

## Exam 1's Practice Problem

- Given the **training** data.

Class	Sex	Survived
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	1	1
3	0	1
2	1	1
3	0	1
2	1	0
2	1	0
3	1	0
2	1	0
3	0	0
3	1	0
2	0	0
1	1	0
2	1	0

Using Gini Index as the measure for impurity to:

- Grow the **maximum tree** with four leaves (a stopping rule!) on the data. Draw the (diagram of) tree.
- Find the misclassification rate on training data of the maximal tree
- Draw all the candidate **subtrees**
- Validate the maximal tree and the subtrees on the following data to select the **optimal tree**. Note: if the chance of Survived is  $1/2$ , predict **Survived**

Class	Sex	Survived
1	0	1
1	1	1
3	1	0
2	0	0

## Solution

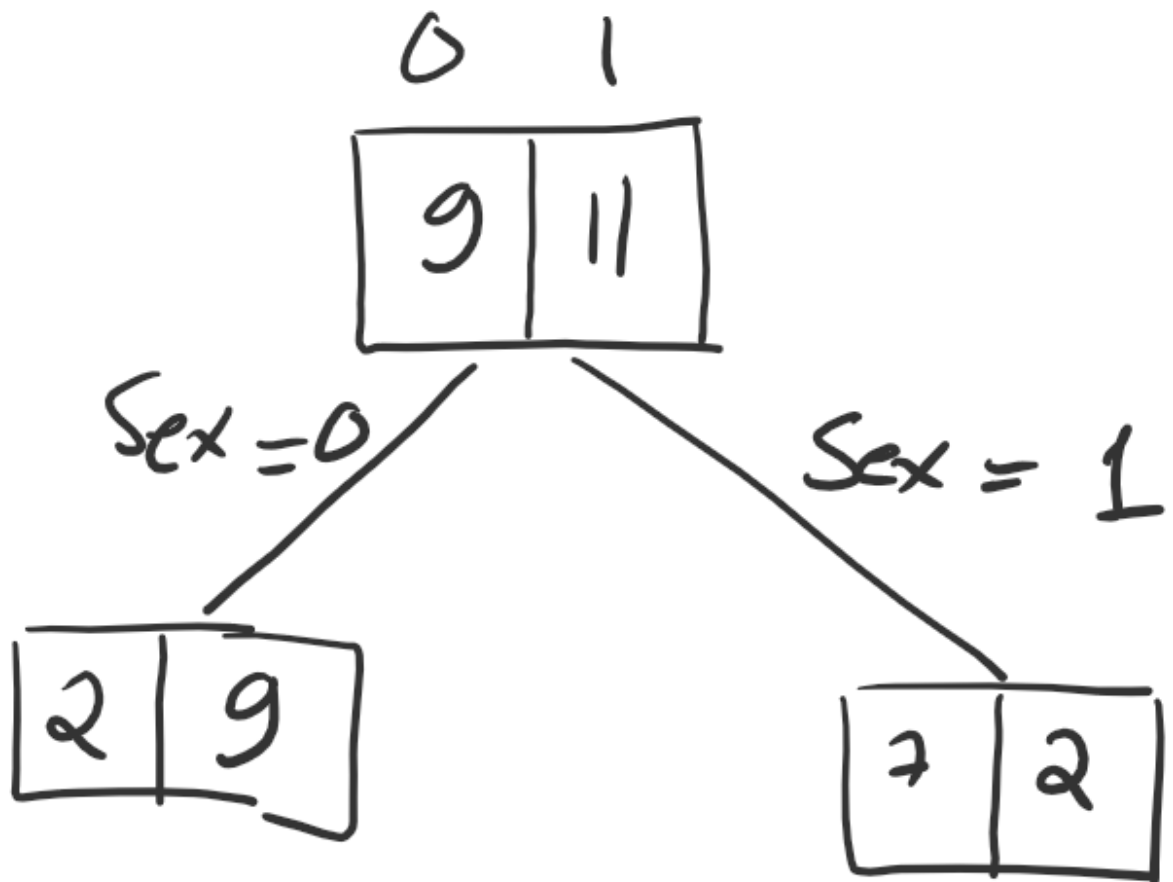
### Step 1: Decide the first split

We need to find the children impurity of all candidate splits for the first. What are candidate first splits?  
Here are all candidate first splits:

- Split 1: Split Sex into Male and Female
- Split 2: Split Class into class 1 and class 2,3
- Split 3: Split Class into class 2 and class 1,3
- Split 4: Split Class into class 3 and class 1,2

