# Data Mining Project 2

Q56061040 湯立婷

# Step 1: Design a set of rules to classify data, e.g., classify students with good performance.

我創造了**4**個Feature，分別是

- mass
- width
- height
- color_score

希望利用以上特徵辨別我買的水果為Orange(Positive) or Lemon(Negative)

# Step 1: Design a set of rules to classify data, e.g., classify students with good performance.

我建造的資料集數量為**35**個，使用的Absolutely right’s rule為

**Orange(Positive):**

mass大於等於200且width大於等於7.5 or

mass小於200且color_score大於等於0.75 or

mass小於200且color_score小於0.75且width大於等於7.5 or

mass小於200且color_score小於0.75且width小於7.5且height小於7.1

# Step 1: Design a set of rules to classify data, e.g., classify students with good performance.

我建造的資料集數量為**35**個，使用的Absolutely right’s rule為

**Lemon(Negative):**

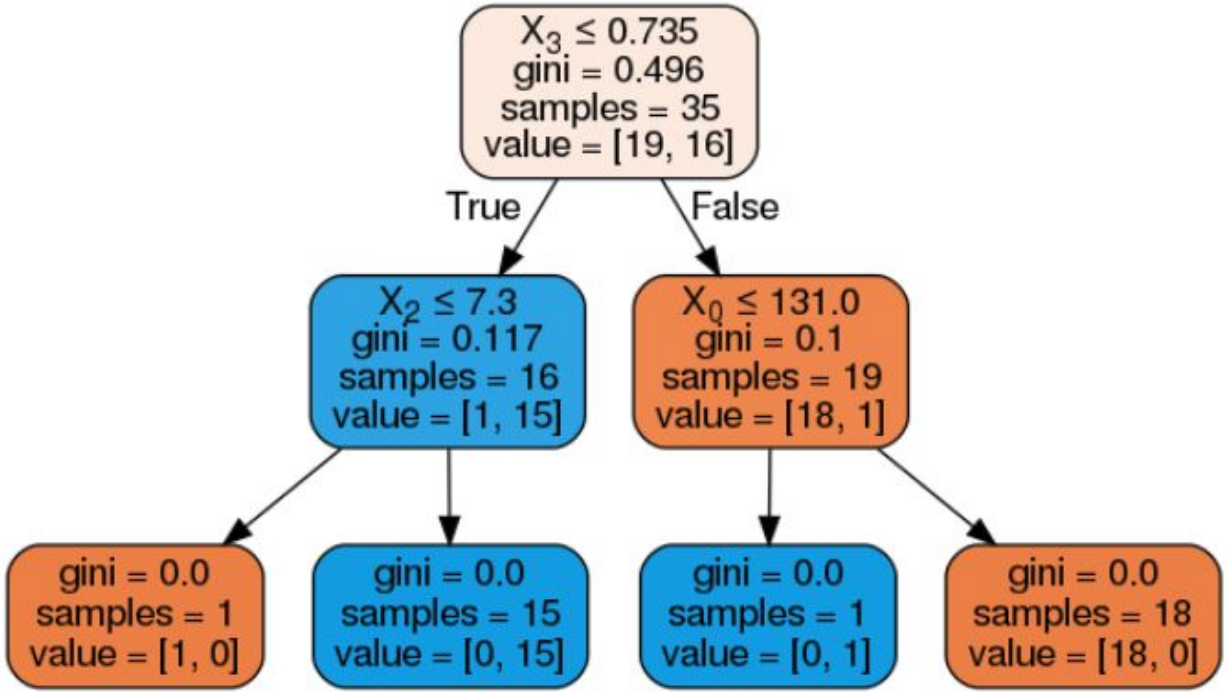
mass大於等於200且width小於7.5 or

mass小於200且color_score小於0.75且width小於7.5且height大於等於7.1

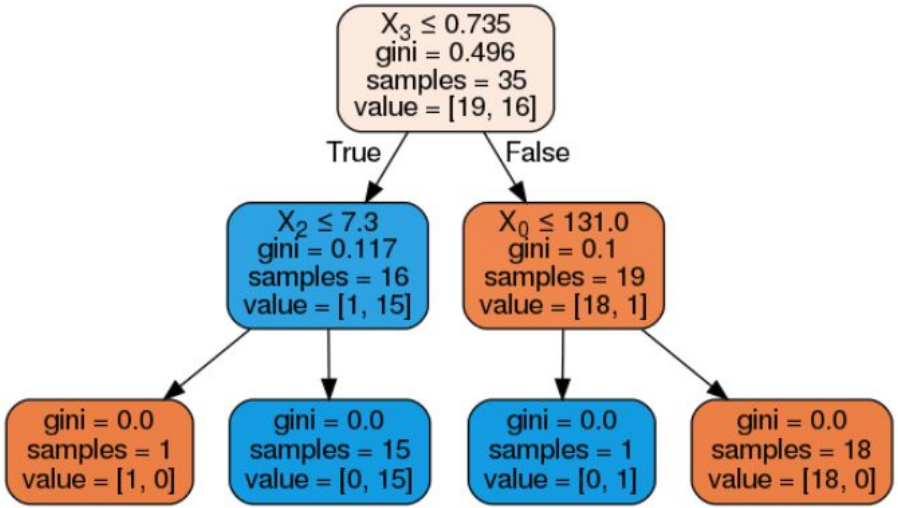
# Step 2: Use the data generated in Step 1 to construct your classification model

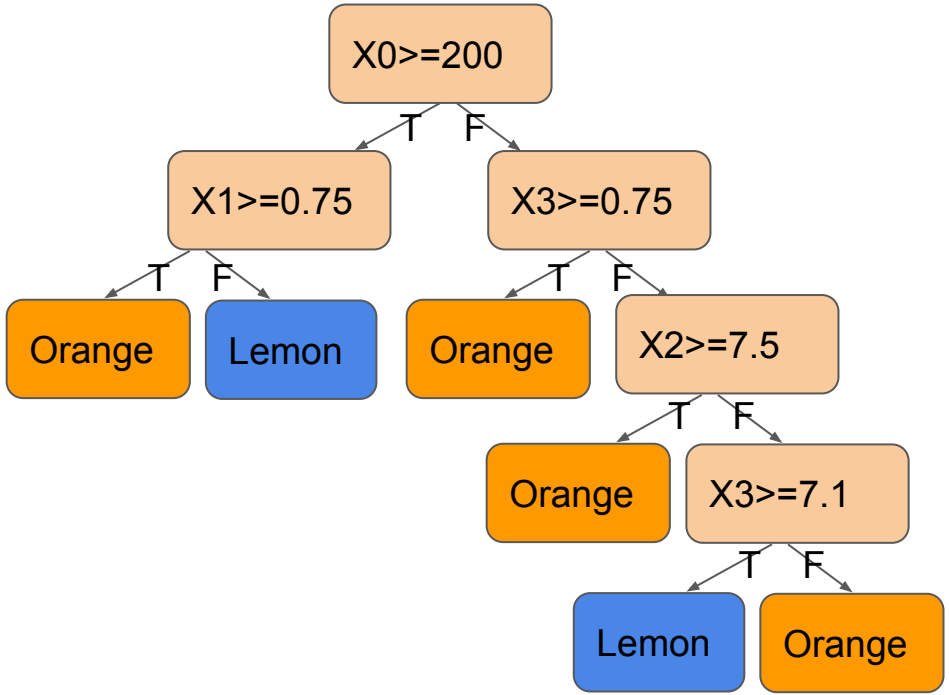
**Decision tree**

X0: mass
X1: width
X2: height
X3: color_score
橘色 : Orange
藍色 : Lemon



# Step 3: Compare the rules in the decision tree from Step 2 and the rules you used to generate your ‘right’ data



# Step 4: Discuss anything you can

- Decision tree會儘量減少樹根到樹葉的距離，以及挑選出真正能劃分的特徵
- 相比我的Absolutely right規則，我將width這個特徵也加進來，但事實上此特徵是多餘的，此外，因為node規則的劃分不好，也讓我從樹根到樹葉的距離比Decision tree多了兩層