2. (50 points) One of the challenges in decision trees is selecting the appropriate splitting criterion. In this problem, we will learn a new splitting criterion called Gini-Index.

(a) Step 1 : Research about Gini-Index and briefly explain how it differs from Information Gain.

(0 - 1)
Entropy = $\sum(-p \log(p))$, p = probability

Entropy: the measurement of the impurity or randomness in the data points.

Information Gain is applied to quantify which feature provides maximal information about the classification based on the notion of entropy
Information Gain = Entropy(parent) - [weighted average]Entropy(children)

(0 - 1)
Gini Index/Impurity = $1 - \sum(p^2)$

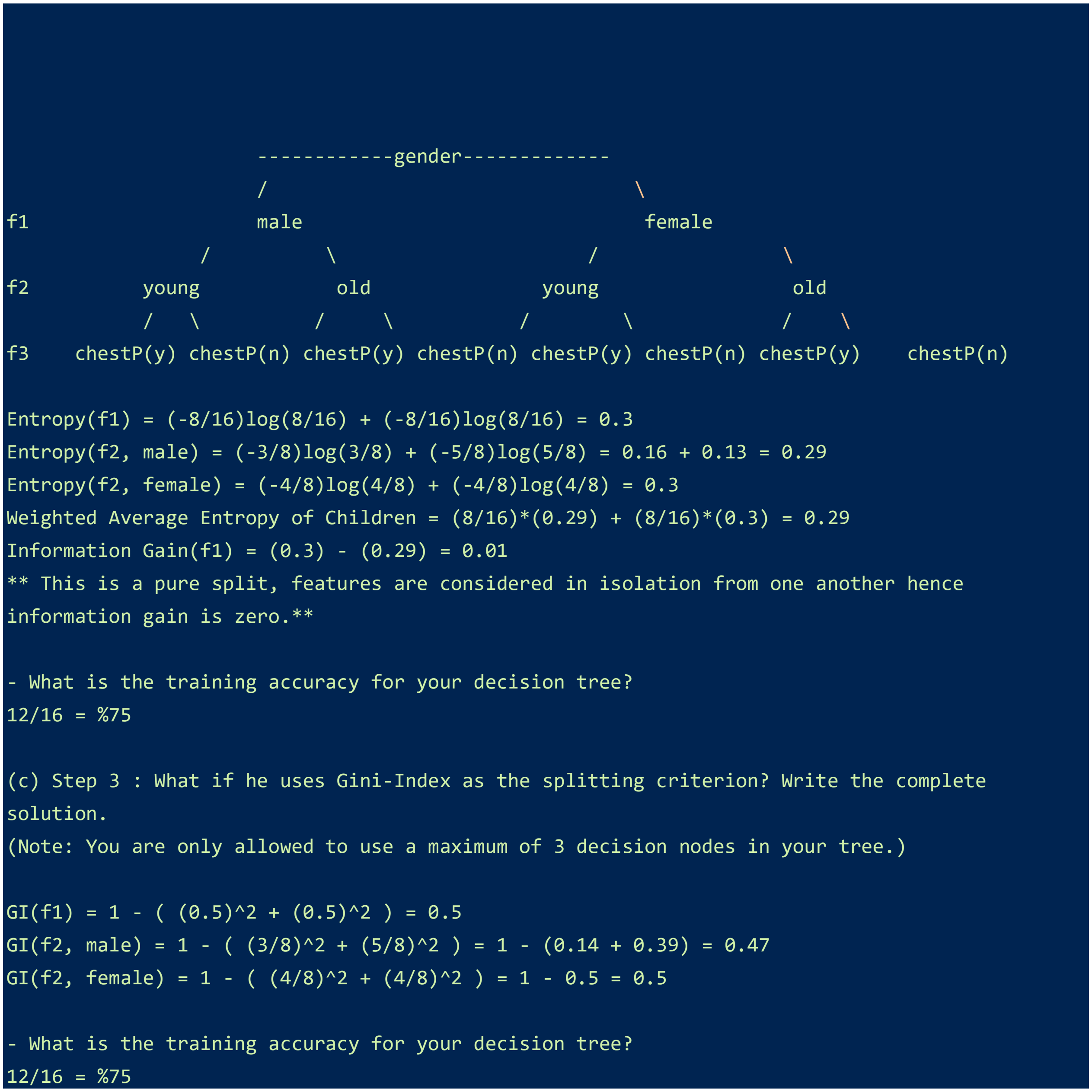
Gini Index/Impurity: calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.
If all the elements are linked with a single class then it can be called pure.

Answer(a):

The Gini Index facilitates the bigger distributions easy to implement on the other hand Information Gain favors lesser distributions having a smaller count with multiple specific values.
Gini index operates on the categorical target variables in terms of “success” or “failure” and performs only binary split,
whereas Information Gain computes the difference between entropy before and after the split and indicates the impurity in classes of elements.

(b) Step 2 : A researcher has discovered a new cure for a disease, however, it can cause some side effects. To predict it’s side effects, he has tested that on 16 volunteer patients.
The table below presents his findings. Now, he wants to use decision trees to make a prediction for future patients using four features (gender, age, smoker, chest pain).
(Note: this is just a toy problem, not a real world data!)

- What is the decision tree for this problem, if he uses Information Gain as splitting criterion? Write the complete solution. (Note: You are only allowed to use a maximum of 3 decision nodes in your tree.)



```
                      ------------gender-------------
                      /                                \
f1                    male                              female
                 /          \                      /                \
f2          young            old               young                 old
            /   \          /     \           /        \             /    \
f3    chestP(y) chestP(n) chestP(y) chestP(n) chestP(y) chestP(n) chestP(y)    chestP(n)

Entropy(f1) = (-8/16)log(8/16) + (-8/16)log(8/16) = 0.3
Entropy(f2, male) = (-3/8)log(3/8) + (-5/8)log(5/8) = 0.16 + 0.13 = 0.29
Entropy(f2, female) = (-4/8)log(4/8) + (-4/8)log(4/8) = 0.3
Weighted Average Entropy of Children = (8/16)*(0.29) + (8/16)*(0.3) = 0.29
Information Gain(f1) = (0.3) - (0.29) = 0.01
** This is a pure split, features are considered in isolation from one another hence
information gain is zero.**

- What is the training accuracy for your decision tree?
12/16 = %75

(c) Step 3 : What if he uses Gini-Index as the splitting criterion? Write the complete
solution.
(Note: You are only allowed to use a maximum of 3 decision nodes in your tree.)

GI(f1) = 1 - ( (0.5)^2 + (0.5)^2 ) = 0.5
GI(f2, male) = 1 - ( (3/8)^2 + (5/8)^2 ) = 1 - (0.14 + 0.39) = 0.47
GI(f2, female) = 1 - ( (4/8)^2 + (4/8)^2 ) = 1 - 0.5 = 0.5

- What is the training accuracy for your decision tree?
12/16 = %75
```