 0.2$

$P_{NO} = 12/15 = 0.8$

$E(L) = 0.722$

sample with slope > 2 = 5

$P_{yes} = 3/5$

$P_{NO} = 2/5$

$E(R) = 0.971$

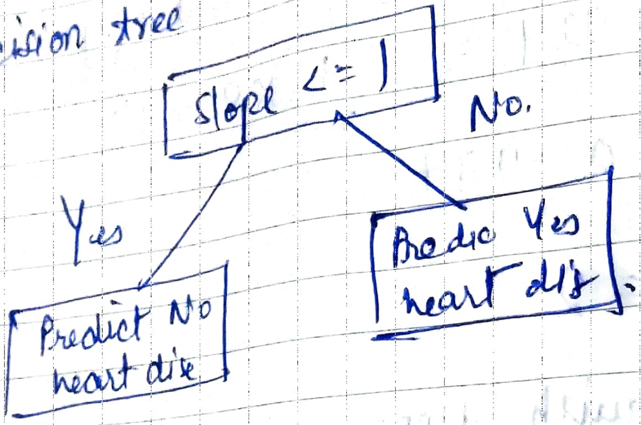
$IG = 0.934 - \left(\frac{15}{20}\right) \times 0.722 - \frac{5}{20}(0.971)$

$= 0.14975$

$IG(i) > IG(ii)$

slope = 1 is better.

Decision tree



9

10

11

12

1

2

3

4

5