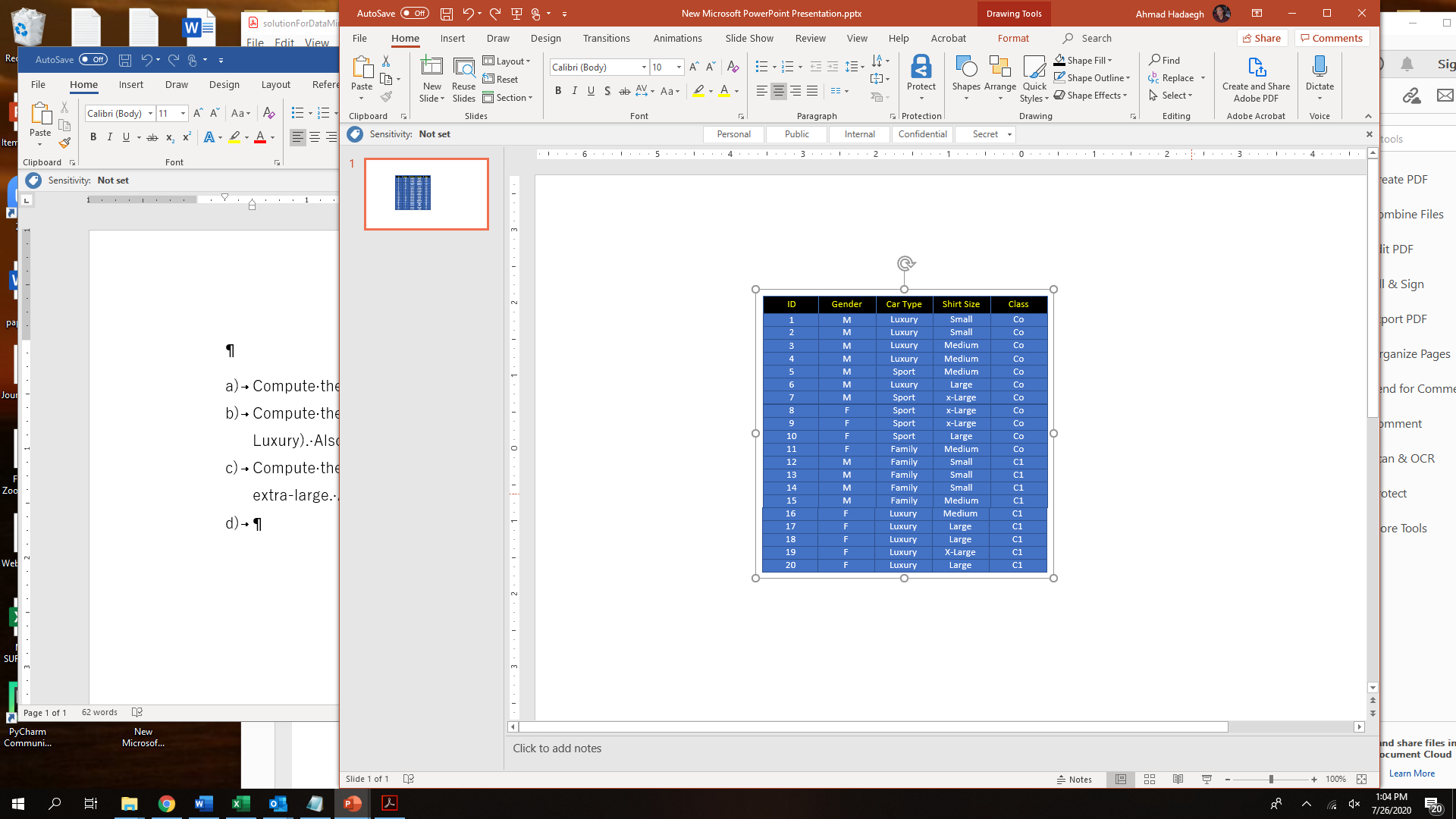
**CS612 – Lab 7**

1. Do the following Question based on the following information



* + 1. Entropy (t) = - Si=0,“c-1” [p(i|t) log2 p(i|t)]
    2. Gini (t) = 1 - Si=0,“c-1” [p(i|t)]2
    3. Classification error (t) = 1 – max [p(i|t)]