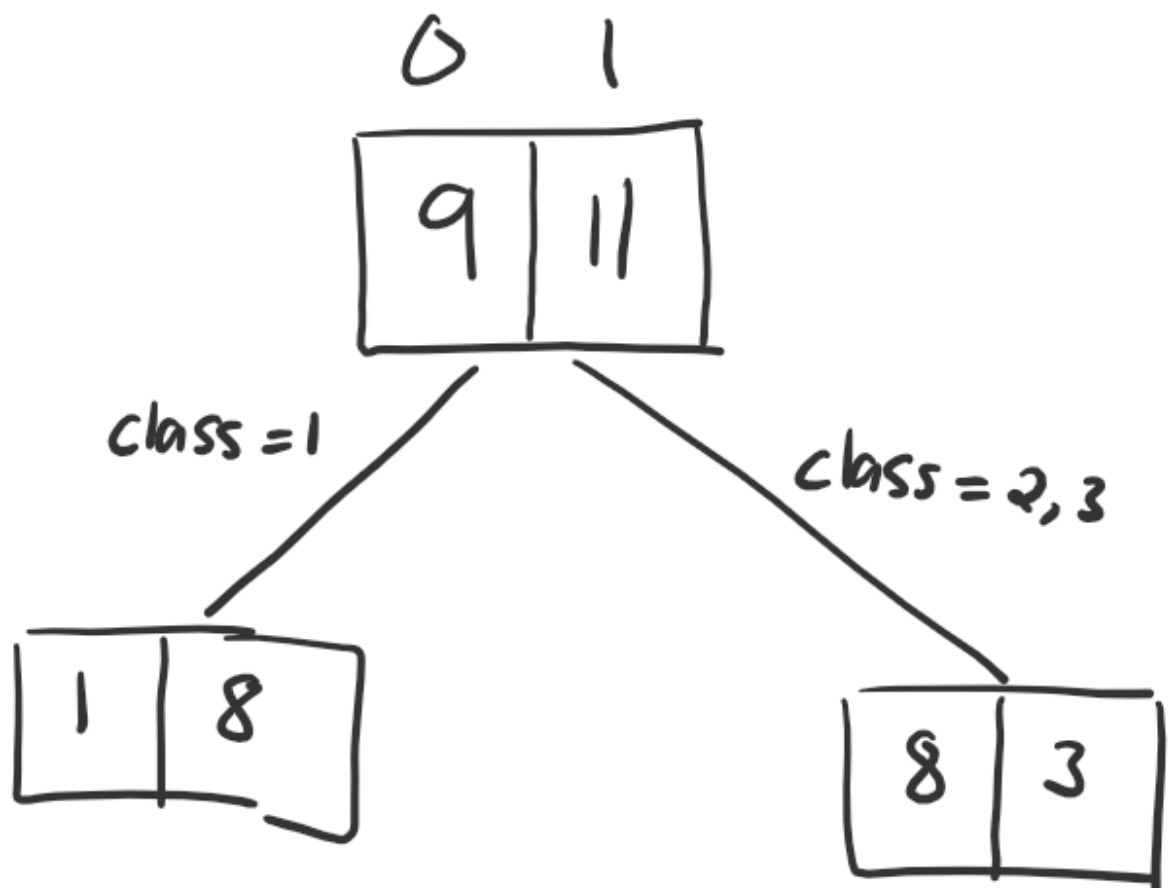
All of this splits have the same parent node. The impurity of the parent is

$$I_{parent} = 1 - \left(\frac{9}{20}\right)^2 - \left(\frac{11}{20}\right)^2 = 0.495$$



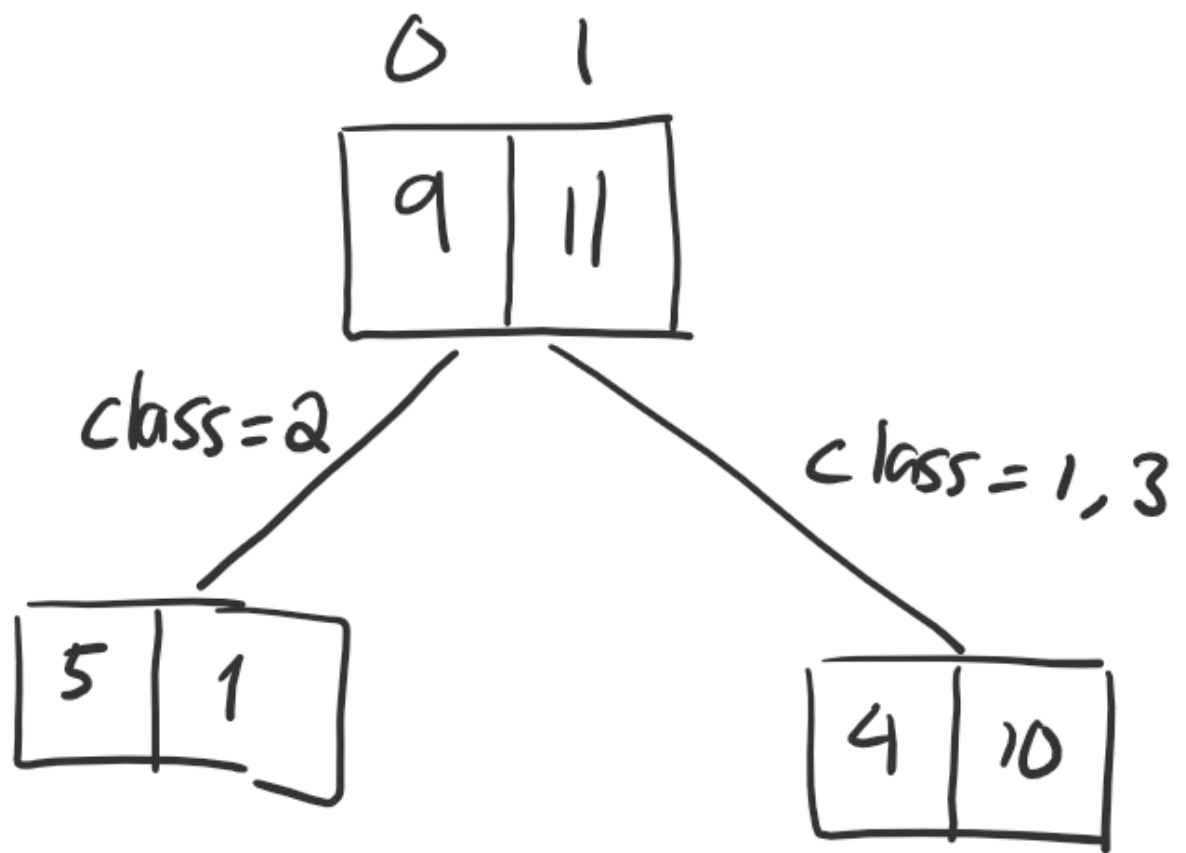
### Split 1:

