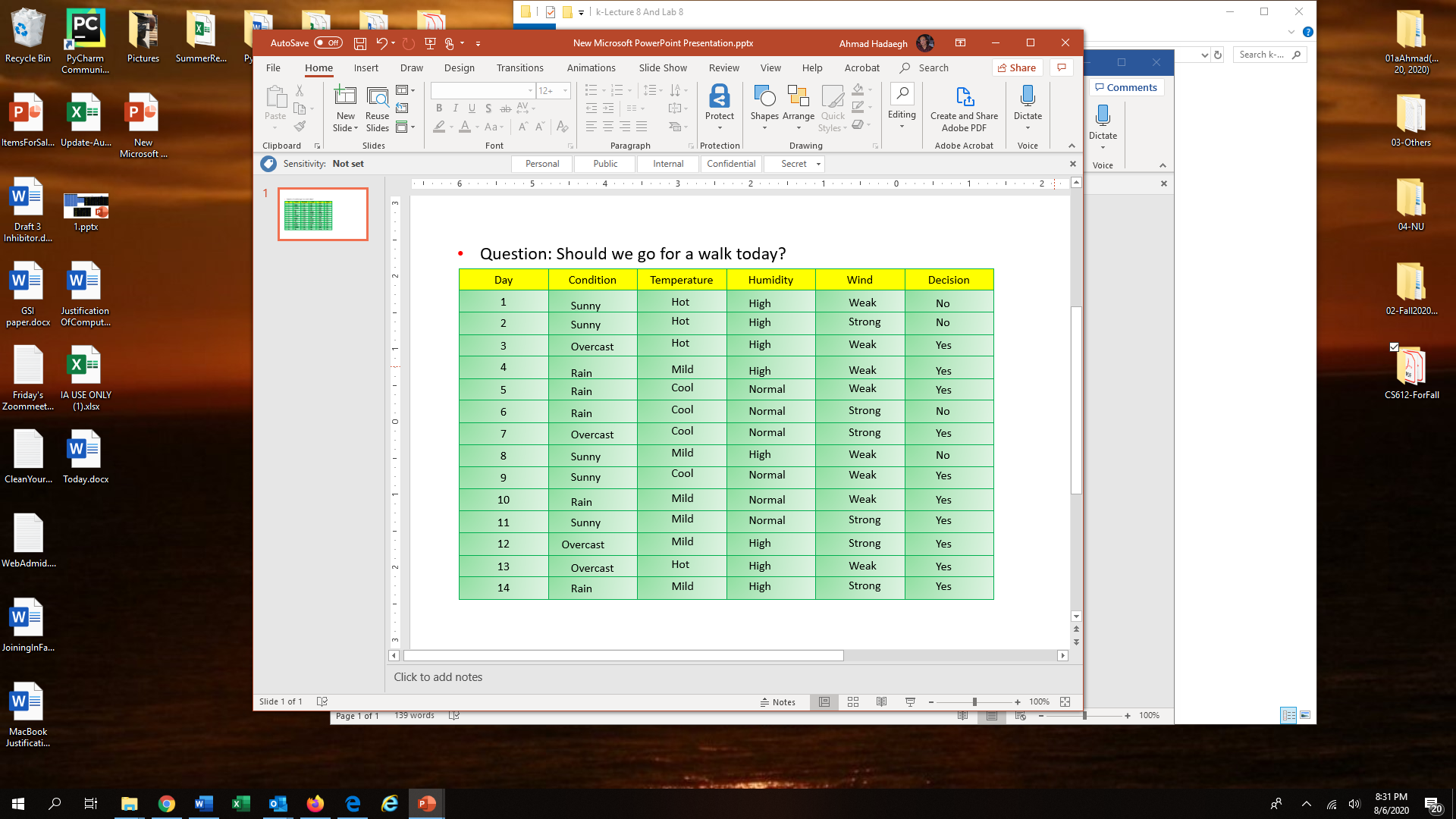
**CS612 – Lab 8**

1. Consider the following data set for a binary class problem.



|  |  |  |  |
| --- | --- | --- | --- |
| Decision | Yes | | No |
|  | | 10 | 4 |
| Entropy | −10/14 l(10/14) − 4/14 (4/14) = **0.8631** | | |

1. How would you split the tree based on the above table at level 1?

*Use entropy to make decision*. Split can be either binary or multiway. For example, if you need to split based on the condition, you need to find entropy based on:

* 1. {sonny}, {Overcast}, and {Rain}
  2. {Sunny and Overcast}, {Rain}
  3. {Sunny and Rain}, {Overcast}
  4. {Rain, Overcast}, {Sunny}

So, based on Condition, you have 4 choices, based on temperature you also have 4 choices, Based on Humidity, you have 2 choices, and based on Wind you have two choices.

