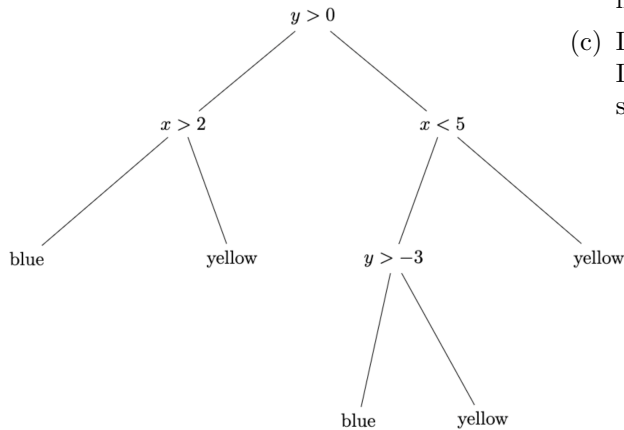


Homework 5: Tree-Based Models and Entropy

Complete the following exercises. Remember to explain your answers.

1. Consider the decision tree below. The left branches are True, and the right branches are False.



- (a) If $x = 1$ and $y = 1$, would the point be classified as blue or yellow?
- (b) If $x = 2$ and $y = -1$, would the point be classified as blue or yellow?
- (c) In the xy -plane, color in the decision regions. Ideally, use blue and yellow. Otherwise, just shade in the blue region.

2. Be sure to complete the worksheet from class on Tues 3/12 before starting this exercise. Consider the data in the table below, taken from the titanic dataset.

Sex	Age	Fare	Survived
male	34	6.4958	0
male	45.5	7.225	0
male	18	8.05	1
female	2	10.4625	0
male	32	15.85	0
male	26	18.7875	1
female	16	7.75	1
male	40	31	1
male	24	7.05	0
female	35	21	1

- (a) Compute the decrease in Gini impurity when you split on whether or not the **person is female**.
- (b) Compute the decrease in Gini impurity when you split on whether or not the **fare was greater than 17**.
- (c) To make an effective decision tree, would it be better to **split on Sex or on Fare** at the start of the tree? Explain your answer.

3. Consider the following training and testing accuracy scores for a classification model, for different values of a parameter.

Parameter value	Training accuracy	Testing accuracy
1	51%	50%
2	57%	55%
3	64%	63%
4	70%	70%
5	81%	79%
6	92%	91%
7	94%	86%
8	96%	82%
9	97%	81%
10	97%	80%

- (a) Give an example of a value for the parameter where we see **overfitting**. Explain your answer.
- (b) Give an example of a value for the parameter where we see **underfitting**. Explain your answer.
- (c) What value do you think is the **best choice** for this parameter? Explain your answer.
- (d) **Plot these accuracy scores**, with the parameter values on the x -axis and the accuracy scores on the y -axis. **Indicate in your graph where the best choice** the the parameter occurs.