The children impurity of Split 1: 0.3191919. The impurity gain of the split is:  $0.495 - 0.3191919 = 0.1758081$



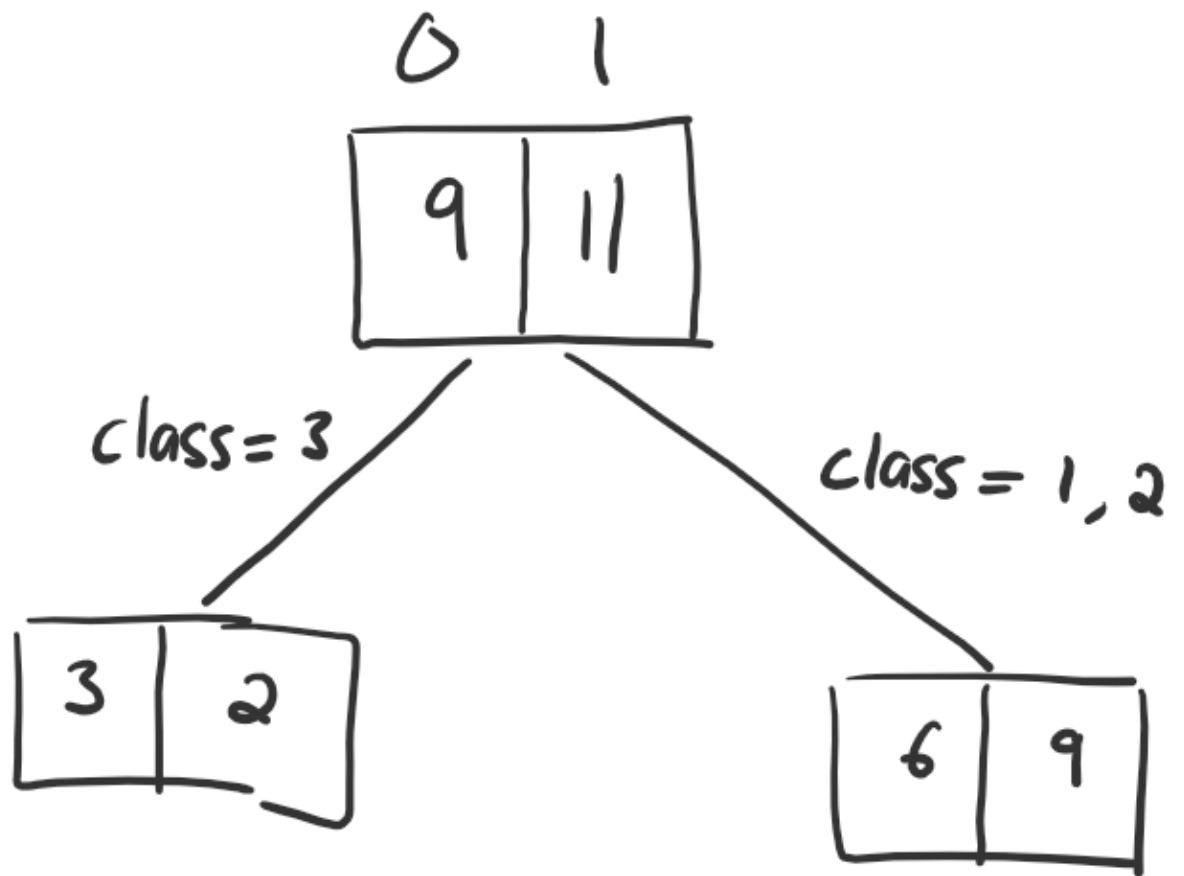
**Split 2:**

The children impurity of Split 2: 0.3070707. The impurity gain of the split is:  $0.495 - 0.3070707 = 0.1879293$



**Split 3:**

The children impurity of Split 3: 0.3690476. The impurity gain of the split is:  $0.495 - 0.3690476 = 0.1259524$



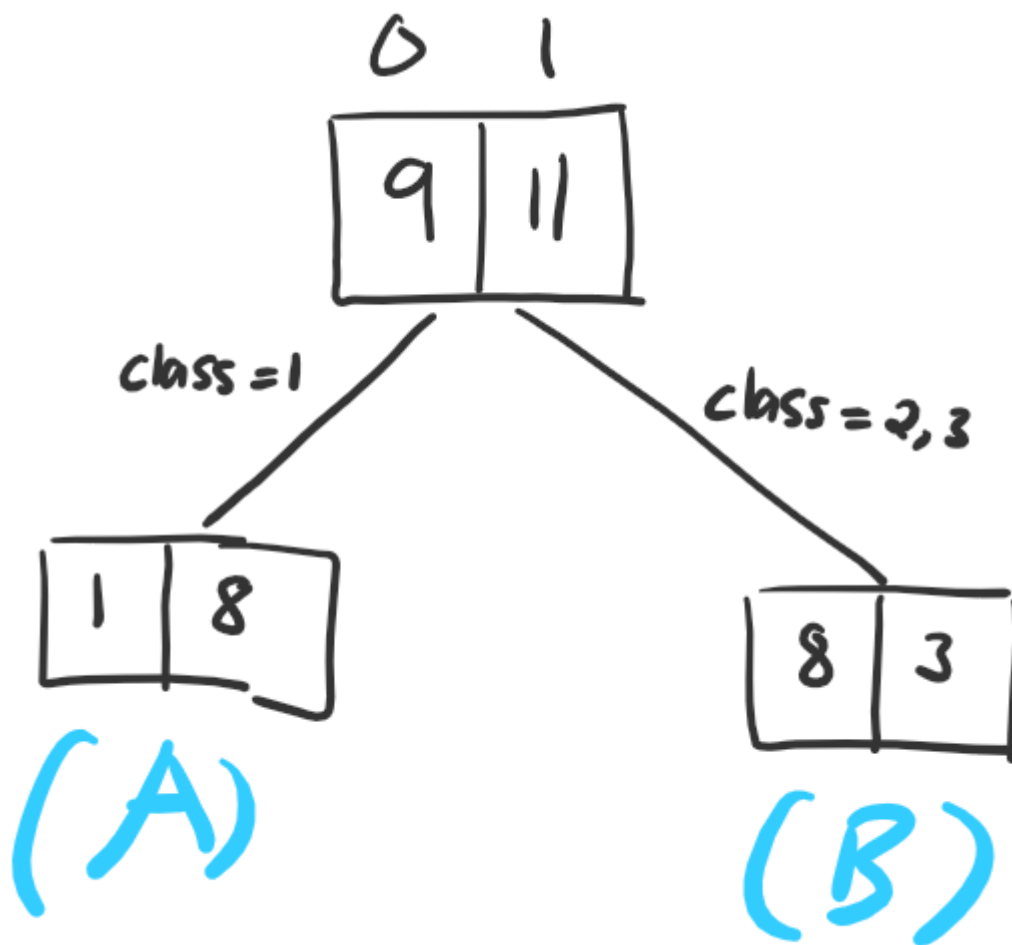
#### Split 4:

