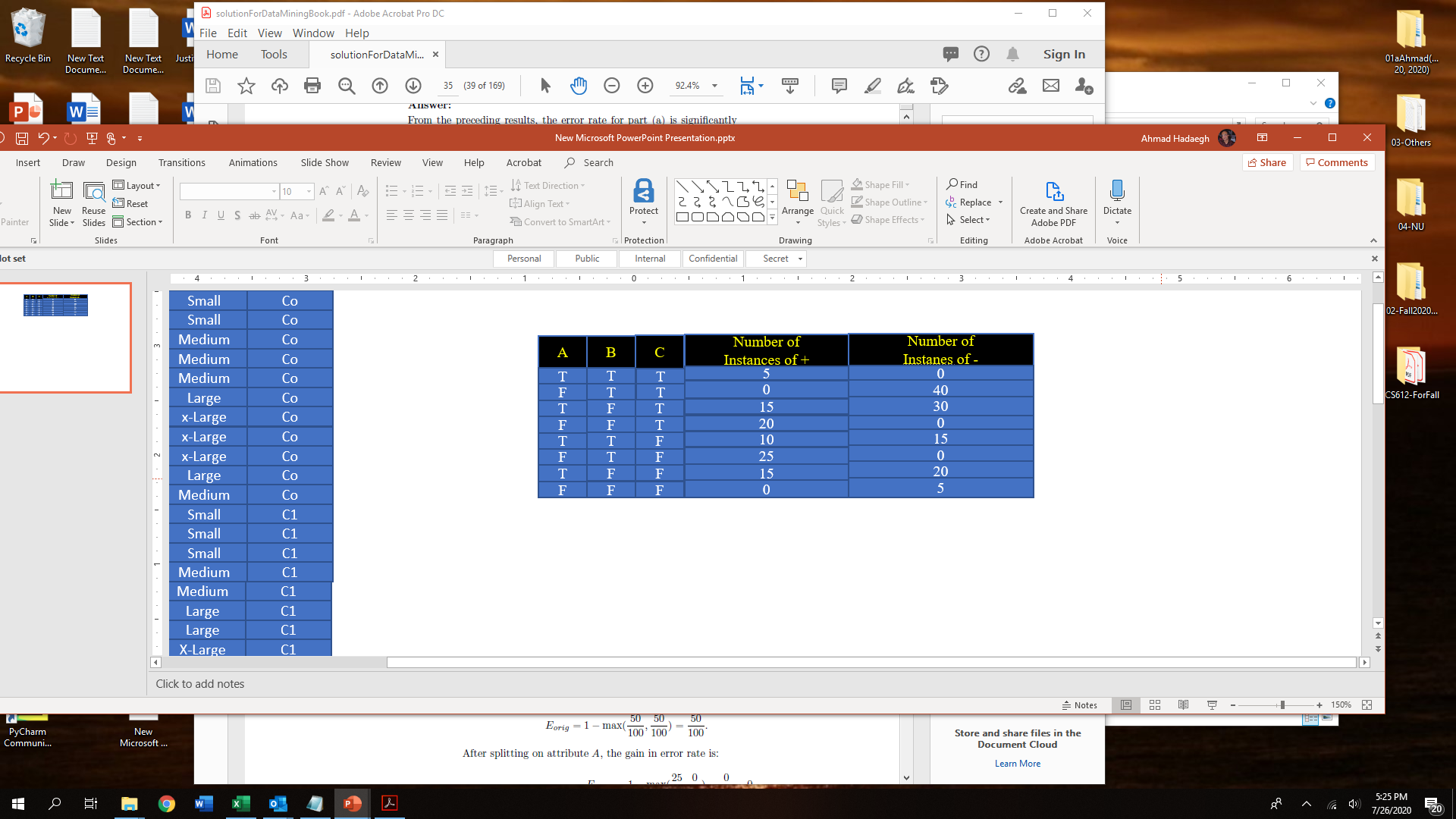
**CS612 – Lab 9**

**Dear Dr. Hadaegh this file is our homework with previous number for question 1, we saw the number of question 1 has been change , so we do it again in the other file in our homework folder. to make sure , we submit both files,old version of lab9 and new version of lab 9. Thank you**

1. The following table summarizes a data set with three attributes *A*, *B*, *C* and two class labels +, *−*. Build a two-level decision tree.



