DECISION TREE USING GINI IMPURITY.

PRATEEK MISHRA
IIT2018199

Dataset:

	Age	Y
1	54	0
2	51	0
3	70	0
4	53	0
5	45	1
6	74	1
7	70	1
8	57	1
9	59	1
10	40	1
11	65	0
12	61	0
13	58	0
14	47	1
15	64	0
16	59	1
17	63	0
18	37	0
19	35	0
20	45	0

sorted according to Ages

1	29	0
2	34	0
3	34	0
4	35	0
5	35	1
6	35	1
7	35	0
8	37	0
9	37	1
10	38	0
11	38	1
12	39	0
13	39	1
14	39	0
15	39	1
16	40	0
17	40	1
18	40	0
19	41	0
20	41	0

now calculating gini impurity for each of these distinct ages:

GINI IMPURITY = Weighted average of less than and greater than probabilities.

LESS THAN
GINI IMPURITY = $1 - (\text{less than Yes Probability})^2 - (\text{less than No Probability})^2$

GREATER THAN
GINI IMPURITY = $1 - (\text{greater than Yes Probability})^2 - (\text{less than Yes Probability})^2$

Greater than Yes Probability = $\frac{\text{Greater than Yes}}{\text{Greater than Yes} + \text{Greater than No}}$

Greater than No Probability = $\frac{\text{Greater than No}}{\text{Greater than No} + \text{Greater than Yes}}$

Similarly formulas for less than Yes and less than No.

The final obtained gini impurities for all ages:

Age	Gini Impurity:
34.5	0.388
35	0.419
36	0.419
37.5	0.410
38.5	0.418
39.5	0.4133
40.5	0.399

Thus the best Rule would be 34.5 with impurity 0.388.

Note that age is a continous feature and we have made it discrete using average of two adjacent values in the dataset.