Based on entropy calculated bellow it is the best to split multiway{sonny},{overcast}, and {rain} based on the condition at level first.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Condition  Have the lowest entropy so is the best choice to split on level 1 | | |
|  |  | Sunny | Overcast | Rain |
| Decision | Yes | 2 | 4 | 4 |
| No | 3 | 0 | 1 |
|  | Entropy | −2/5(2/5) – 3/5 (3/5) =0.9710 | −4/4(4/4) −0/4 (0/4) =0 | −4/5(4/5) – 1/5 (1/5) =0.7219 |
| Total Entropy | | 5/14\*0.9710+4/14\*0+5/14\*0.7219=0.6046 | | |
|  | Gain info | 0.8631-0.6046=0.2585 | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Condition | |
|  |  | Sunny and Overcast | Rain |
| Decision | Yes | 6 | 4 |
| No | 3 | 1 |
|  | Entropy | −6/9(6/9) – 3/9 (3/9) =0.9183 | −4/5(4/5) – 1/5 (1/5) = 0.7219 |
| Total Entropy | | 9/14\*0.9183+5/14\*0.7219=0.8481 | |
|  | Gain info | 0.8631-0.8481= 0.015 | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Condition | |
|  |  | Overcast | Sunny and Rain |
| Decision | Yes | 4 | 6 |
| No | 0 | 4 |
|  | Entropy | −4/4(4/4) – 0/4 (0/4) = 0 | −6/10(6/10) – 4/10 (4/10) =09710 |
| Total Entropy | | 0\*4/14+10/14\*0.9710= 0.6936 | |
|  | Gain info | 0.8631-0.6936= 0.1695 | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Condition | |
|  |  | Sunny | Overcast and Rain |
| Decision | Yes | 2 | 8 |
| No | 3 | 1 |
|  | Entropy | −2/5(2/5) – 3/5 (3/5) =0.9710 | −8/9(8/9) – 1/9 (1/9) = 0.5032 |
| Total Entropy | | 5/14 \*0.9710+9/14\* 0.5032=0.6703 | |
|  | Gain info | 0.8631-0.6703=0.1928 | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Temperature | | |
|  |  | Hot | Mild | Cool |
| Decision | Yes | 2 | 5 | 3 |
| No | 2 | 1 | 1 |
|  | Entropy | −2/4(2/4) – 2/4 (2/4) = 1 | −5/6(5/6) −1/6 (1/6) =0.65 | −3/4(3/4) – 1/4 (1/4) =0.8113 |
| Total Entropy | | 4/14\*1+6/14\*0.65+4/14\*0.8113=0.7961 | | |
|  | Gain info | 0.8631-0.7961=0.067 | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Temperature | |
|  |  | Hot and Mild | Cool |
| Decision | Yes | 7 | 3 |
| No | 3 | 1 |
|  | Entropy | −7/10(7/10) – 3/10 (3/10) =0.8813 | −3/4(3/4) – 1/4 (1/4) = 0.8113 |
| Total Entropy | | 10/14\*0.8813+4/14\*0.8113=0.8613 | |
|  | Gain info | 0.8631-0.8613 = 0.0018 | |