The children impurity of Split 4: 0.48. The impurity gain of the split is:  $0.495 - 0.48 = 0.015$

Split	Parent Impurity	Children impurity	Impurity Gain
1	0.495	0.3191919	0.1758081
2	0.495	0.3070707	0.1879293
3	0.495	0.3690476	0.1259524
4	0.495	0.48	0.015

The best split is the split with the highest impurity gain. Thus, split 2 is the first split of the tree.

We have the first Split as follow



## Step 2: Decide the second split

The next split could happen at a split at node A or node B

Let's look at the data in node A (Class = 1):

Class	Sex	Survived
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	1	1
1	1	0

What are candidate splits at node A? The only candidate split at Node A is splitting at Sex.

Let's look at the data in node B (Class = 2, 3):

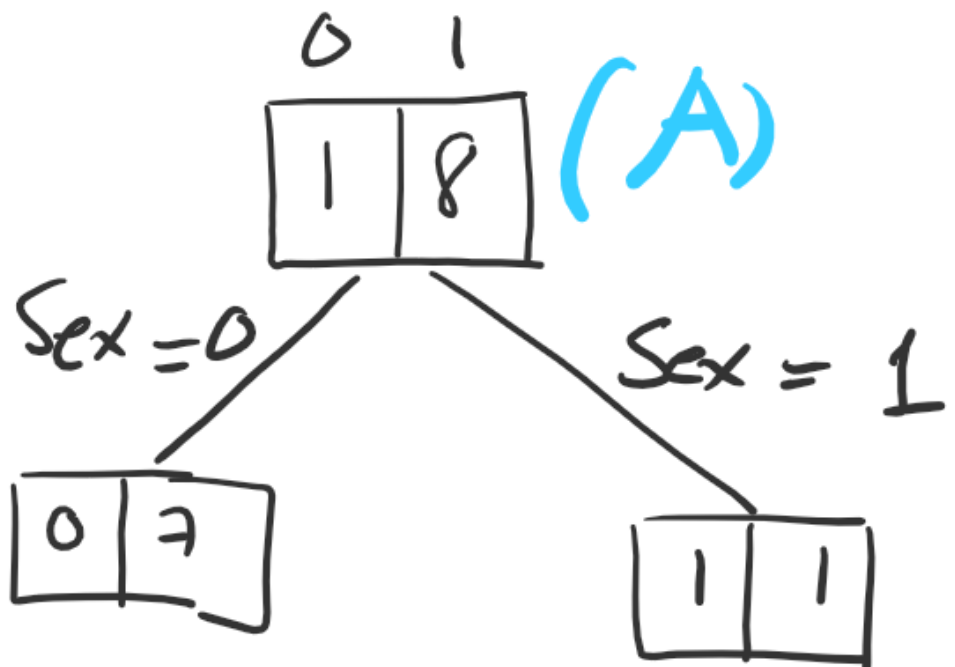
Class	Sex	Survived
3	0	1
2	1	1
3	0	1
2	1	0
2	1	0
3	1	0
2	1	0
3	0	0
3	1	0
2	0	0
2	1	0

Node B can be splitted at Sex and Class

Therefore, the candidates splits for the second split are

- Split node A at Sex
- Split node B at Sex
- Split node B at Class

We will compute the children impurity of all the candidate splits.