1. According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

Error rate before splitting at first level= 1- max[90/200, 110/200] = 1-110/200= 0.45

|  |  |  |
| --- | --- | --- |
|  | A= T | A= F |
| + | 5+15+10+15=45 | 0+20+25+0=45 |
| - | 0+30+15+20=65 | 40+0+0+5=45 |
| Error rate | 1-max[45/110,65/110]= 45/110 | 1-max[45/90,45/90]=45/90=0.4 |
| Total error rate when split on A | 110/200\*45/110+90/200\*45/90= 0.45 | |
| Error gain when split on A | 0.45-0.45=0 | |

|  |  |  |
| --- | --- | --- |
|  | B= T | B= F |
| + | 5+0+10+25=40 | 15+20+15+0=50 |
| - | 0+40+15+0=55 | 30+0+20+5=55 |
| Error rate | 1-max[40/95,55/95]= 40/95 | 1-max[50/105,55/105]=50/105 |
| Total error rate when split on B | 95/200\*40/95+105/200\*50/105=0.45 | |
| Error gain when split on B | 0.45-0.45= 0 | |

|  |  |  |
| --- | --- | --- |
|  | C= T | C= F |
| + | **5+0+15+20=40** | **10+25+15+0=50** |
| - | **0+40+30+0 =70** | **15+0+20+5=40** |
| Error rate | 1-max[40/110,70/110]= 40/110 | 1-max[40/90,50/90]=40/90 |
| Total error rate when split on C | 110/200\*40/110+90/200\*40/90=0.4 | |
| Error gain when split on C | 0.45- 0.4= 0.05 | |

At first level C attribute is best to split because C has minimum error rate and maximum Error gain

1. Repeat for the two children of the root node.

**When C= T the error gain is:**

|  |  |  |  |
| --- | --- | --- | --- |
| **A** | **B** | **+** | **-** |
| T | T | 5 | 0 |
| F | T | 0 | 40 |
| T | F | 15 | 30 |
| F | F | 20 | 0 |

**Total Error Rate before splitting at second level is= 1- max[40/110,70/110]= 40/110**

|  |  |  |
| --- | --- | --- |
|  | A= T | A= F |
| + | 5+15= 20 | 20+0=20 |
| - | 30+0 = 30 | 0+40=40 |
| Error rate | 1-max[25/50,40/50]=20/50 | 1-max[20/60,5/60]=20/60 |
| Total error rate when split on A | 50/110\*20/50+60/110\*20/60=40/110 | |
| **Error gain when split on A at second level and C= T** | 40/110-40/110=0 | |
|  |  | |

|  |  |  |
| --- | --- | --- |
|  | B= T | B= F |
| + | 5+0=5 | 15+20=35 |
| - | 40+0=40 | 0+30= 30 |
| Error rate | 1-max[5/45,40/45]= 5/45 | 1-max[35/65,30/65]=30/65 |
| Total error rate when split on B | 45/110\*5/45+65/110\*30/65=35/110 | |
| **Error gain when split on B at second level and c=T** | 40/110-35/110= 10/110= 1/11 | |

On second level when C=T the best split is B because of its error rate is lower than A and its Error gain in more than A’s error gain