where “c” is the number of classes

|  |  |  |
| --- | --- | --- |
|  | **Gender** | |
|  | **Female** | **male** |
| **C0** | 4 | 7 |
| **C1** | 5 | 4 |
| a)Gini | 1-(4/9)2-(5/9)2 = **0.4938** | 1-(7/11)2-(4/11)2 =**0.4628** |
| a)Overall Gini | (9/20)0.493+(11/20)0.463=**0.4767** | |
| b)Entropy | -(4/9) (4/9)-(5/9) (5/9)=  -4/9\*-1.170-5/9\*-0.8480 = **0.9910** | -(7/11) (7/11)-(4/11) (4/11)=  -7/11\*-0.652 -4/11 \* -1.460= **0.9456** |
| Total Entropy | 9/20\*[-(4/9) (4/9)-(5/9) (5/9)]+11/20[-(7/11) (7/11)-(4/11) (4/11)]= 9/20\*0.9910+11/20\* 0.9456=**0.9660** | |
| c)Error | 1-max[4/9,5/9]=**0.444** | 1-max[7/11,4/11]=**0.363** |
| Total Error | 9/20[1-MAX[4/9,5/9]]+11/20[1-MAX[7/11,4/11]=9/20\*0.444+11/20\*0.363=**0.4** | |

1. **Ovelal gini for par a : 0.4767**
2. Computer the **Entropy for part a : 0.966**
3. Computer **the error for part a : 0.4**
4. Compute the Gini index for the Car Type attribute **using multiway 3 directions** (Family, Sport, Luxury). Also find the overall Gini index

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Car type | | | |
|  | Family | | Sport | luxury |
| C0 | 1 | | 5 | 5 |
| C1 | 4 | | 0 | 5 |
| 1. Gini | 1-(1/5)2-(4/5)2=0.32 | | 1-(5/5)2-0= 0 | 1-(5/10)2-(5/10)2=0.5 |
| **d)Overal Gini** | | **(5/20)\*0.32 + (5/20)\*0 + (10/20)\*0.5=0.33** | | |
| e)Entropy | -(1/5) (1/5)-(4/5) (4/5)=-1/5\*-2.322 – 4/5\*-0.322 = 0.7219 | | -(5/5) (5/5)-0=**0** | -(5/10) (5/10)-(5/10) (5/10)=  -5/10\*-1 – 5/10 \* -1 = 1 |
| **Total Antropy** | 5/20\*0.7219+5/20\*0+10/20\*1=**0.6804** | | | |
| f)Error | 1-max[(1/5),(4/5)]=0.2 | | 1-max[1,0]=0 | 1-max[0.5,0.5]=0.5 |
| **Total Error** | 0.2\*5/20+0\*5/20+0.5\*10/20=**0.3** | | | |

**Ovelal gini for part d :0.33**

1. Computer the **Entropy for part d : 0.680**
2. Computer **the error for part d: 0.3**
3. Compute the Gini index for the Shirt Size attribute using multiway: **small, medium. Larger, and extra-large**. Also find the overall Gini index

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Shirt Size** | | | |
|  | **small** | **Medium** | **Larger** | **Extra large** |
| **C0** | 2 | 4 | 2 | 3 |
| **C1** | 3 | 2 | 3 | 1 |
| **g)Gini** | 1-(2/5)2-(3/5)2=0.48 | 1-(4/6)2-(2/6)2=0.4444 | 1-(2/5)2-(3/5)2=0.48 | 1-(3/4)2-(1/4)2=0.375 |
| **g)Overal Gini** | (5/20)\*0.48+(6/20)\*0.44+(5/20)\*0.48+(4/20)\*0.375=**0.4483** | | | |
| **h)Entropy** | -(2/5) (2/5)-(3/5) (3/5)=  -2/5\* -1.322 – 3/5\*-0.737= **0.9710** | -(4/6) (4/6)-(2/6) (2/6)=  -4/6\*-0.585- 2/6 \* -1.585=  **0.9182** | -(2/5) (2/5)-(3/5) (3/5)=  **-2/5\*-1.322-3/5\*-0.737= 0.9710** | -(3/4) (3/4)-(1/4) (1/4)=  **-3/4\* -0.415 – ¼ \* -2= 0.8112** |
| **Total Antropy** | 5/20\*0.9710+6/20\*0.9182+5/20\*0.9710+4/20\*0.8112=**0.9232** | | | |
| **i)Error** | 1-Max[(2/5),(3/5)]=**0.4** | 1-max[(4/6),(2/6)]=**0.3333** | 1-max[(2/5),(3/5)]=**0.4** | 1-max[(3/4),(1/4)]=**.25** |
| **Total Error Rate** | 5/20\*0.4+6/20\*0.33+5/20\*0.4+4/20\*0.25=**0.3499** | | | |
|  |  | | | |

