Problem 2

1. Build a decision tree:
   The entropy equation is:
   
   \[ E = -\sum_j P(w_j) \log P(w_j) \]

   According to the data, we have the total labels 9+, 4-. So the entropy for the whole data is:

   \[ E = -\frac{9}{13} \log \frac{9}{13} - \frac{4}{13} \log \frac{4}{13} = 0.8905 \]

   If we classify the data with Wakeup:
   Early: 4+
   \[ E = -\frac{4}{4} \log \frac{4}{4} = 0 \]
   Normal: 2+, 3-
   \[ E = -\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5} = 0.9710 \]
   Late: 3+, 1-
   \[ E = -\frac{3}{4} \log \frac{3}{4} - \frac{1}{4} \log \frac{1}{4} = 0.8113 \]

   \[ \text{Gain}(\text{Wakeup}) = 0.8905 - \frac{4}{13} \cdot 0 - \frac{5}{13} \cdot 0.9710 - \frac{4}{13} \cdot 0.8113 = 0.2674 \]

   If we classify the data with HaveTalk:
   Yes: 6+, 1-

   \[ \]
\[ E = -\frac{6}{7} \log \frac{6}{7} - \frac{1}{7} \log \frac{1}{7} = 0.5917 \]

No: 3+, 3-

\[ E = -\frac{3}{6} \log \frac{3}{6} - \frac{3}{6} \log \frac{3}{6} = 1 \]

\[ \text{Gain}(\text{HaveTalk}) = 0.8905 - \frac{7}{13} \cdot 0.5917 - \frac{6}{13} \cdot 1 = 0.1104 \]

If we classify the data with Weather:

Sunny: 6+, 2-

\[ E = -\frac{6}{8} \log \frac{6}{8} - \frac{2}{8} \log \frac{2}{8} = 0.8113 \]

Rain: 3+, 2-

\[ E = -\frac{3}{5} \log \frac{3}{5} - \frac{2}{5} \log \frac{2}{5} = 0.9710 \]

\[ \text{Gain}(\text{Weather}) = 0.8905 - \frac{8}{13} \cdot 0.8113 - \frac{5}{13} \cdot 0.9710 = 0.0178 \]

Since \( \text{Gain}(\text{Wake up}) \) is the largest one, we choose Wakeup in this step.

The following steps are skipped since they are quite similar to this one.

The final decision tree is:
2. According to the tree learned, the sample should be classified to NO.

3. Yes, the sample can be classified with missing data: *marginalize*. If Wakeup = Early, definitely the student will go to school. If Wakeup = Normal, since HaveTalk = Yes, the student will go to school. If Wakeup = Late, since Weather = Sunny, the student will go to school. So for this sample, no matter what the wake up time is, the student will go to school.