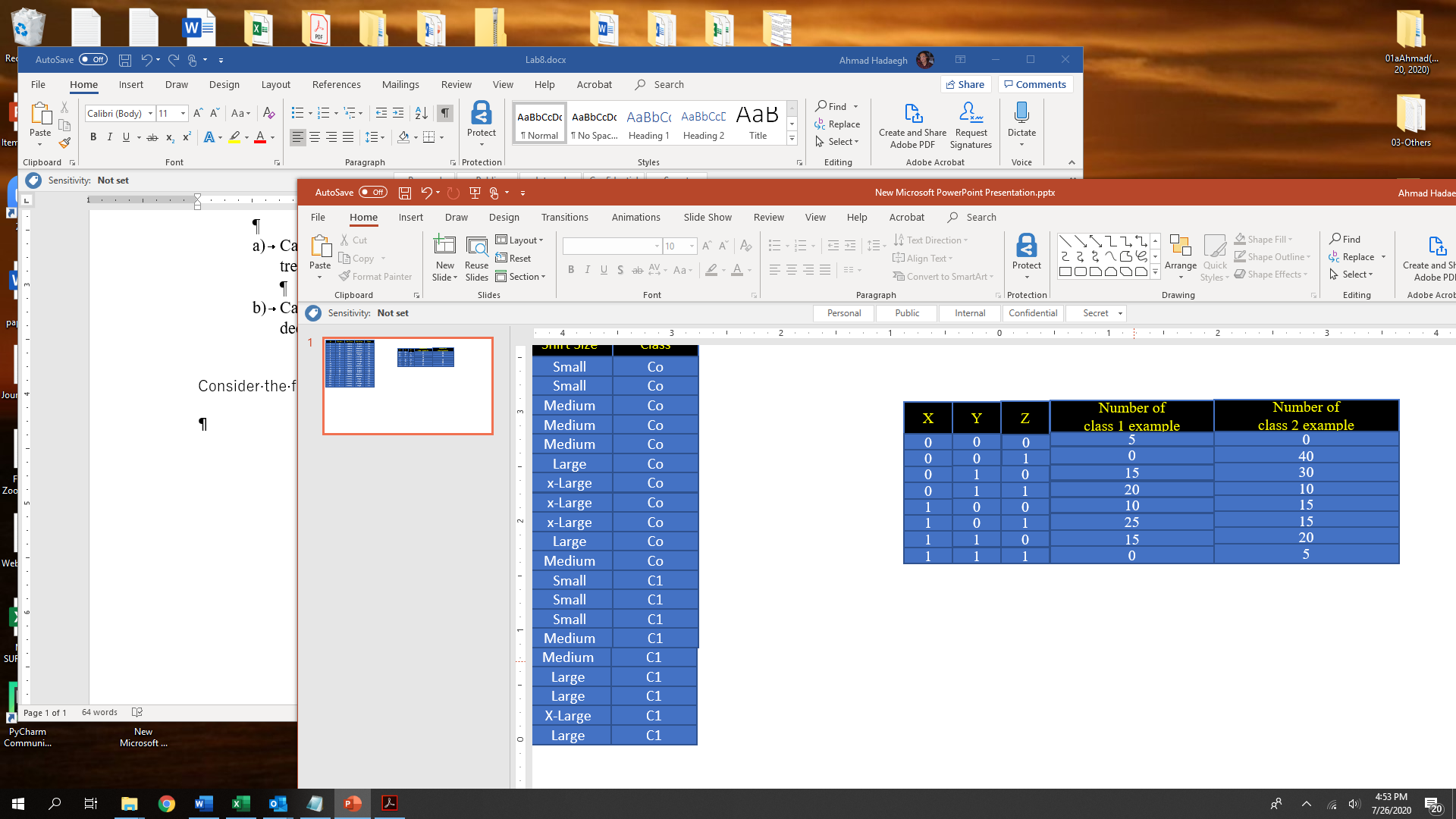
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Temperature | |
|  |  | Hot | Mild and Cool |
| Decision | Yes | 2 | 8 |
| No | 2 | 2 |
|  | Entropy | −2/4(2/4) – 2/4 (2/4) = 1 | −8/10(8/10) – 2/10 (2/10) =0.7219 |
| Total Entropy | | 1\*4/14+10/14\*0.7219= 0.8013 | |
|  | Gain info | 0.8631-0.8013= 0.0618 | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Temperature | |
|  |  | Mild | Hot and Cool |
| Decision | Yes | 5 | 5 |
| No | 1 | 3 |
|  | Entropy | −5/6(5/6) – 1/5 (1/5) =0.65 | −5/8(5/8) – 3/8 (3/8) = 0.9544 |
| Total Entropy | | 6/14 \*0.65+8/14\* 0.9544=0.8239 | |
|  | Gain info | 0.8631-0.8239=0.0392 | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Humidity | |
|  |  | High | Normal |
| Decision | Yes | 5 | 5 |
| No | 3 | 1 |
|  | Entropy | −5/8(5/8) – 3/8 (3/8) =0.9544 | −5/6(5/6) – 1/6 (1/6) = 0.65 |
| Total Entropy | | 8/14 \*0.9544+6/14\* 0.65=0.8239 | |
|  | Gain info | 0.8631-0.6703=0.0392 | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Wind | |
|  |  | Weak | Strong |
| Decision | Yes | 6 | 4 |
| No | 2 | 2 |
|  | Entropy | −6/8(6/8) – 2/8 (2/8) =0.8113 | −4/6(4/6) – 2/6 (2/6) = 0.9183 |
| Total Entropy | | 8/14 \*0.8113+6/14\* 0.9183=0.8571 | |
|  | Gain info | 0.8631-0.8571=0.006 | |