**Overall gini for part g: 0.4483**

1. **Computer the Entropy for part g : 09232**
2. **Computer the error for part g: 0.3499**

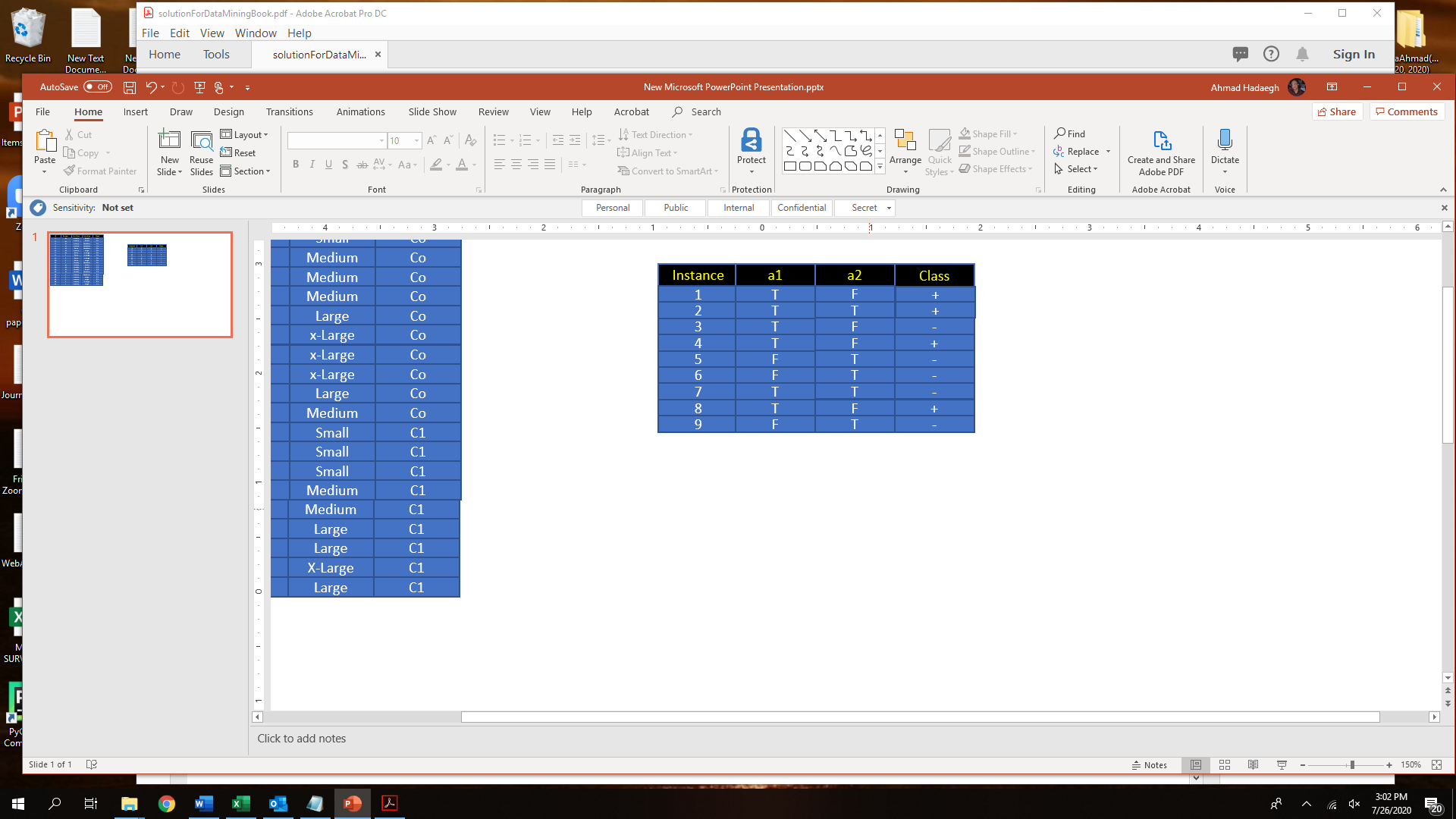
**Which of the above would you choose to do the split based on the result of your Gini?****Car type has has the lowest overall gini index so based on the result of our gini index we would choose car type to do the split**

