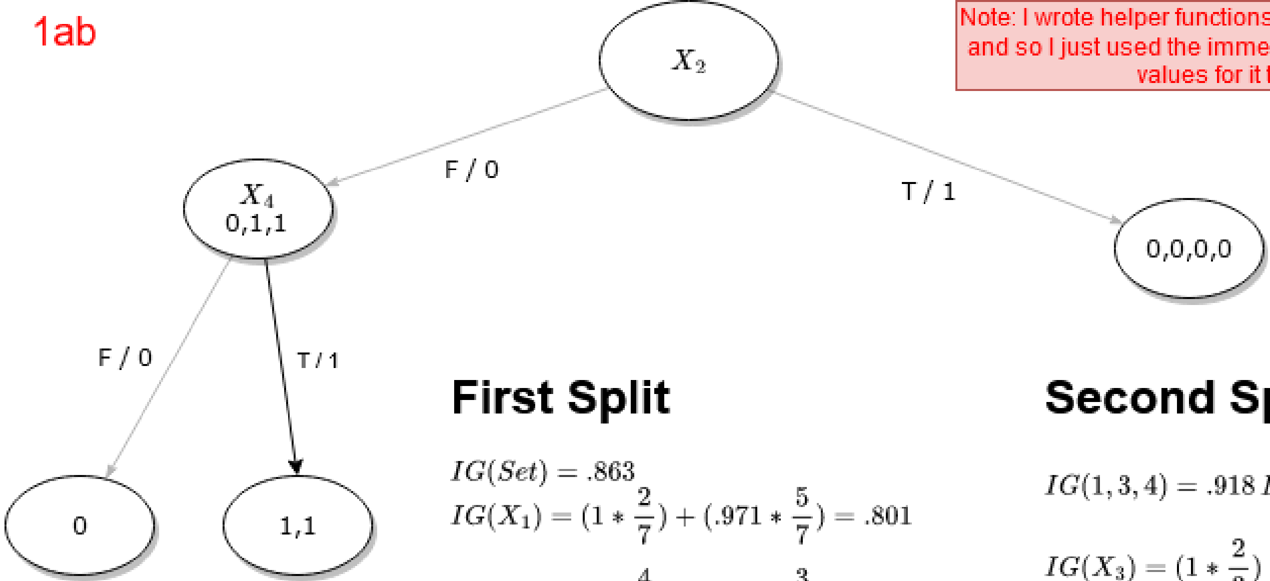
1ab

Note: I wrote helper functions to help with the written part and so I just used the immediate window and input the values for it to calculate



First Split

$IG(Set) = .863$

$IG(X_1) = (1 * \frac{2}{7}) + (.971 * \frac{5}{7}) = .801$

$IG(X_2) = (0 * \frac{4}{7}) + (.918 * \frac{3}{7}) = .393$

$IG(X_3) = (.917 * \frac{3}{7}) + (.924 * \frac{4}{7}) = .861$

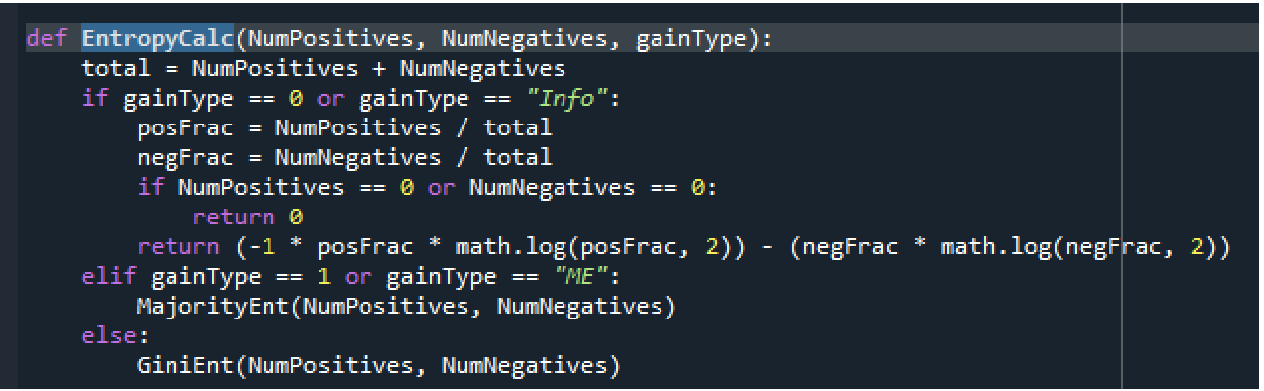
$IG(X_4) = (.917 * \frac{3}{7}) + (.924 * \frac{4}{7}) = .861$