**When C=F:**

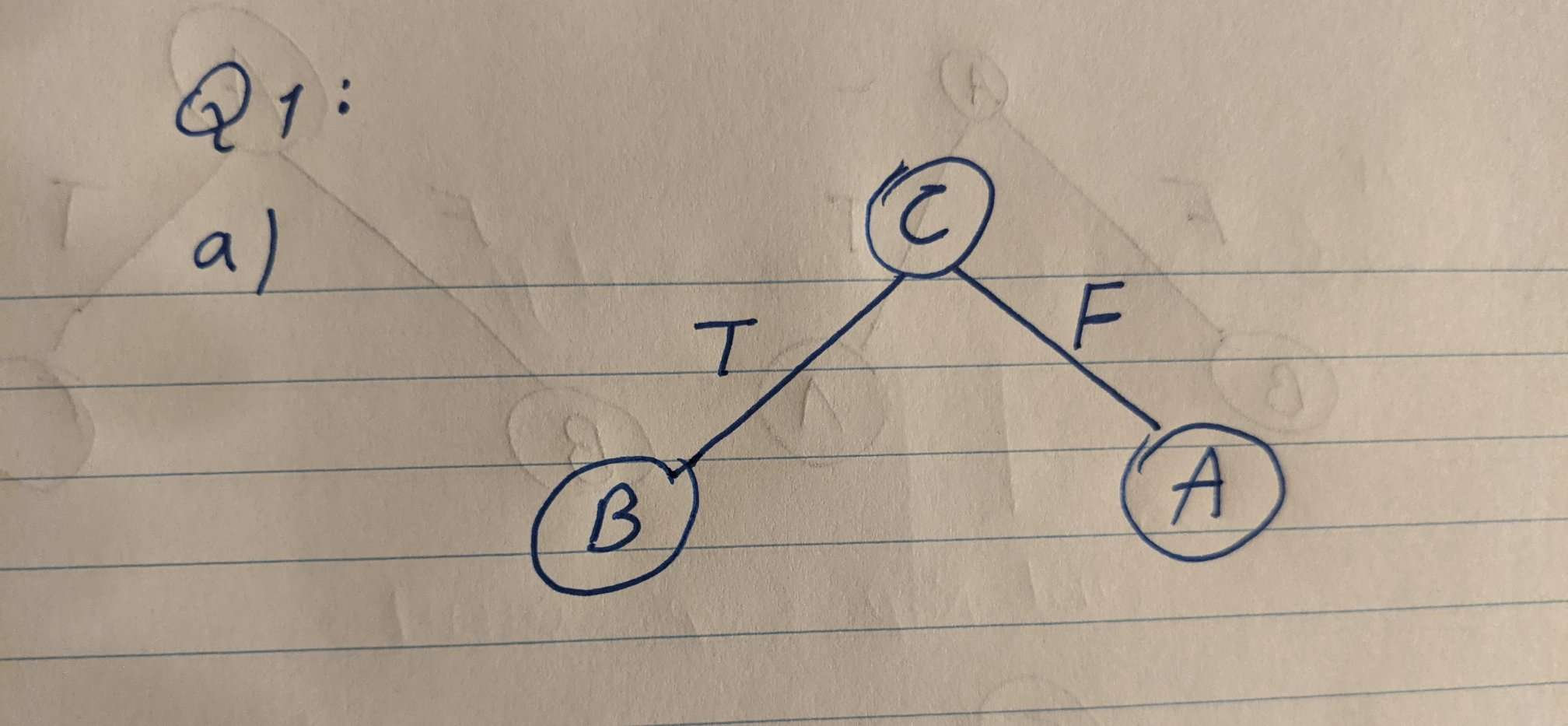
**Total error before split on second level: 1- max[50/90,40/90]= 40/90**

|  |  |  |  |
| --- | --- | --- | --- |
| A | B | + | - |
| T | T | 10 | 15 |
| F | T | 25 | 0 |
| T | F | 15 | 20 |
| F | F | 0 | 5 |