1. Consider the following set of training examples.



1. How would you split the tree at level 1? Based on X or Y or Z?

Splitting Attribute at Level 1. To determine the test condition at the root node, we need to compute the error rates ,Gini, and Entropy for attributes X, Y , and Z.

For attribute X, the corresponding counts are:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | C1 | C2 | error | Gini | Entropy |
| 0 | 40 | 80 | 1-80/120= 0.3333 | 1-(40/120)2-(80/120)2=0.4444 | 0.9183 |
| 1 | 50 | 55 | 1-55/105= 0.4762 | 1-(50/105)2-(55/105)2=0.4989 | 0.9984 |
| Total gini | 120/225\*0.4444+105/225\*0.4989=0.4699 | | | | | |
| Total error: | (50 + 40)/225 = 0.4 | | | | | |

Total entropy : 0.9557

For attribute Y, the corresponding counts are:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Y | C1 | C2 | Error | Gini | entropy |
| 0 | 40 | 70 | 1-70/110= 0.3636 | 0.4628 | 0.9457 |
| 1 | 50 | 65 | 1-65/115= 0.4348 | 0.4915 | 0.9877 |

Total gini: 110/225\*0.4628+115/225\*0.4915=0.4775

Total error rate using attribute Y is (40 + 50)/225 =0.4

Total Entropy: 0.9672

For attribute Z, the corresponding counts are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | C1 | C2 | Gini | Entropy |
| 0 | 5+15+10+15=45 | 0+30+15+20 = 65 | 0.4835 | 0.9760 |
| 1 | 45 | 40+10+15+5=70 | 0.4764 | 0.9656 |

Total gini index:110/225\*0.4835+115/225\*0.4764= 0.4799

Therefore, the error rate using attribute Y is (45 + 45)/225 = 0.4

Total entropy : 0.9707

**All of them have same error rate but since X gives the lowest Entropy and gini index, it is chosen as the splitting attribute at level 1.**

1. How would you split the tree at level 2?

After splitting on attribute X, the subsequent test condition may involve either attribute Z or Y. This depends on the training examples distributed to the X= 0 and X= 1 child nodes

FOR X=0,

the corresponding counts for attributes Z and Y are the same, as shown in the table below. Error rate in both cases(x and Y) are 0.3333 but **Gini index and Entropy of Z in smaller than Y.**

**So for x=0 , Z is the best split.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Z | C1 | C2 | Gini | Entropy |
| 0 | 20 | 30 | 0.48 | 0.9710 |
| 1 | 20 | 50 | 0.4081 | 0.8631 |

Error rate: (20+20)/120= 0.3333

Total gini:0.4380

Total Entropy:0.9080

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Y | C1 | C2 | Gini | Entropy |
| 0 | 5 | 40 | 0.1975 | 0.5033 |
| 1 | 35 | 70 | 0.4444 | 0.9182 |

Total error: (5+35)/120=0.3333

Total Gini:0.4629

Total Entropy:0.9922

For X=1,

the corresponding Error rate for attrebiutes X and Y are the same in the tables below But **gini index and entropy Y is smaller than Z .So for X=1 Y is the best split.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Z | C1 | C2 | gini | Entropy |
| 0 | 25 | 35 | 0.4861 | 0.9799 |
| 1 | 25 | 20 | 0.4938 | 0.9911 |