Second Split

$IG(1,3,4) = .918\ IG(X_1) = (1 * \frac{2}{3}) + (0 * \frac{1}{3}) = .667$

$IG(X_3) = (1 * \frac{2}{3}) + (0 * \frac{1}{3}) = .667$

$IG(X_4) = (0 * \frac{2}{3}) + (0 * \frac{1}{3}) = 0$

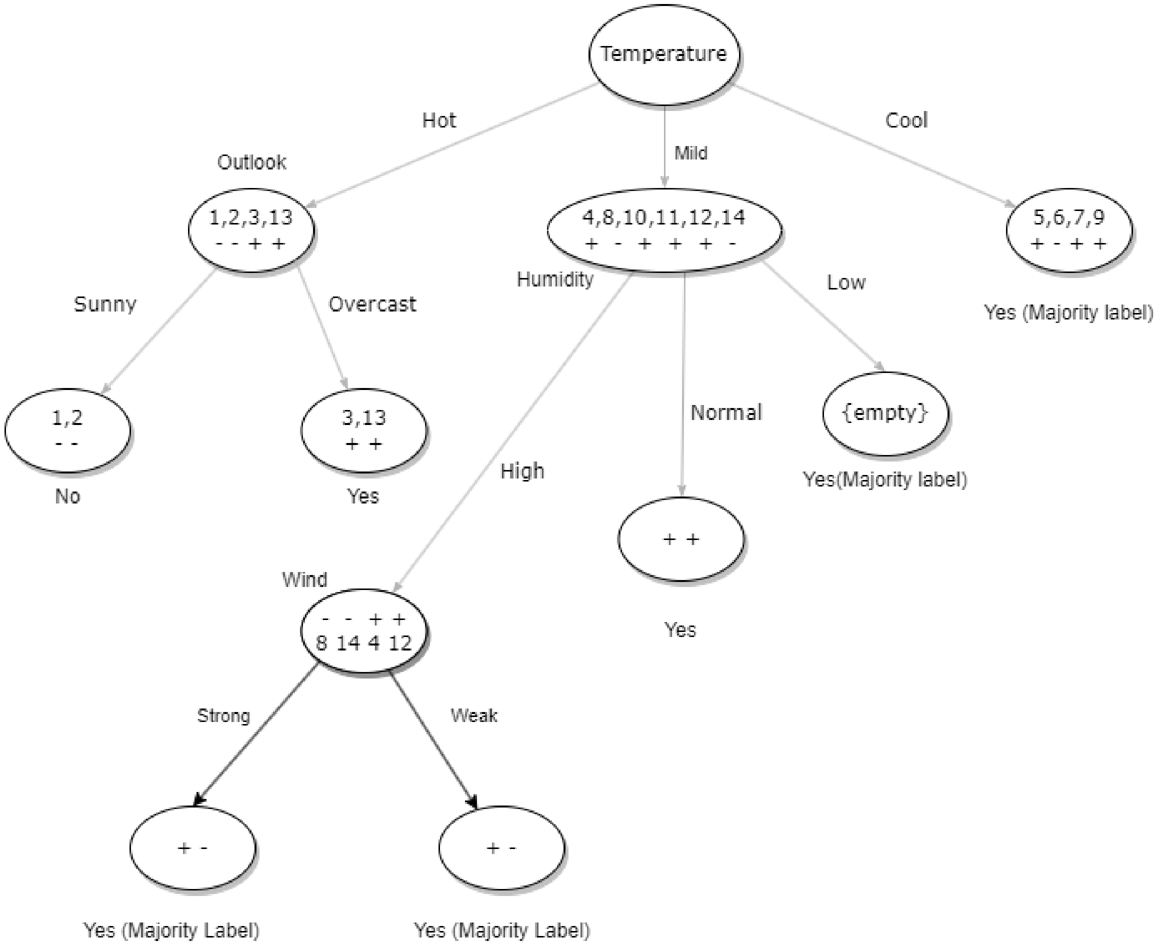


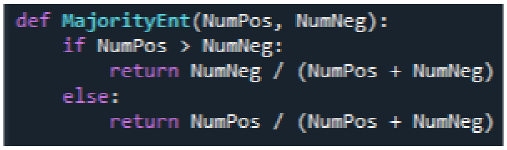
```
def EntropyCalc(NumPositives, NumNegatives, gainType):
    total = NumPositives + NumNegatives
    if gainType == 0 or gainType == "Info":
        posFrac = NumPositives / total
        negFrac = NumNegatives / total
        if NumPositives == 0 or NumNegatives == 0:
            return 0
        return (-1 * posFrac * math.log(posFrac, 2)) - (negFrac * math.log(negFrac, 2))
    elif gainType == 1 or gainType == "ME":
        MajorityEnt(NumPositives, NumNegatives)
    else:
        GiniEnt(NumPositives, NumNegatives)
```