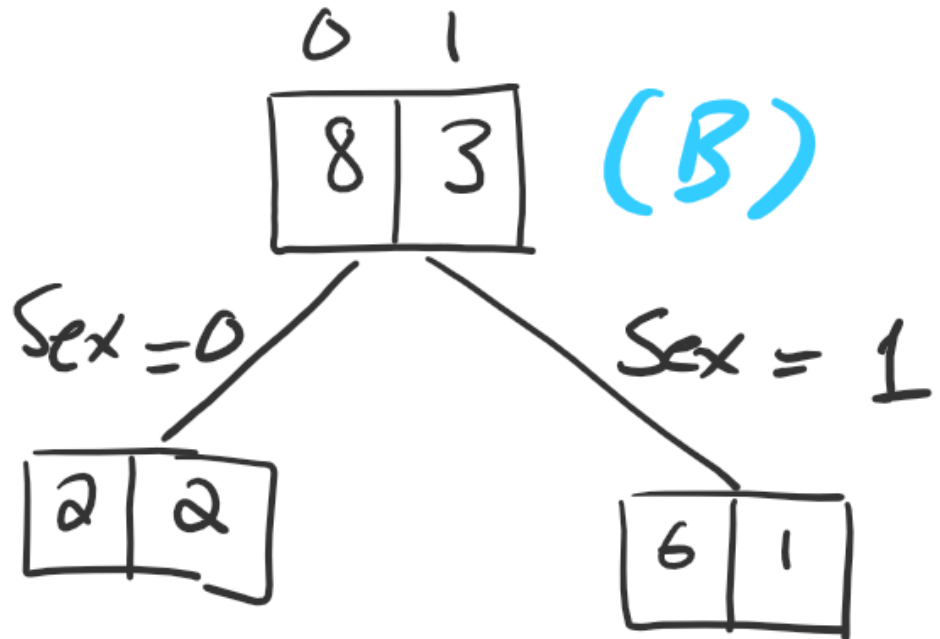


**Splitting node A at Sex:**

The children impurity of this split: 0.1111111.

$$I_{parent} = 1 - \left(\frac{1}{9}\right)^2 - \left(\frac{8}{9}\right)^2 = 0.1975309$$

The impurity gain of the split is  $0.1975309 - 0.1111111 = 0.0864198$

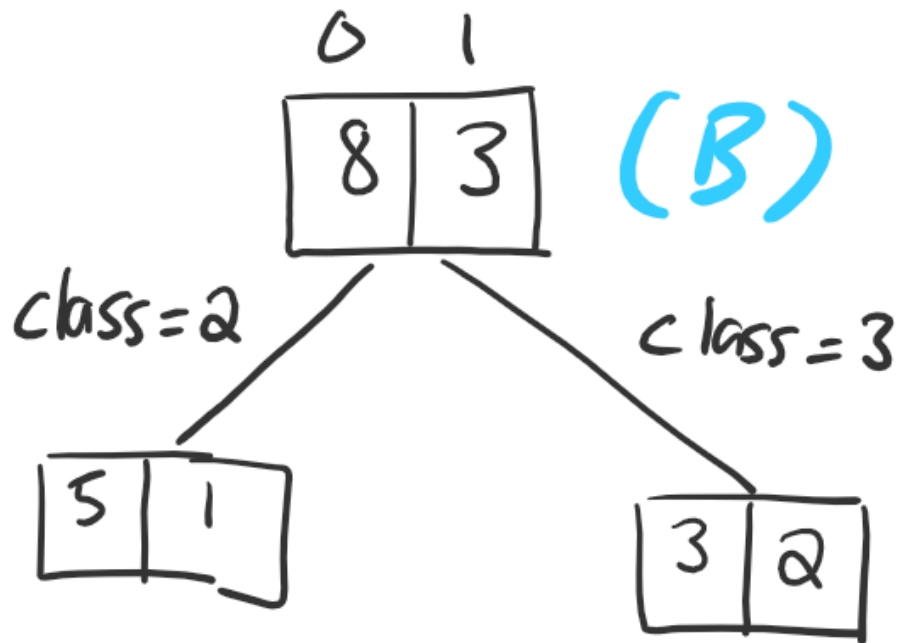


**Splitting node B at Sex:**

$$I_{parent} = 1 - \left(\frac{3}{11}\right)^2 - \left(\frac{8}{11}\right)^2 = 0.3966942$$

The children impurity of this split: 0.3376623. The impurity gain of the split is  $0.3966942 - 0.3376623 = 0.0590319$