.

**Which of the above would you choose to do the split based on the result of your Entropy? Car type has has the lowest overall entropy so based on the result of our entropy we would choose car type to do the split**

**Which of the above would you choose to do the split based on the result of your error? Car type has has the lowest overall error rate so based on the result of our error rate we would choose car type to do the split**

1. Consider the training examples shown in above table for a binary classification problem.



* 1. What is the entropy of this collection of training examples with respect to the positive class?

There are four positive examples and five negative examples. Thus,

P(+) = 4/9 and P(−)=5/9.

The entropy of the training examples : −4/9 l(4/9) − 5/9 (5/9) = **0.9911.**

* 1. What are the information gains of a1 and a2 relative to these training examples?

|  |  |  |
| --- | --- | --- |
| **A1** | **T** | **F** |
| **+** | **4** | **0** |
| **-** | **2** | **3** |
| **The Overal entropy for a1 is:** | 6 /9[ − (4/6) (4/6) − (2/6) (2/6)] + 3 /9 [ − (0/3) (0/3) − (3/3) (3/3)]= **0.6121** | |
| **The information gain for a1:** | 0.9911 – 0.6121 = **0.379** | |
| **Overal gini index for a1:** | 6/9[ 1 – (4/6)2 – (2/6)2] +3/9 [1 – (0/3)2 – (3/3)2] =**0.2962** | |
| **Overal Error for a1 is:** | 6 /9[ 1 −max[4/6,2/6]] + 3/9 [1 −max[0/3,3/3]] = **0.2222** | |

|  |  |  |
| --- | --- | --- |
| **A2** | **T** | **F** |
| **+** | **1** | **3** |
| **-** | **4** | **1** |
| **The Overal entropy for a2 is:** | 5 /9[ − (1/5) (1/5) − (4/5) (4/5)] +4 /9 [ − (3/4) (3/4) − (1/4) (1/4)]=0.410 +0.3605= **0.7616** | |
| **The information gain for a2:** | 0.9911 – 0.7616 = **0.2295** | |
| **Overal gini index for a2:** | 5/9[ 1 – (1/5)2 – (4/5)2] +4/9 [1 – (3/4)2 – (1/4)2] =0.178+0.167= **0.3444** | |
| **Overal Error for a2 is:** | 5 /9[ 1 −max[1/5,4/5]] + 4/9 [1 −max[3/4,1/4]] = **0.2222** | |

* 1. What is the best split (among a1, a2) according to the information gain?

**According to information gain among a1 and a2 , a1 has greater gained information so a1 is the best split**

* 1. What is the best split (between a1 and a2) according to the classification error rate?

**The classification error rate a1 and a2 are same value :0.2222. so** **according to the classification error rate no matter to select a1 or a2 to split**

* 1. What is the best split (between a1 and a2) according to the Gini index?

**For attribute a1, the gini index: 0.2962**

**For attribute a2, the gini index :0.34444.**

**Since the gini index for a1 is smaller, it produces the better split.**

**What to submit:**

* **Place Questions 1 and Question 2 in one file called it Lab7.docx**
* **Place the above two files in a folder called it based on your name and your partner’s name (ex: NancyJack-Lab7)**
* **Zip it and submit it through the cougar courses.**