Mini Quiz 1

下の決定木で、100個のデータが特徴量F1, F2によって、50個 (Group 1)、40個 (Group 2)、10個 (Group 3) の3つに 分けられたとする。scikir-learnの定義にしたがった各特徴量の重要度(feature importance) は、F1の方がF2よりも [A (大きい / 小さい)] 。

```
(100 data)
F1: 30歳以上？ (Gini不純度減少: 0.7)
  Yes:
    F2: 年収1000万以上？ (Gini不純度減少: 0.3)
      Yes:
        Group 1 (50 data)
      No:
        Group 2 (40 data)
  No:
    Group 3 (10 data)
```

Ans. of Mini Quiz 1

下の決定木で、100個のデータが特徴量F1, F2によって、50個 (Group 1)、40個 (Group 2)、10個 (Group 3) の3つに 分けられたとする。scikir-learnの定義にしたがった各特徴量の重要度(feature importance) は、F1の方がF2よりも [A 大きい] 。

(100 data)
F1: 30歳以上？ (Gini不純度減少: 0.7) ← F1: 100 data 中 100 data を分けた
  Yes:
    F2: 年収1000万以上？ (Gini不純度減少: 0.3) ← F2: 100 data 中 90 data を分けた
      Yes:
        Group 1 (50 data)
      No:
        Group 2 (40 data)
  No:
    Group 3 (10 data)

各特徴量の重要度
 = ジニ不純度減少分 **x**
  木に入力されたデータ数に対する分けたデータ数の割合

Mini Quiz 1

In the following decision tree, 100 data are divided into 50 data (Group 1), 40 data (Group 2), and 10 data (Group 3) by the features F1 and F2. According to the definition of feature importance in scikit-learn, the feature importance of F1 is [A. (larger / smaller)] than that of F2.

```
(100 data)
F1: >= 30 years old ? (decrease of Gini impurity: 0.7)
  Yes:
    F2: Annual income >= 10,000,000 Yen? (Gini impurity: 0.3)
      Yes:
        Group 1 (50 data)
      No:
        Group 2 (40 data)
  No:
    Group 3 (10 data)
```

Ans. of Mini Quiz 1

In the following decision tree, 100 data are divided into 50 data (Group 1), 40 data (Group 2), and 10 data (Group 3) by the features F1 and F2. According to the definition of feature importance in scikit-learn, the feature importance of F1 is [A. larger] than that of F2.

F1: divide 100 data out of 100 data

(100 data)
F1: >= 30 years old ? (decrease of Gini impurity: 0.7)
  Yes:
    F2: Annual income >= 10,000,000 Yen? (Gini impurity: 0.3)
      Yes:
        Group 1 (50 data)
      No:
        Group 2 (40 data)
  No:
    Group 3 (10 data)

F2: divide 90 data out of 100 data

Feature importance
= the decrease of Gini impurity
 x the proportion of the number of separated data to the number of input data