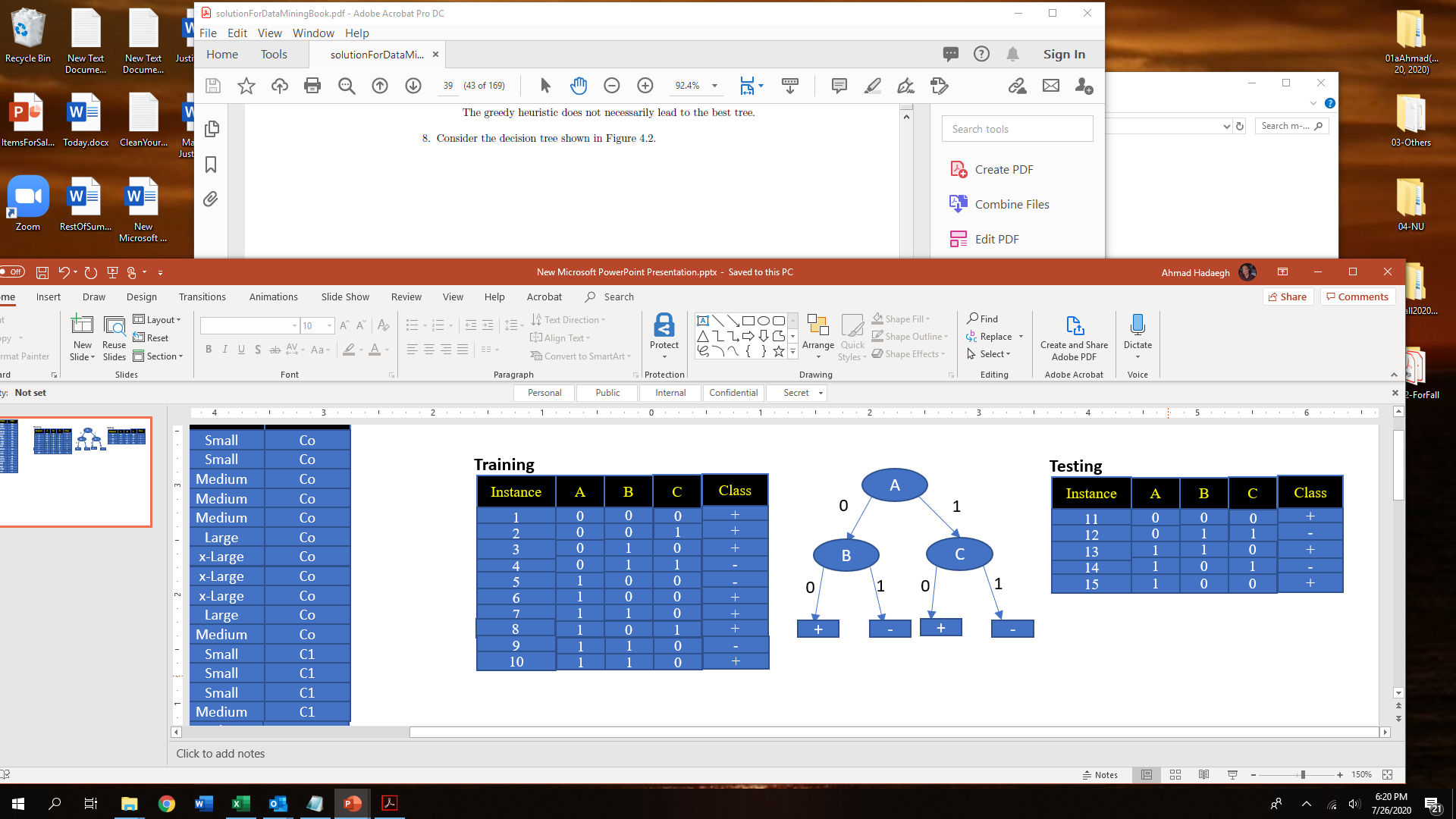
|  |  |  |
| --- | --- | --- |
|  | A= T | A= F |
| + | 10+15= 25 | 25+0=25 |
| - | 15+20 = 35 | 0+5=5 |
| Error rate | 1-max[25/60,35/60]=25/60 | 1-max[25/30,5/30]=5/30 |
| Total error rate when split on A | 60/90\*25/60+30/90\*5/30=30/90 | |
| **Error gain when split on A at second level and C= T** | 40/90-30/90=10/90= 1/9 | |
| **The Overal entropy for A is:** | 60/90[ − (25/60) (25/60) − (35/60) (35/60)] +30 /90 [ − (25/30) (25/30) − (5/30) (5/30)]=0.410 +0.3605= **0.8699** | |

|  |  |  |
| --- | --- | --- |
|  | B= T | B= F |
| + | 10+25=35 | 15+0=15 |
| - | 15+0=15 | 20+5= 25 |
| Error rate | 1-max[35/50,15/50]= 15/50 | 1-max[15/40,25/40]=15/40 |
| Total error rate when split on B | 50/90\*15/50+40/90\*15/40=30/90 | |
| **Error gain when split on B at second level and c=T** | 40/90-30/90=10/90= 1/9 | |
| **The Overal entropy for B is:** | 50/90[ − (35/50) (35/50) − (15/50) (15/50)] +40 /90 [ − (15/40) (15/40) − (25/40) (25/40)]= **0.9138** | |

On second level when C=F the best split is A because with same error rate ,A has lower entropy than B



