Example: Building a decision tree using Hunt's Algorithm

Impurity measure: Gini index

Splitting approach: Binary split

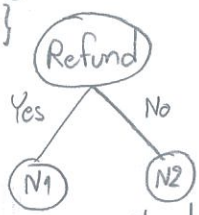
$$Gini(t) = 1 - \sum_j [p(j|t)]^2$$

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Normal attributes | class attribute

Step 1 Determining the best split for the root node

(1) Refund = {Yes, No}



Cheat → Yes = 0 Cheat → Yes = 3

Cheat → No = 3 Cheat → No = 4

$$GINI(N1) = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = 0$$

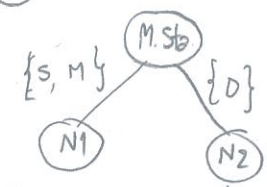
$$GINI(N2) = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0.49$$

$$GINI_{split} = \left(\frac{3}{10}\right)(0) + \left(\frac{7}{10}\right)(0.49) = 0.343$$

---

(2) Marital Status = {Single, Married, Divorced}

(2.1)



Cheat → Yes = 2 Cheat → Yes = 1

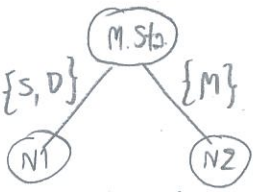
Cheat → No = 6 Cheat → No = 1

$$GINI(N1) = 1 - \left[\left(\frac{2}{8}\right)^2 + \left(\frac{6}{8}\right)^2\right] = 0.375$$

$$GINI(N2) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$$

$$GINI_{split} = \left(\frac{8}{10}\right)(0.375) + \left(\frac{2}{10}\right)(0.5) = 0.4$$

(2.2)



Cheat → Yes = 3 Cheat → Yes = 0

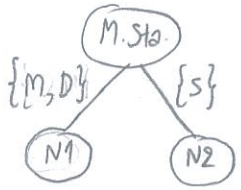
Cheat → No = 3 Cheat → No = 4

$$GINI(N1) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 0.5$$

$$GINI(N2) = 1 - \left[\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2\right] = 0$$

$$GINI_{split} = \left(\frac{6}{10}\right)(0.5) + \left(\frac{4}{10}\right)(0) = 0.3$$

(2.3)



Cheat → Yes = 1 Cheat → Yes = 2

Cheat → No = 5 Cheat → No = 2

$$GINI(N1) = 1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right] = 0.28$$

$$GINI(N2) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$$

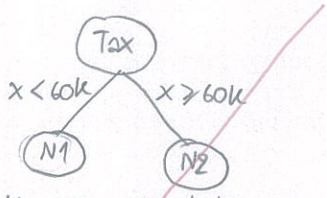
$$GINI_{split} = \left(\frac{6}{10}\right)(0.28) + \left(\frac{4}{10}\right)(0.5) = 0.368$$

Cheat → N N N Y Y Y N N N N

(3) Taxable Income = {60k, 70k, 75k, 85k, 90k, 95k, 100k, 120k, 125k, 220k}

In this case, we have 10 distinct values, thus we have to use these values as a splitting point.

(3.1)



Cheat → Yes = 0 Cheat → Yes = 3

Cheat → No = 0 Cheat → No = 7

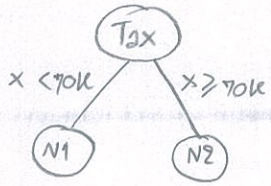
$GINI(N1) = 1 - \left[\left(\frac{0}{0}\right)^2 + \left(\frac{0}{0}\right)^2\right] \approx 1$ (indetmed)

$GINI(N2) = 1 - \left[\left(\frac{3}{10}\right)^2 + \left(\frac{7}{10}\right)^2\right] = 0.42$

$GINI_{split} = \left(\frac{0}{10}\right)(1) + \left(\frac{10}{10}\right)(0.42)$

$= 0.42$

(3.2)



Cheat → Yes = 0 Cheat → Yes = 3

Cheat → No = 1 Cheat → No = 6

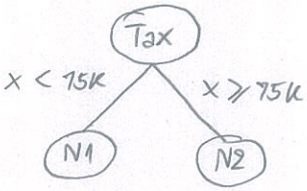
$GINI(N1) = 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right] = 0$

$GINI(N2) = 1 - \left[\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right] = 0.44$

$GINI_{split} = \left(\frac{1}{10}\right)(0) + \left(\frac{9}{10}\right)(0.44)$

$= 0.4$

(3.3)