1b. Boolean

$\neg X_2 \wedge X_4$

2a Majority Error





```
def MajorityEnt(NumPos, NumNeg):
    if NumPos > NumNeg:
        return NumNeg / (NumPos + NumNeg)
    else:
        return NumPos / (NumPos + NumNeg)
```

First Split

$$ME(S) = 5/14$$

$$Outlook(S) = \frac{5}{14} - (\frac{4}{14} * \frac{1}{2}) - 0 - (\frac{1}{3} * \frac{6}{14}) = .071$$

$$Temperature(S) = \frac{5}{14} - (\frac{1}{2} * \frac{4}{14}) - (\frac{1}{3} * \frac{6}{14}) - (\frac{1}{4} * \frac{4}{14}) = .142$$

$$Humidity(S) = \frac{5}{14} - (\frac{3}{7} * \frac{7}{14}) - (\frac{1}{7} * \frac{7}{14}) = .0714$$

$$Wind(S) = \frac{5}{14} - (\frac{1}{2} * \frac{6}{14}) - (\frac{1}{4} * \frac{8}{14}) = 0$$

Second Split

$$ME(S) = 1/2$$

$$Outlook(S) = \frac{2}{4} - 0 - 0 = 1/2$$

I stopped here because at best another attribute would be a tie and I'd just pick randomly

Third Split

$$+ - + + + -$$

$$ME(S) = \frac{1}{3}$$

$$Humidity(S) = \frac{1}{3} - (\frac{1}{2} * \frac{4}{6} - 0) = 0$$

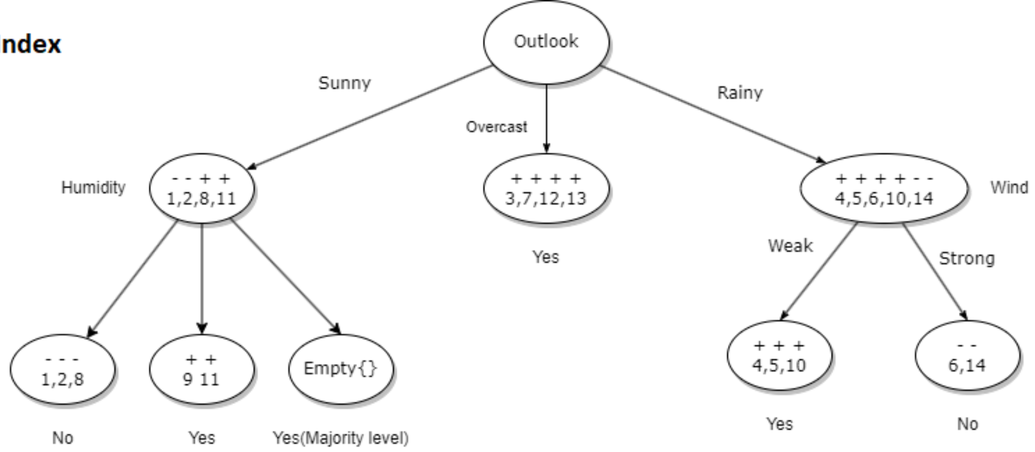
$$Wind(S) = \frac{1}{3} - (\frac{1}{3} * \frac{1}{2}) - (\frac{1}{2} * \frac{1}{3}) = 0$$

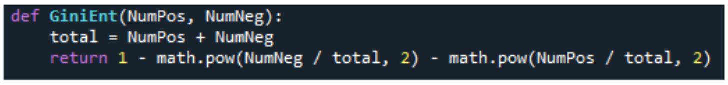
Tie so I chose the first attribute, Humidity

Last Attribute

Wind is the last attribute so I didn't do any computation
and assigned everything that was a tie or empty with the overall set majority label of positive(yes)

2bc Gini Index





```
def GiniEnt(NumPos, NumNeg):
    total = NumPos + NumNeg
    return 1 - math.pow(NumNeg / total, 2) - math.pow(NumPos / total, 2)
```