Total error: (25+20)/120=0.375

Total Gini:0.4281

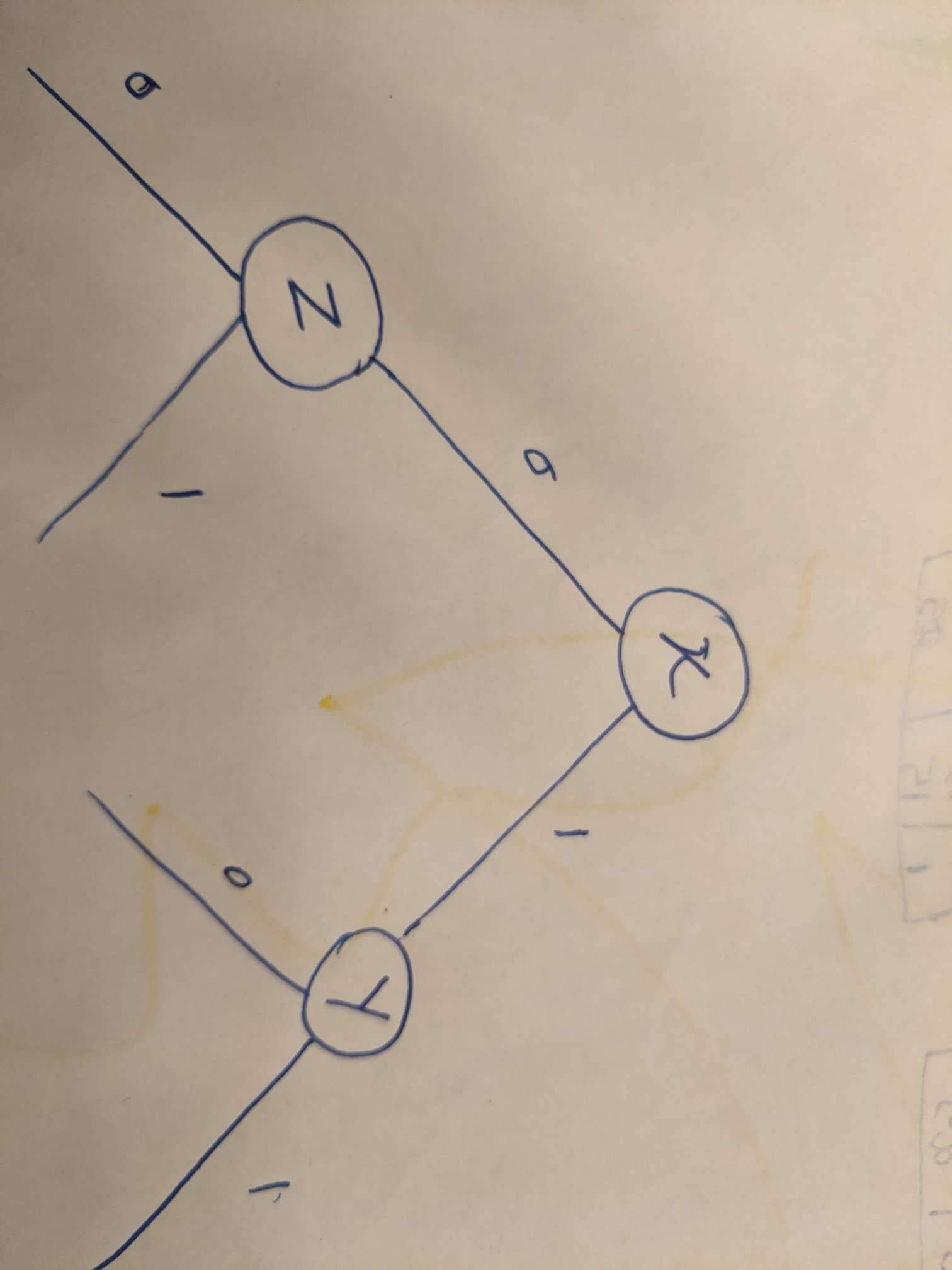
Total Entropy:0.8616

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Y | C1 | C2 | gini | Entropy |
| 0 | 35 | 30 | 0.4970 | 0.9957 |
| 1 | 15 | 25 | 0.4688 | 0.9544 |

