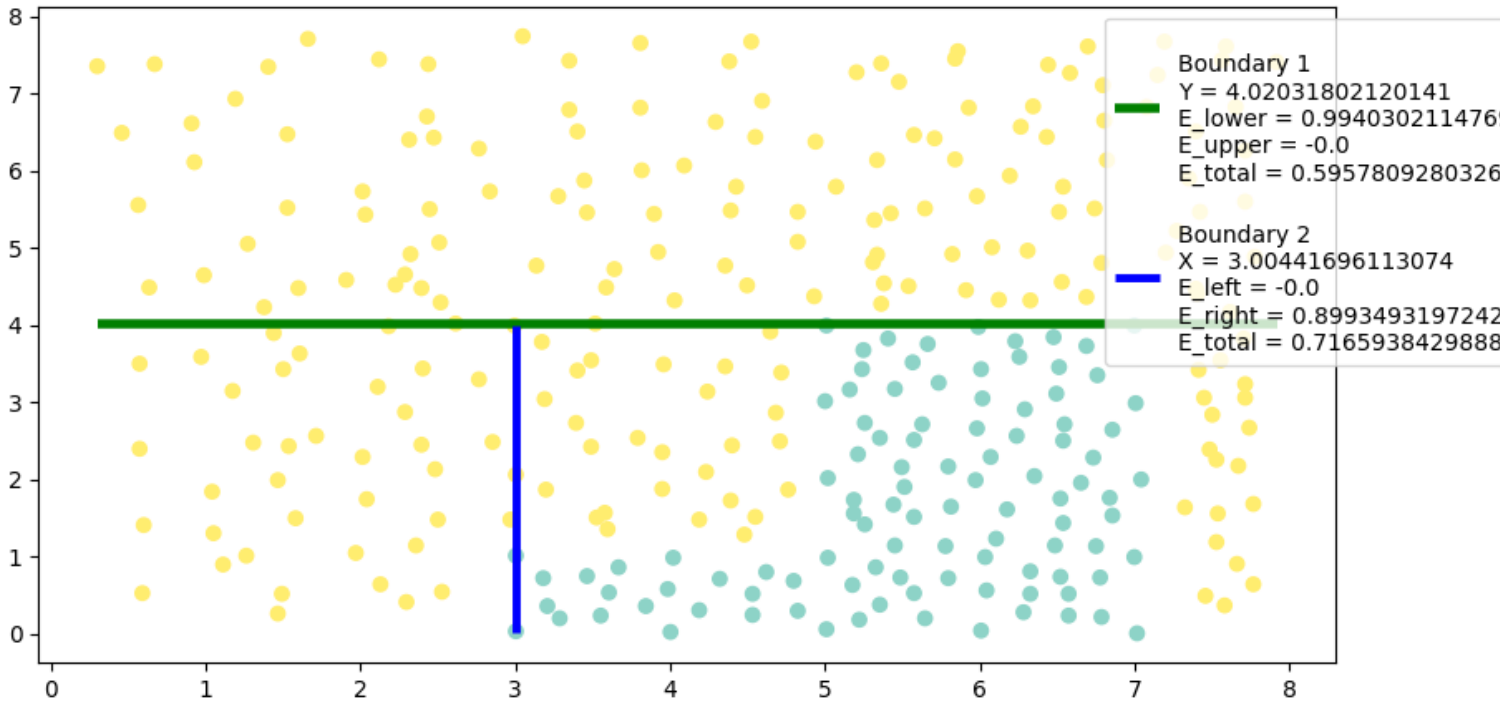
CMPE 481 ASSIGNMENT 2

ADALET VEYİS TURGUT - 2017400210

0. Algorithm

This is a recursive algorithm.
At each consecutive call, the level parameter is increased by 1 to indicate current depth.
There are two stop conditions: when depth(level) is equals to the maxlevel or when splitted subdatas are pure.
Possible split points are the data points. I chose this approach instead of taking the averages of two consecutive data points.
The point where information gain is maximum is decided as the split point.