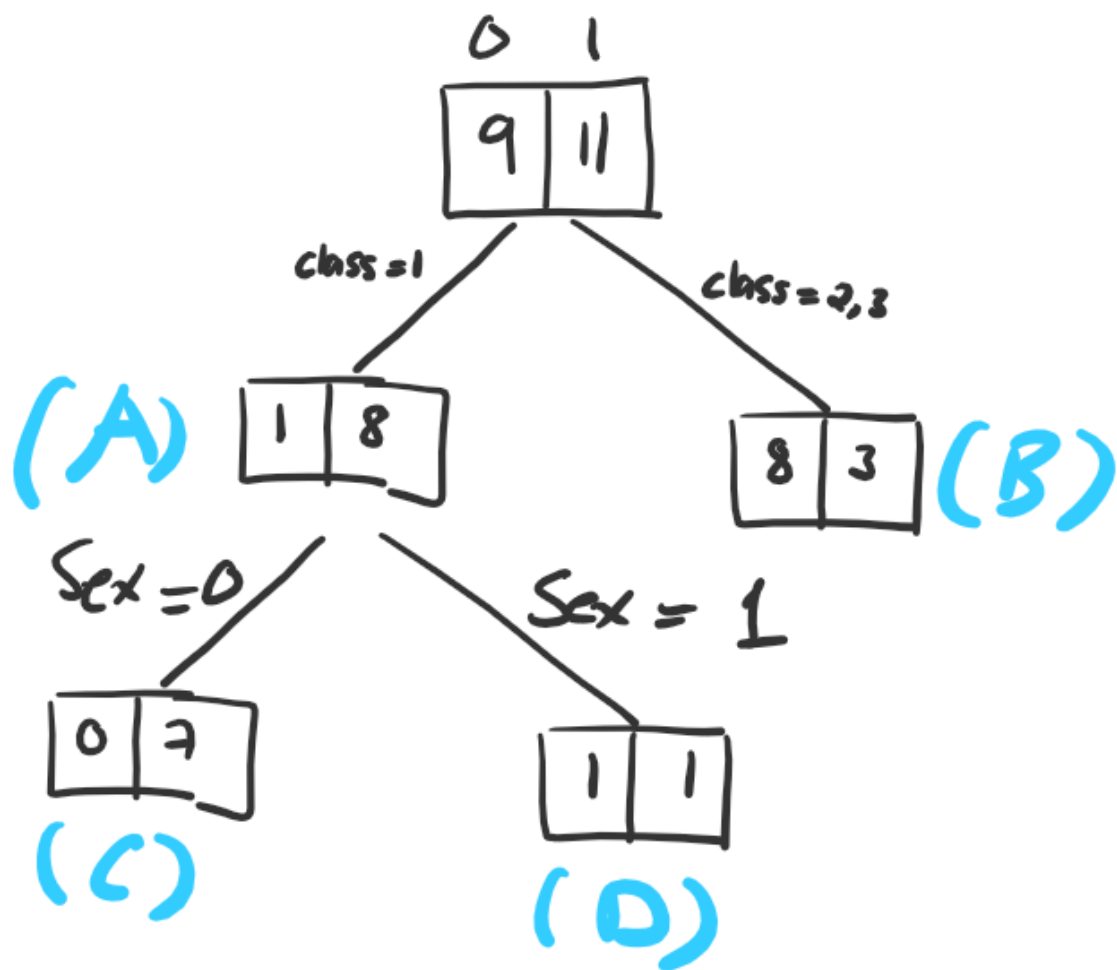
**Splitting node B at Class:**

$$I_{parent} = 1 - \left(\frac{3}{11}\right)^2 - \left(\frac{8}{11}\right)^2 = 0.3966942$$

The children impurity of this split: 0.369697. The impurity gain of the split is  $0.3966942 - 0.369697 = 0.0269972$

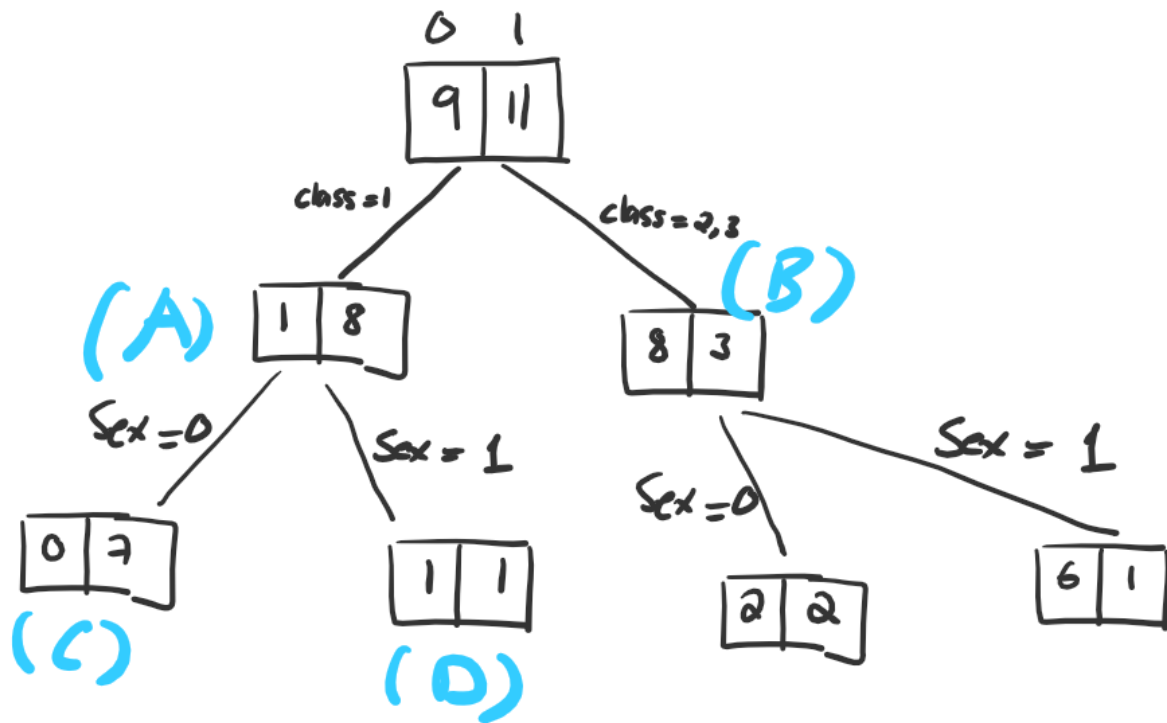
Candidate Split	Parent Impurity	Children impurity	Impurity Gain
Split node A at Sex	0.1975309	0.1111111	0.0864198
Split node B at Sex	0.3966942	0.3376623	0.0590319
Split node B at Class	0.3966942	0.369697	0.0269972

The best split is the split with the highest impurity gain. Thus, the second split is splitting node A at Sex  
Update the tree