Cheat → Yes = 0 Cheat → Yes = 3

Cheat → No = 2 Cheat → No = 5

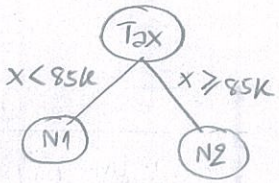
$GINI(N1) = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$

$GINI(N2) = 1 - \left[\left(\frac{3}{8}\right)^2 + \left(\frac{5}{8}\right)^2\right] = 0.469$

$GINI_{split} = \left(\frac{2}{10}\right)(0) + \left(\frac{8}{10}\right)(0.469)$

$= 0.375$

(3.4)



Cheat → Yes = 0 Cheat → Yes = 3

Cheat → No = 3 Cheat → No = 4

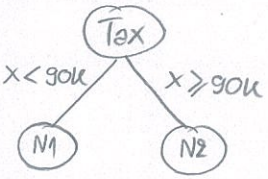
$GINI(N1) = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = 0$

$GINI(N2) = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0.49$

$GINI_{split} = \left(\frac{3}{10}\right)(0) + \left(\frac{7}{10}\right)(0.49)$

$= 0.343$

(3.5)



Cheat → Yes = 1 Cheat → Yes = 2

Cheat → No = 3 Cheat → No = 4

$GINI(N1) = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right]$

$= 0.375$

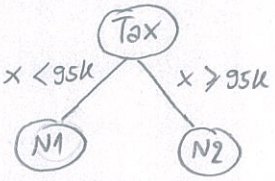
$GINI(N2) = 1 - \left[\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right]$

$= 0.44$

$GINI_{split} = \left(\frac{4}{10}\right)(0.375) + \left(\frac{6}{10}\right)(0.44)$

$= 0.414$

(3.6)



Cheat → Yes = 2 Cheat → Yes = 1

Cheat → No = 3 Cheat → No = 4

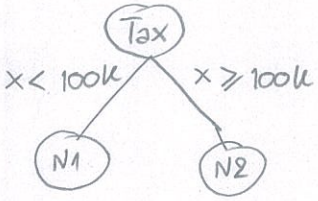
$GINI(N1) = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right] = 0.48$

$GINI(N2) = 1 - \left[\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right] = 0.32$

$GINI_{split} = \left(\frac{5}{10}\right)(0.48) + \left(\frac{5}{10}\right)(0.32)$

$= 0.4$

(3.7)



Cheat → Yes = 3 Cheat → Yes = 0

Cheat → No = 3 Cheat → No = 4

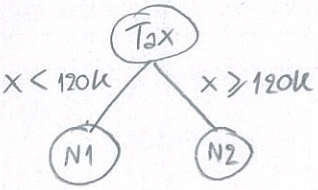
$GINI(N1) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 0.5$

$GINI(N2) = 1 - \left[\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2\right] = 0$

$GINI_{split} = \left(\frac{6}{10}\right)(0.5) + \left(\frac{4}{10}\right)(0)$

$= 0.3$

(3.8)



Cheat → Yes = 3 Cheat → Yes = 0

Cheat → No = 4 Cheat → No = 3

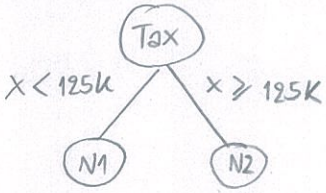
$GINI(N1) = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0.49$

$GINI(N2) = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = 0$

$GINI_{split} = \left(\frac{7}{10}\right)(0.49) + \left(\frac{3}{10}\right)(0)$

$= 0.343$

(3.9)



Cheat → Yes = 3 Cheat → Yes = 0

Cheat → No = 5 Cheat → No = 2

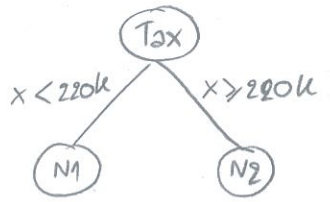
$GINI(N1) = 1 - \left[\left(\frac{3}{8}\right)^2 + \left(\frac{5}{8}\right)^2\right] = 0.469$

$GINI(N2) = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$

$GINI_{split} = \left(\frac{8}{10}\right)(0.469) + \left(\frac{2}{10}\right)(0)$

$= 0.375$

(3.10)