First Split

$$GiniEnt(S) = .459$$

$$Sunny --+++$$

$$Overcast ++++$$

$$Rainy +++--$$

$$Outlook(Sunny) = .48$$

$$Outlook(Overcast) = 0$$

$$Outlook(Rainy) = .444$$

$$Outlook(S) = .459 - (.48 * \frac{5}{14}) - 0 * \frac{4}{14} - \frac{5}{14} * .48 = .116$$

$$Hot --++$$

$$Mild --++++$$

$$Cool -+++$$

$$Temperature(H) = .5$$

$$Temperature(M) = .444$$

$$Temperature(C) = .375$$

$$.459 - (\frac{1}{2} * \frac{4}{14}) - (.444 * \frac{6}{14}) - (.375 * \frac{4}{14}) = .019$$

$$High ----+++$$

$$Normal -++++++$$

$$Low$$

$$Humidity(H) = .490$$

$$Humidity(N) = .245$$

$$Humidity(L) = 0$$

$$Humidity(S) = .459 -. (49 * \frac{7}{14}) -. (245 * \frac{7}{14}) = .0915$$

First Split Continued

$$Strong ---+++$$

$$Weak --++++++$$

$$Wind(S) = .5$$

$$Wind(W) = .375$$

$$.459 - (.5 * \frac{6}{14}) - (.375 * \frac{8}{14}) = .0304$$

Second Split

New Set

$$++--$$

$$Hot --$$

$$Mild -+$$

$$Cool +$$

$$Temp(H) = 0$$

$$Temp(M) = .5$$

$$Temp(C) = 0$$

$$.5 - (.5 * \frac{2}{5}) = .28$$

$$High ---$$

$$Normal ++$$

$$Humidity(H) = 0$$

$$Humidity(N) = 0$$

$$Humidity(S) = .5$$

This is the best it can get only hope is a tie and randomly choose so I'll stop here

Last Split

New Set

$$+++--$$

$$Wind(Str) = 0$$

$$Wind(W) = 0$$

$$Wind(S) = .48$$

This is a perfect split and all samples have the same label so I stopped here

2c. The initial split is different for IG/GI vs ME. This makes IG/GI identical while ME differs quite a bit in the entire structure.

3a.Note: Still using the code posted previously for calculations

Outlook
Positives

Rain
Rain
Overcast
Sunny
Rain
Sunny
Overcast
Overcast
—--------------
Sunny
Sunny
Rain
Sunny
Rain

5 Sunny, 5 Rain, 4 Overcast
Choosing Sunny as majority feature
New Entropy .940

Outlook is the best feature

$$Outlook(S) = .940 - .918(\frac{6}{15}) = .572$$

Outlook(O) = 0

Outlook(S) = .918

Outlook(R) = .721
—---

$$Temperature(S) = .940 - (1 * (\frac{4}{15})0.863 * (\frac{7}{15}) - .811 * (\frac{4}{15}) = .0543$$

$Temperature(H) = 1$
$Temperature(M) = .863$
$Temperature(c) = .811$
—--

$$Humidity(S) = .940 - .985 * (\frac{7}{15}) - .544 * (\frac{8}{15}) = .19$$

$Humidity(H) = .985$

$Humidity(N) = .544$

$Humidity(L) = 0$

$$Wind(Set) = .940 - 1 * (\frac{6}{15}) - .764 * (\frac{9}{15}) = .0816$$

$Wind(Str) = 1$

$Wind(W) = .764$

—--

3b

Outlook is the best split

$$Outlook(S) : .940 - .971 * (\frac{5}{15}) - .722 * (\frac{5}{15}) = .376$$

$Outlook(Sun) := .971$

$Outlook(O) = 0$

$Outlook(R) = .722$

$Temp(S) = .0543$

And all other variables will be the same because the unknown only affects Outlook

—---

3c.

Positive Outlook values

3 4 5 7 9 10 11 12 13

O R R O S R S O O

$$Outlook(Set) := .940 - .995 * (\frac{5}{14}) - .6944 * (\frac{5}{14}) = .336$$

$$Outlook(O) + + + + (4 + \frac{4}{14})^{+}), 0^{-} = 0$$

$$Outlook(S) : - - - + + (2 + \frac{5}{14})^{+}, 2^{-} = .995$$

$$Outlook)R_{:} + + - + + (4 + \frac{5}{14}^{+}), 1^{-} = .6944$$

All the rest of the attribute splits should be the same because only outlook was affected by the missing data