1. Consider the decision tree shown below:



1. Compute the generalization error rate of the tree using the optimistic approach.

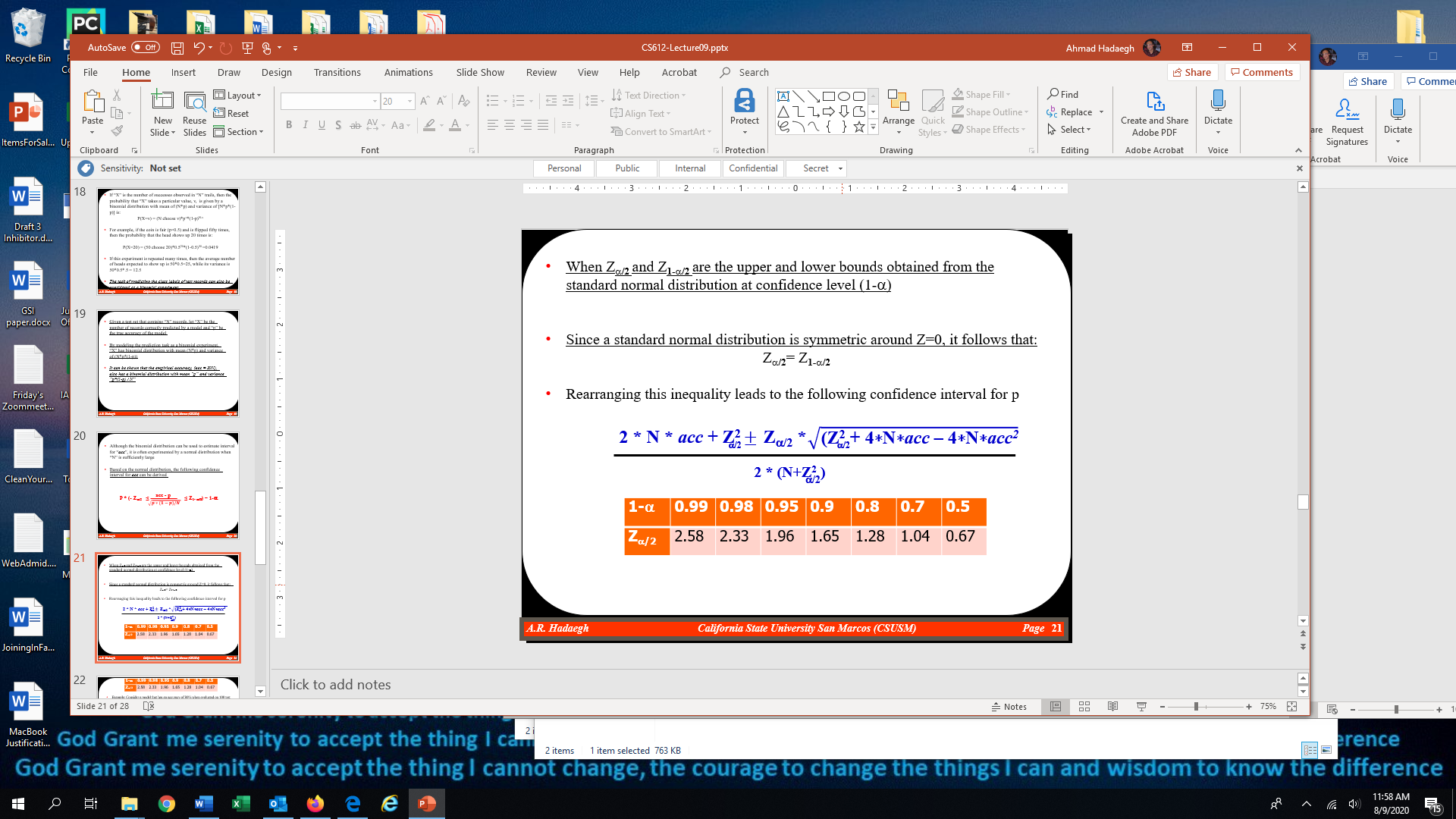
**The decision tree returns '+' ,if (not A and not B) or (A and not C)**statement well.

According to optmistic approach, using the above ,we can observe that instances 3,5,8,9 are misclassified ,Hence Generalization error rate = Training Error rate= 4/10 = 0.4

1. Compute the generalization error rate of the tree using the pessimistic approach. (For simplicity, use the strategy of adding a factor of 0.5 to each leaf node.)

 According to the pessimistic approach, the generalization error rate is (4+ (4 × 0.5))/10 = 0.6