Cheat → Yes = 3 Cheat → Yes = 0

Cheat → No = 6 Cheat → No = 1

$GINI(N1) = 1 - \left[\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right] = 0.44$

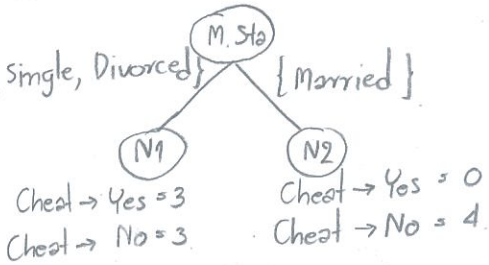
$GINI(N2) = 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right] = 0$

$GINI_{split} = \left(\frac{9}{10}\right)(0.44) + \left(\frac{1}{10}\right)(0)$

$= 0.4$

Minimum { ① 0.343, ②.① 0.4, ②.② 0.3, ②.③ 0.368, ③.① 0.42, ③.② 0.4, ③.③ 0.375, ③.④ 0.343, ③.⑤ 0.414, ③.⑥ 0.4, ③.⑦ 0.3, ③.⑧ 0.343, ③.⑨ 0.375, ③.⑩ 0.4 } = 0.3 of ③

∴ Since the minimum $GINI_{split}$ is obtained from test condition 2.2 and 3.7, we can use either test condition 2.2 or 3.7 as a test condition for the root node. In this case, I choose 2.2.



Step 2 : Check whether we should stop splitting the records at node N1.

Two main stopping criteria :

1) Stop expanding a node when all the records belong to the same class
2) Stop expanding a node when all the records have identical attribute values

∴ Since all the records at node N1 do not satisfy the two main stopping criteria, then we need to split these records further.

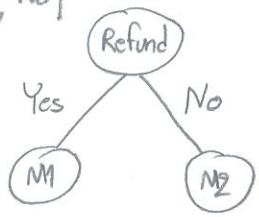
Step 3 : Check whether we should stop splitting the records at node N2.

∴ Since all the records at node N2 satisfy the first main stopping criterion, then we stop splitting and declare this node as a leaf node of value No.

Step 4 : Determine the best split for node N1. (Now, we have only 6 records left at node N1)

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

① Refund = {Yes, No}



Cheat → Yes = 0 Cheat → Yes = 3

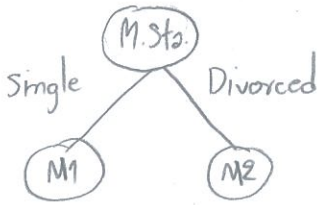
Cheat → No = 2 Cheat → No = 1

$GINI(N1) = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$

$GINI(N2) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$

$GINI_{split} = \left(\frac{2}{6}\right)(0) + \left(\frac{4}{6}\right)(0.375) = 0.25$

(2) Marital Status = { Single, Divorced }



Cheat → Yes = 2 Cheat → Yes = 1

Cheat → No = 2 Cheat → No = 1

$GINI(N1) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$

$GINI(N2) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$

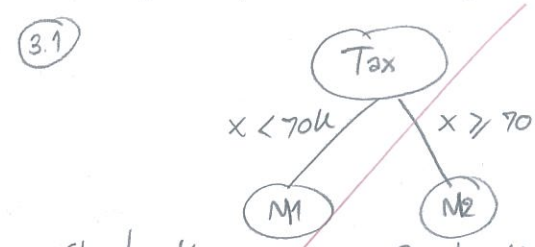
$GINI_{split} = \left(\frac{4}{6}\right)(0.5) + \left(\frac{2}{6}\right)(0.5)$

$= 0.5$

(3) Taxable Income = { 70k (N), 85k (Y), 90k (Y), 95k (Y), 125k (N), 220k (N) }

(4)

(3.1)



Cheat → Yes = 0 Cheat → Yes = 3

Cheat → No = 0 Cheat → No = 3

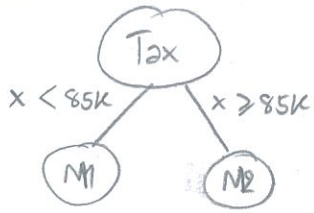
$GINI(N1) = 1 - \left[\left(\frac{0}{0}\right)^2 + \left(\frac{0}{0}\right)^2\right] \approx 1$ (undefined)

$GINI(N2) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 0.5$

$GINI_{split} = \left(\frac{0}{6}\right)(1) + \left(\frac{6}{6}\right)(0.5)$

$= 0.5$

(3.2)