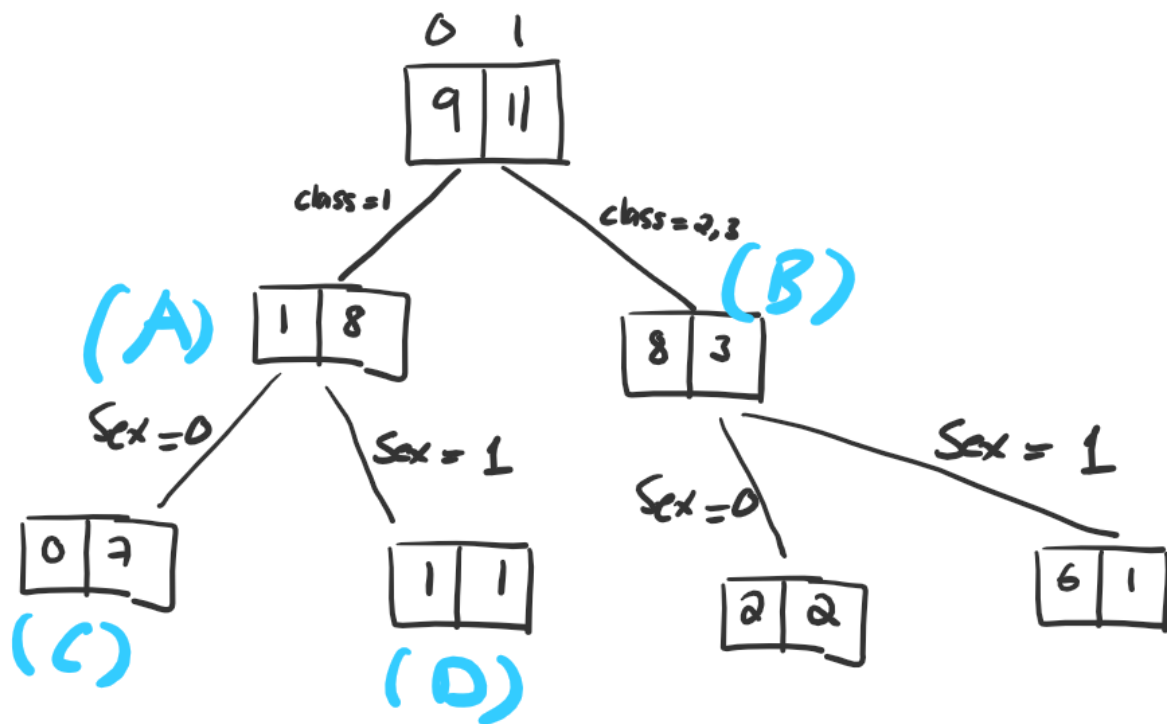


In Step 2, we already that nodes B can be splitted in two ways and the one with higher impurity is the split at Sex. Therefore, the third split is splitting node B at Sex

## Maximal Tree



## 2. Misclassification Rate



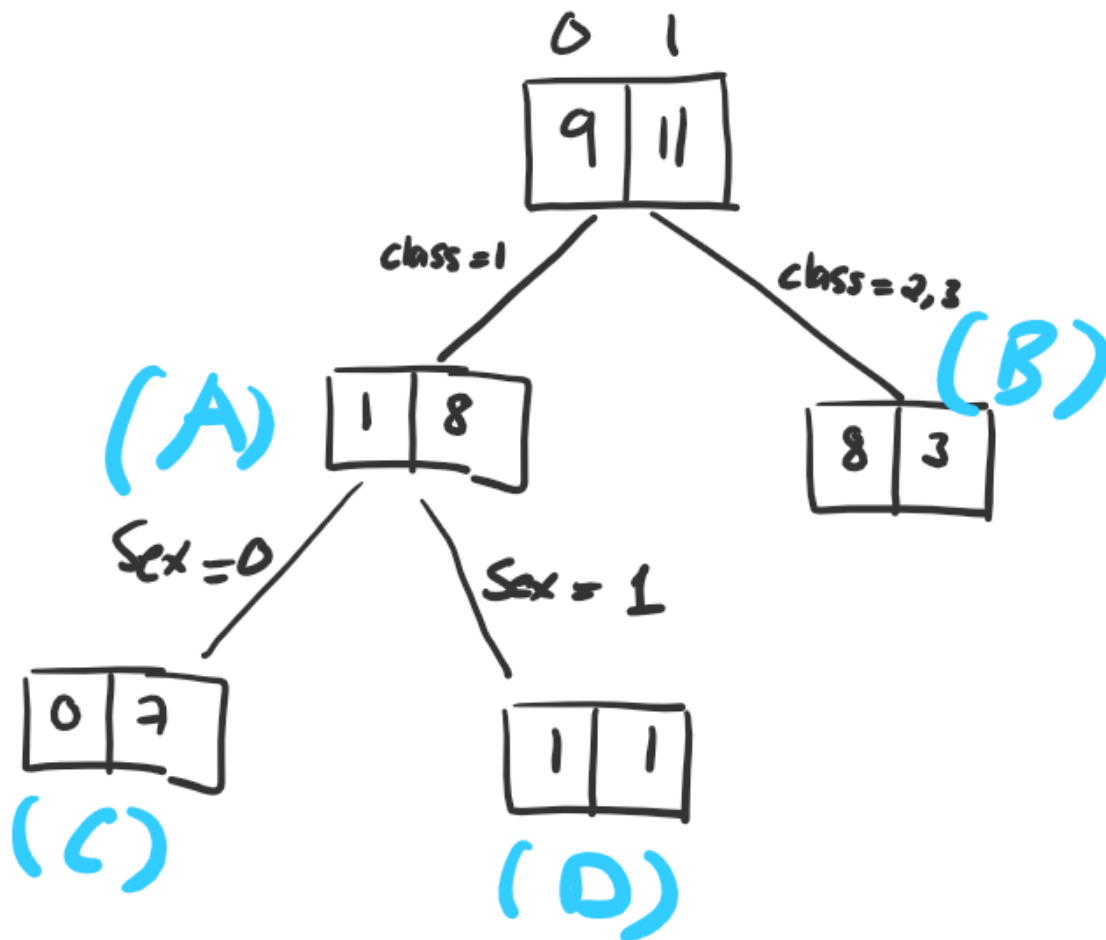
The prediction of each leaf is the majority. Thus, the misclassification of a leaf is the minority in the leaf.

