ML ASSIGNMENT 00 - DECISION TREE

Kartic Chouley IIT2018181

Problem 1: Solve with gini impurity for age & label as feature.

S. No	Age	Label
1	37	0
2	41	0
3	44	0
4	48	1
5	48	0
6	49	0
7	52	0
8	53	1
9	54	0
10	56	0
11	56	1
12	56	0
13	57	0
14	57	0
15	57	0
16	62	1
17	63	1
18	63	0
19	67	1
20	67	1

Label = 0 (No Heart Disease), Label = 1 (Heart Disease)

⇒ As There is only one feature (age), thats why we will have one level decision tree.

⇒ Now we have to calculate the threshold value of root node, which is the least gini impurity if we consider all the value which are present in the table on page number 1.

* For a leaf node, gini impurity is given as :

$$\text{Gini Impurity} = 1 - (\text{Probability of Yes})^2 - (\text{Probability of NO})^2$$

* For a non leaf node, the gini impurity is the weighted average of gini impurity of leaf nodes.

* Let T → threshold value
  - N_{left} → Number of samples in left node (age ≤ T).
  - N_{right} → Number of samples in right node (age > T)
  - $(GNI)_{left}$ → Gini Impurity of left node.
  - $(GNI)_{right}$ → Gini Impurity of right node
  - $(GNI)_{tot}$ → Total Gini Impurity.

⇒ Formula is :

$$(GNI)_{tot} = \frac{N_{left}}{N} \times (GNI)_{left} + \frac{N_{Right}}{N} \times (GNI)_{right}.$$

GINI IMPURITY CALCULATIONS FOR ALL AGES

T	N_{Left}	N_{right}	$(GNI)_{left}$	$(GNI)_{right}$	$(GNI)_{tot}$
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	0.43
56	12	8	0.37	0.5	0.42
57	15	5	0.32	0.32	0.32
62	16	4	0.37	0.37	0.375
63	18	2	0.40	0	0.36

⇒ Consider T = 37

⇒ Number of samples with age ≤ 37 : 1

$P_{Yes} = 0/1 = 0$, $P_{No} = 1/1 = 1$

$(GNI)_{left} = 1 - 0^2 - 1^2 = 0$

⇒ Number of samples with age > 37 : 19

$P_{Yes} = (7/19) = 0.36$

$P_{No} = (12/19) = 0.63$

$(GNI)_{right} = 1 - (0.36)^2 - (0.63)^2$

$= 0.465$

$$(GNI)_{tot} = \frac{N_{left}}{N} \times (GNI)_{left} + \frac{N_{right}}{N} \times (GNI)_{right}$$

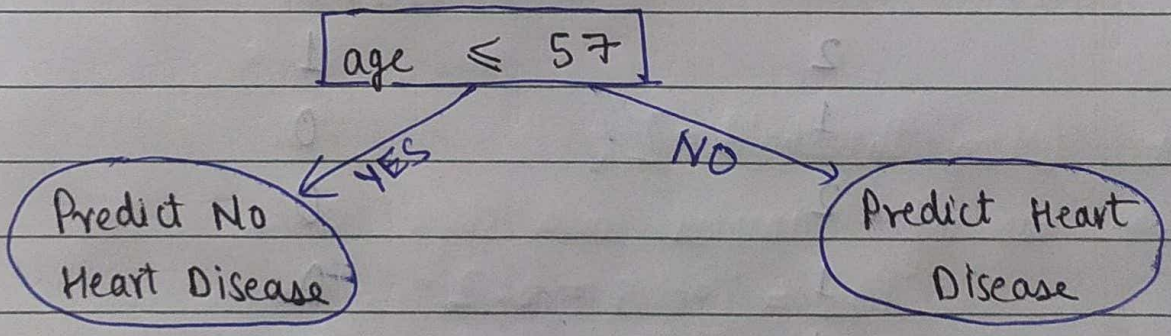
$$= \frac{1}{20} \times 0 + \frac{19}{20} \times 0.465$$

$$= 0.442.$$

⇒ Similarly, we can calculate GNI for all thresholds. I calculated it and mentioned it in Table on page 3.

⇒ From the table, we can conclude that T is optimal for. Age = 57. ($GNI_{tot} = 0.32$) → Least Gini Impurity.

DECISION TREE



Problem 2: Solve with Information Gain for slope & label as feature.

S.NO	Slope	Label
1	3	0
2	1	0
3	1	0
4	3	1
5	1	0
6	1	0
7	1	0
8	3	1
9	1	0
10	2	0
11	2	1
12	1	0
13	2	0
14	1	0
15	1	0
16	3	1
17	2	1
18	3	0
19	2	1
20	2	1

Label = 0 (No Heart Disease), Label = 1 (Heart Disease)

⇒ As there is only one feature (slope), thatswhy we will have one level decision tree.

⇒ We need to find that slope, for which the information gain is the highest.

→ Value of threshold can be 1 or 2. If we take threshold = 3 then it will include all the samples. So we need to find best threshold energy among 1 & 2.

Formula Used :-

$$\text{Entropy } (S) = -P(\text{Yes}) \cdot \log P(\text{Yes}) - P(\text{No}) \cdot \log P(\text{No})$$

$$\text{Information Gain } (IG) = S - (\text{Average entropy of children})$$

$$\text{Total Entropy, } E(S) = -P(\text{Yes}) \cdot \log P(\text{Yes}) - P(\text{No}) \cdot \log P(\text{No})$$

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

$$= 0.934.$$

* When threshold = 1,

→ No. of Samples with slope <= 1 : 9.

$P_{yes} = 0$, $P_{No} = 1$

Entropy (left) = 0 (As $-0.\log 0 - 1.\log 1 = 0$)

→ No. of Samples with slope > 1 : 11

$P_{yes} = 7/11$, $P_{No} = 4/11$

$$\text{Entropy (Right)} = -\left(\frac{7}{11}\right)\log\left(\frac{7}{11}\right) - \left(\frac{4}{11}\right)\log\left(\frac{4}{11}\right) = 0.946$$

→ Information Gain (IG) = E(S) − Avg Entropy of Children

$$= 0.934 - 0 - \frac{11}{20} \times 0.946$$

$$= 0.4137$$

* When threshold = 2,

→ No of Samples with slope <= 2 : 15

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

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

$E(L) = 0.722$ $(-0.2 \cdot \log(0.2) - 0.8 \log(0.8))$

→ No. of Samples with slope > 2 : 5.

$P_{yes} = 3/5 = 0.6$

$P_{No} = 2/5 = 0.4$

$E(R) = 0.971$ $(-0.6 \cdot \log(0.6) - 0.4(\log(0.4)))$

→ Information Gain (IG) = E(S) - Average Entropy of Children

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

$$= 0.14975.$$

★ IG for Slope = 1 > IG for Slope = 2

Hence Slope = 1 is the better threshold.

DECISION TREE