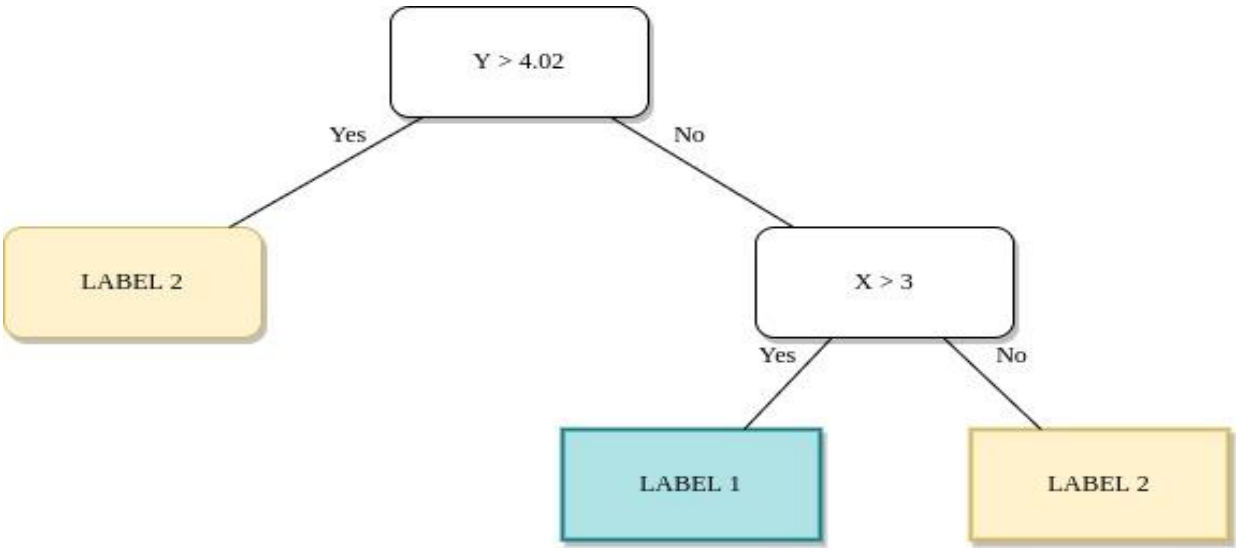
1. Decision Boundaries



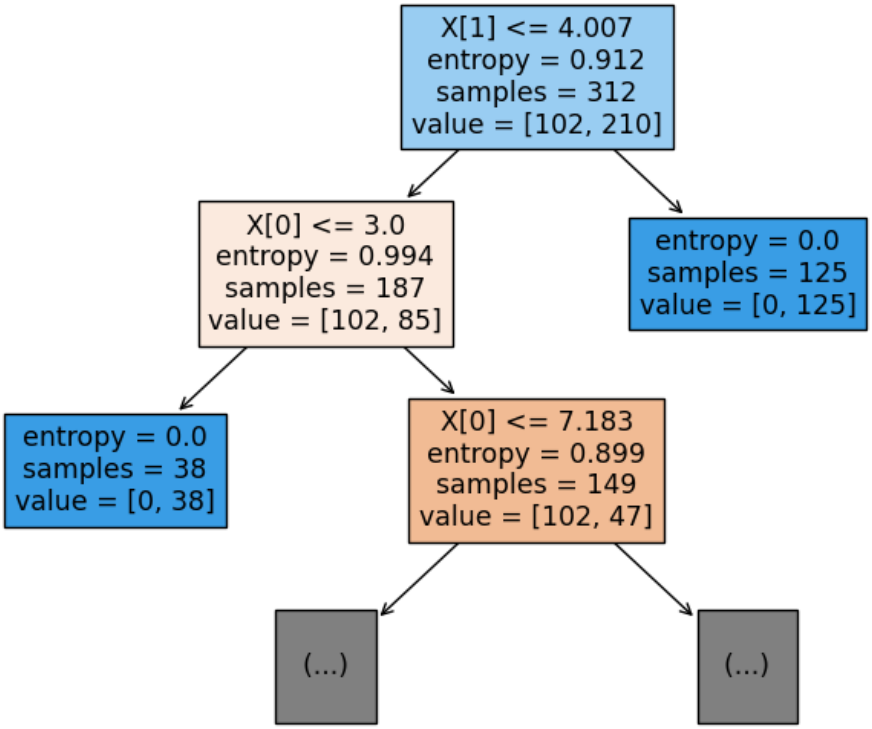
Yellow colored points are labeled 2, turquoise colored points are labeled 1.