- Total misclassification is:  $0+1+2+1=4$
- Misclassification Rate is  $4/20 = 20\%$

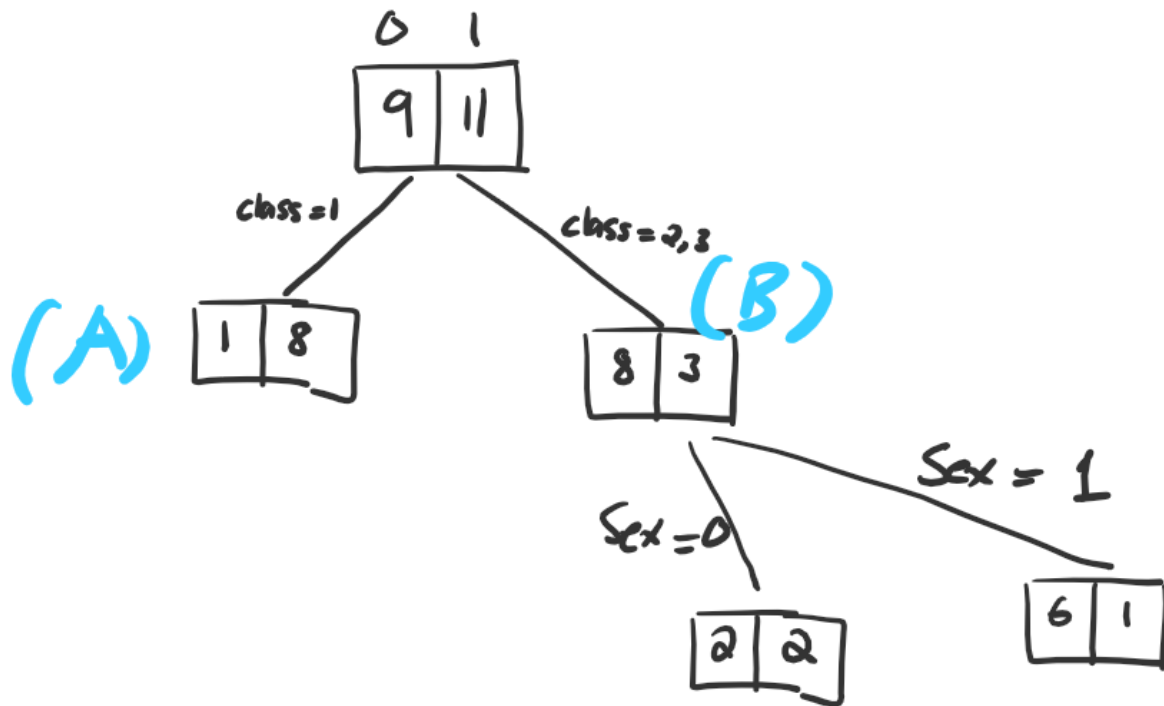
### 3. Subtrees

There are three subtrees that can be pruned from the maximal tree

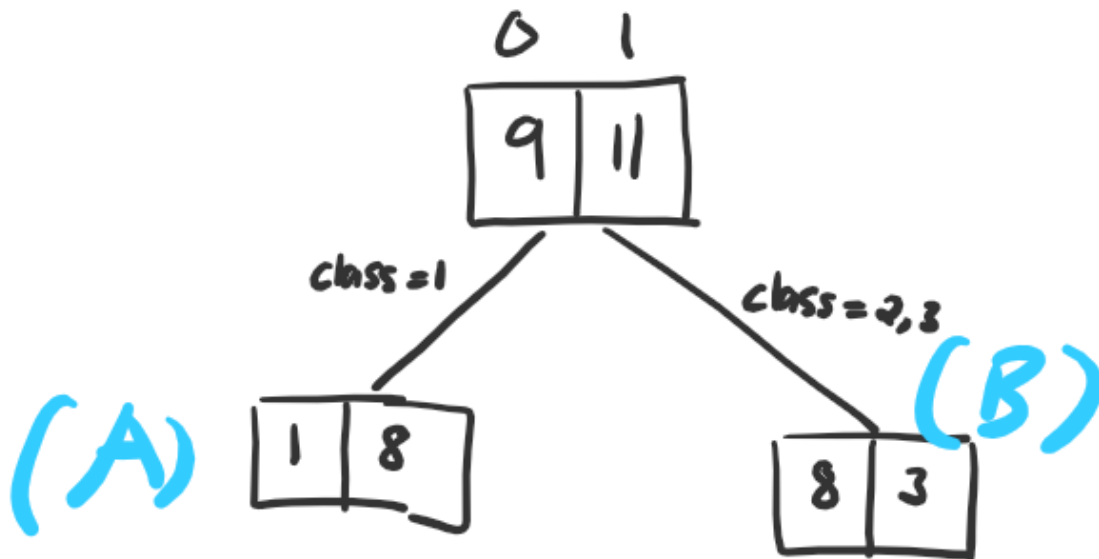
Subtree 1



Subtree 2



Subtree 3



## 4. Optimal Tree

Predictions of all the subtrees and the maximal tree

Class	Sex	Survived	Subtree1 Predicts	Subtree2 Predicts	Subtree3 Predicts	Maximal Tree Predicts
1	0	1	1	1	1	1
1	1	1	1	1	1	1
3	1	0	0	0	0	0
2	0	0	0	1(Missed!)	0	1 (Missed!)

- We see that Subtree 1 and Subtree 3 both have perfect predictions! (0 misclassification)
- Since Subtree 3 has less leaves than subtree 1, it is simpler than subtree 3.
- We select Subtree 3 as the optimal tree

Winning Model!