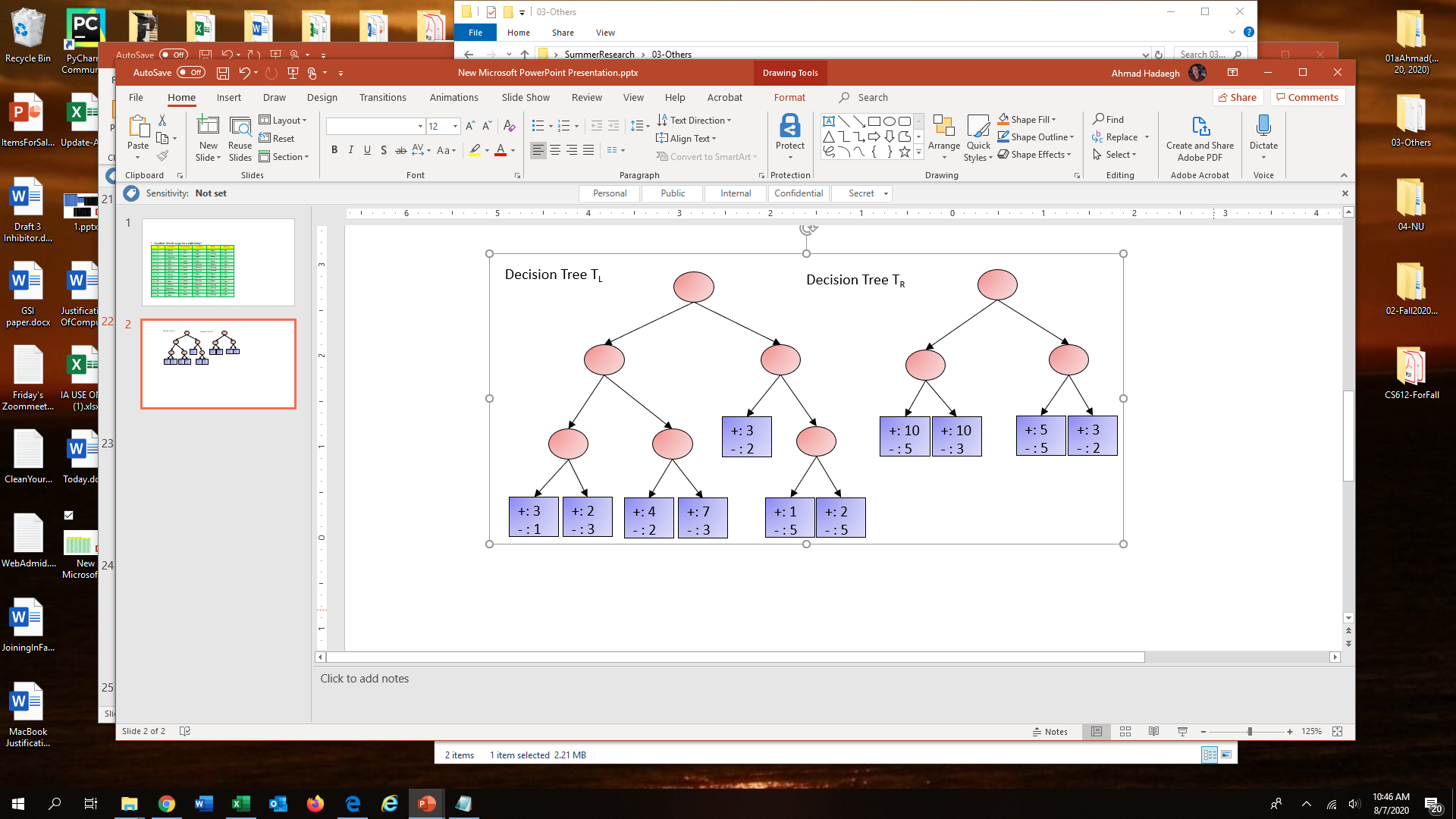
Total error: (30+15)/120= 0.375

Total Gini:0.4255

Total Entropy:0.8575



1. Consider the following two trees. Find the optimistic and pessimistic algorithms of the trees. Assume the penalty of each leaf is 0.75 Which one should be selected for the classification?.



**Optimistic Error :**

**Based on optimistic error left tree is consider better for classification**

**E(TL)= (1+2+2+3+1+2+2)/(4+5+6+10+5+6+7)=0.3023 E(TR) = (5+3+5+2)/(15+13+10+5)=0.3488**

**Pesimistic Error:**

**Eg(TL)=0.3032 + 7 \*0.75=5.5532 Eg(TR)=0.3488 + 4\*0.75= 3.3488**

**Based on Pesimistic Error, Right tree is consider better for classification**

**What to submit:**

* **Place Questions 1 and Question 2 and Question 3 in one file called it Lab8.docx**
* **Name the folder based on your name and your partner’s name and your partner’s name (ex: NancyJack-Lab8)**
* **Zip it and submit it through the cougar courses.**