First two levels of the decision boundaries are shown, boundary 1 is thicker than the boundary 2. Corresponding entropy levels are in the figure.

2. Plot of the Decision Tree



3. Comparison with scikit-learn



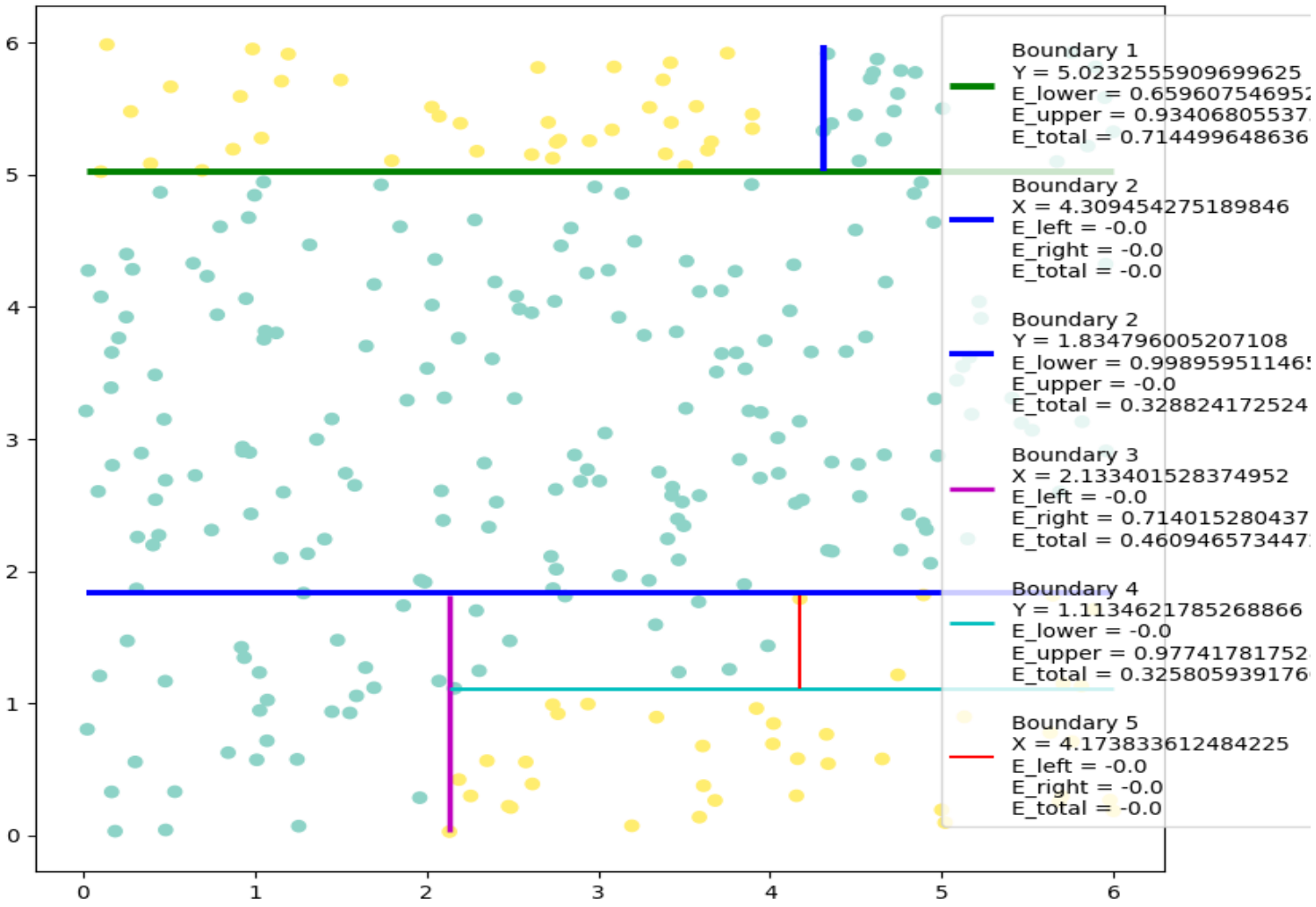
Results are similar, only differ in the percentile of the numbers. I think the reason is I took points as separators instead of the average of two consecutive points.