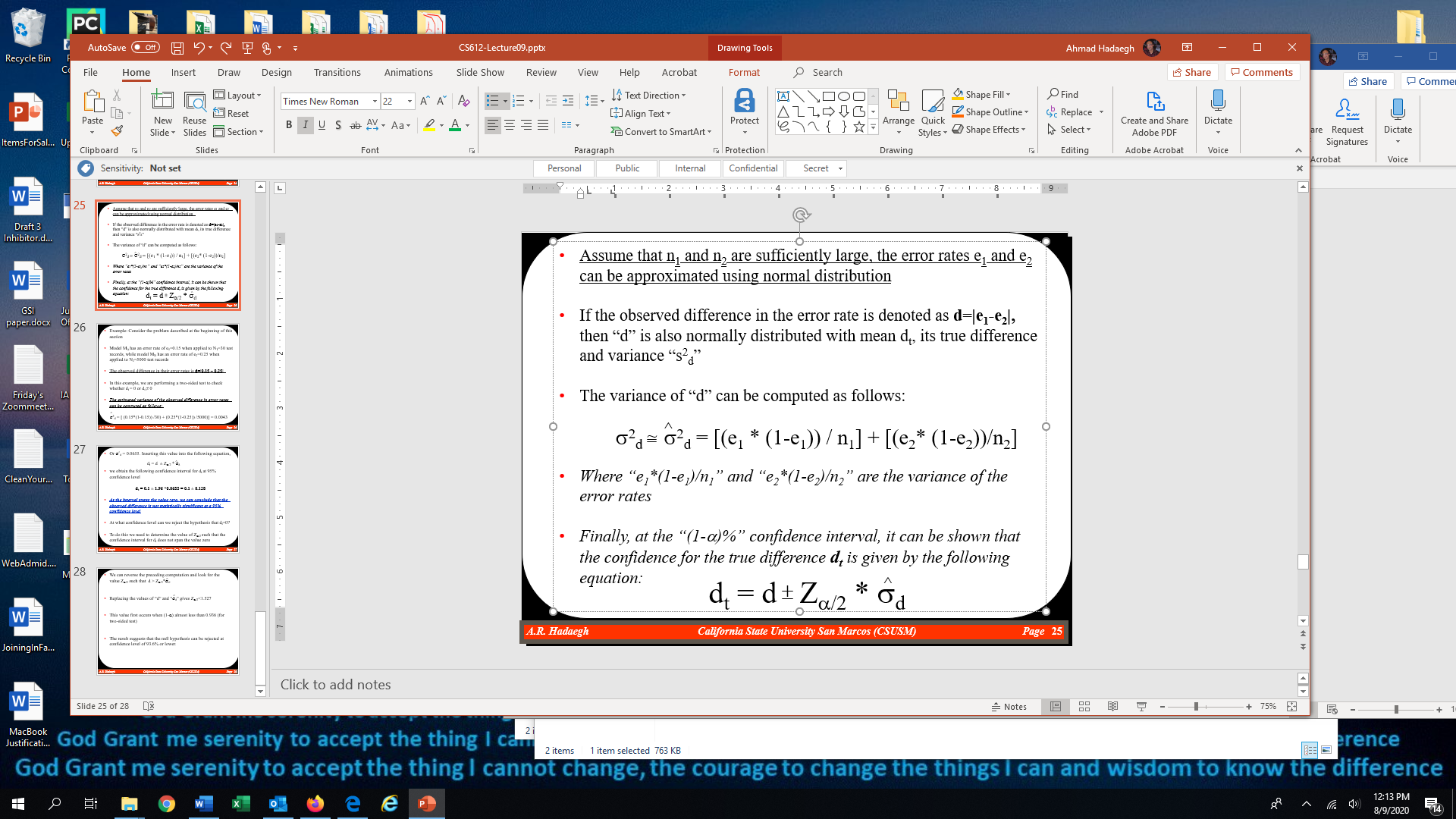
1. Suppose we have the following equation is to find confidence internal.



Where N=1000 is the number of records, acc= 75% is the accuracy, standard normal distribution at confidence level (1-a) is 0.99. Calculate the interval based on the given information.

**(2\*1000\* 0.75 + 2.582 2.58 \* sqrt (2.582 + 4\*1000\*0.75 – 4\*1000\*0.752)/2\*(1000+2.582) = interval = %0.7131 , %0.7836**

1. Suppose model MA has an error rate of e1=0.20 when applied to N1=3000 test records, while model MB has an error rate of e2=0.25 when applied to N2=6000 test records. Use the following formula to find the estimated variance of the observed difference in error.



Find If the observed difference in the error “d”, the mean dt, its true difference and variance “s2d” with confidence level of 98%. Does the interval span the value zero?

**The observed difference in their error rates is d=|0.20 – 0.25|= 0.05**

**s2d = [ (0.20\*(1-0.20)) /3000) + (0.25\*(1-0.25)) /6000)] = 0.00008458**

**s^d = 0.009197 . Inserting this value into the following equation,**

**we obtain the following confidence interval for dt at 95% confidence level**

**dt = d ± Za/2 \* sd**

**dt = 0.05 ± 2.33 \*0.009197 = 0.05 ± 0.02143= 0.02857 , 0.07143**

***As the interval does not span the value zero, we can conclude that the observed difference is statistically significant at a 95% confidence level.***