Cheat → Yes = 0 Cheat → Yes = 3

Cheat → No = 1 Cheat → No = 2

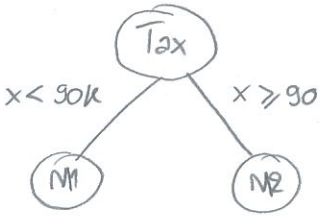
$GINI(N1) = 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right] = 0$

$GINI(N2) = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0.48$

$GINI_{split} = \left(\frac{1}{6}\right)(0) + \left(\frac{5}{6}\right)(0.48)$

$= 0.4$

(3.3)



Cheat → Yes = 1 Cheat → Yes = 2

Cheat → No = 1 Cheat → No = 2

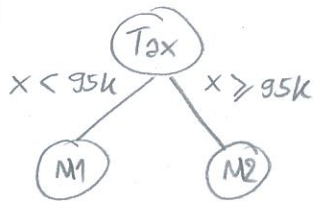
$GINI(N1) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$

$GINI(N2) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$

$GINI_{split} = \left(\frac{2}{6}\right)(0.5) + \left(\frac{4}{6}\right)(0.5)$

$= 0.5$

(3.4)



Cheat → Yes = 2 Cheat → Yes = 1

Cheat → No = 1 Cheat → No = 2

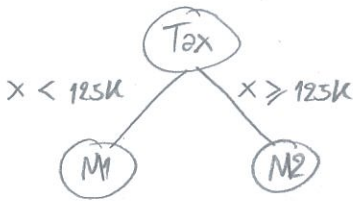
$GINI(N1) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.44$

$GINI(N2) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.44$

$GINI_{split} = \left(\frac{3}{6}\right)(0.44) + \left(\frac{3}{6}\right)(0.44)$

$= 0.44$

(3.5)



Cheat → Yes = 3 Cheat → Yes = 0

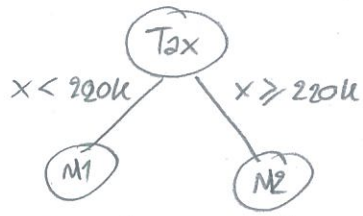
Cheat → No = 1 Cheat → No = 2

$GINI(N1) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$

$GINI(N2) = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$

$GINI_{split} = \left(\frac{4}{6}\right)(0.375) + \left(\frac{2}{6}\right)(0)$

$= 0.25$



Cheat → Yes = 3 Cheat → Yes = 0

Cheat → No = 2 Cheat → No = 1

$GINI(N1) = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0.48$

$GINI(N2) = 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right] = 0$

$GINI_{split} = \left(\frac{5}{6}\right)(0.48) + \left(\frac{1}{6}\right)(0)$

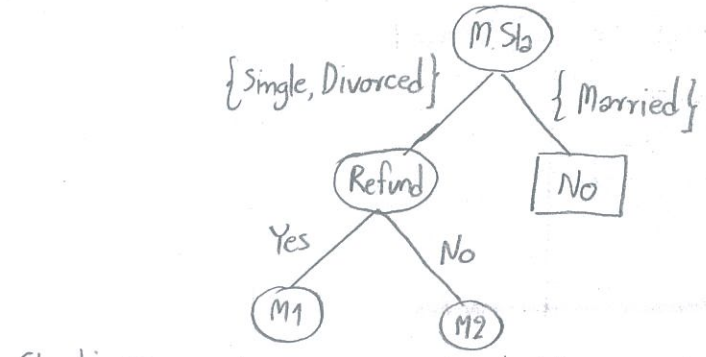
$= 0.4$

Minimum { (1) 0.25, (2) 0.5, (3.1) 0.5, (3.2) 0.4, (3.3) 0.5, (3.4) 0.44, (3.5) 0.25, (3.6) 0.4 }

$= 0.25$

Since the minimum $GINI_{split}$ of 0.25 is obtained from test condition 1 and 3.5, thus we can use either test condition 1 or 3.5 as a test condition for node N1.

In this case, I choose test condition 1.



Step 5 : Check whether we should stop splitting the records at node M1.

∴ Since all the records at node M1 satisfy the first main stopping criterion, then we stop splitting and declare this node as a leaf node of value No.

Step 6 : Check whether we should stop splitting the records at node M2.

∴ Since all the records at node M2 do not satisfy the two main stopping criteria, then we need to split thess records further.

Repeat step 4, 5, and 6. until we cannot split further.