I labeled [X>3: yes] node as LABEL 1 since it was the most frequent term in that part, but scikit-learn wanted to grow the tree full instead of labeling it.

Also entropy values are the same: my E_lower from boundary 1 is equals to the [X[0] <=3] node of the scikit-learn, my E_right from Boundary 2 is equals to the [X[0] <=7.183] node of scikit-learn.

4. My Own Dataset

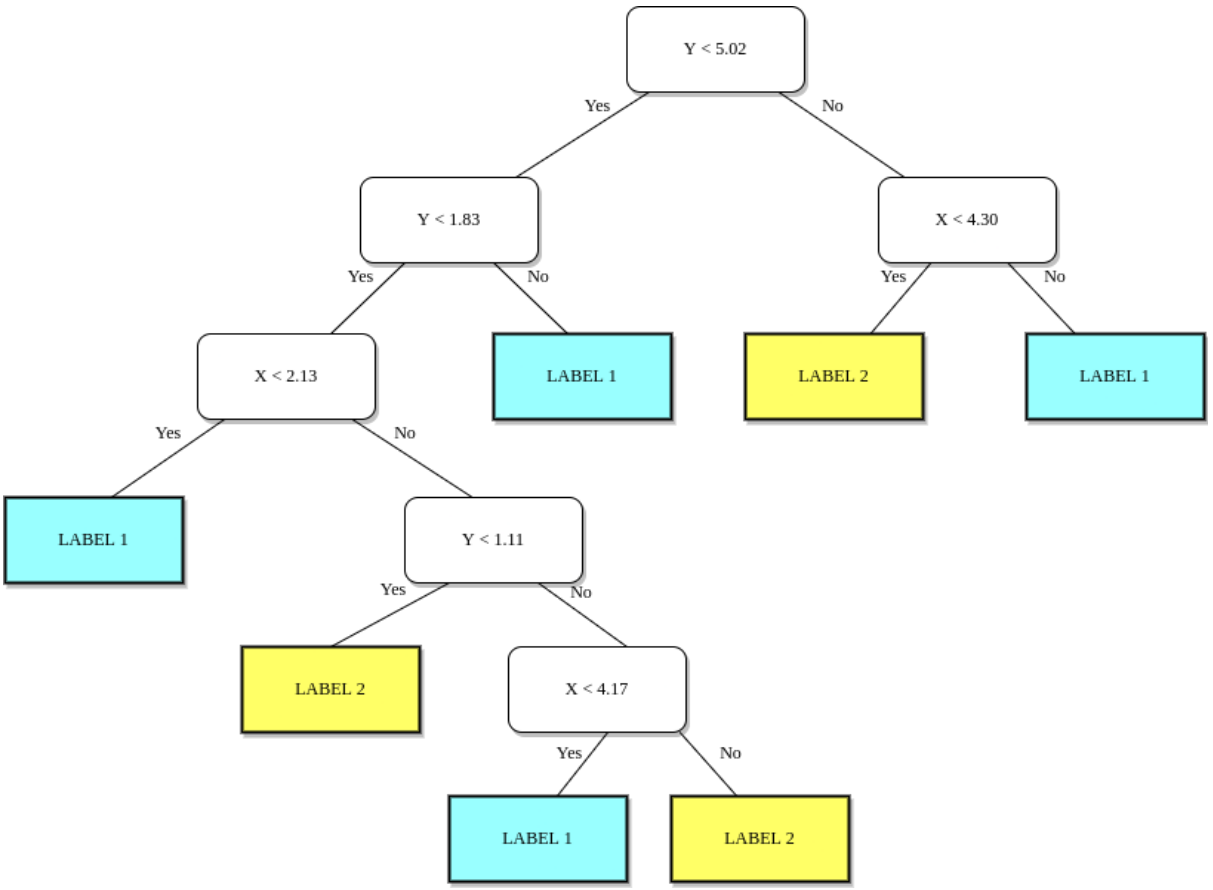
4.1 Decision Boundaries



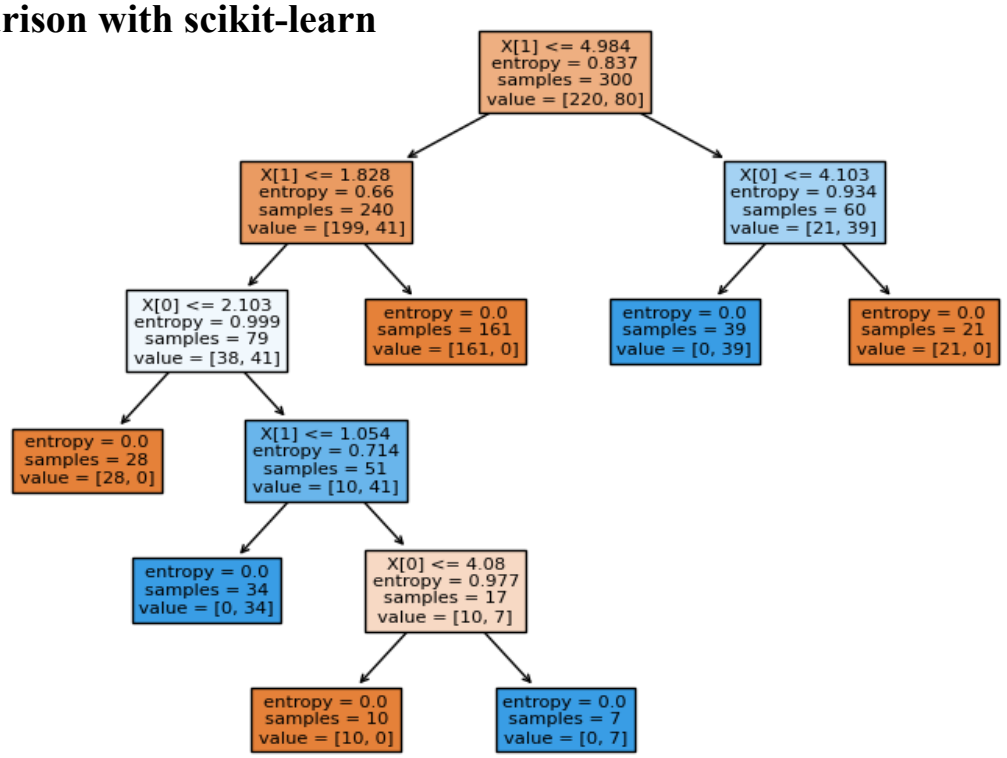
I used python random library for generating the dataset. Generator code is the following:

```
def myData():
    random.seed(10) # make seed 10
    data = []
    for _ in range(300): # generate 300 data points
        x = random.random()*6 # x values are between 0,6
        y = random.random()*6 # y values are between 0,6
        if (x < 4 and y > 5) or (x > 4 and y < 2) or (2<x and x<4 and y<1):
            # label as 1 if the condition above is true
            data.append([(x,y),1])
        else:
            data.append([(x,y),0])
    return data
```

4.2 Plot of the Decision Tree



4.3 Comparison with scikit-learn



Results are almost the same. Again, like the first dataset, split points differ slightly. But entropy calculations are the same.