Assignment 2 Report

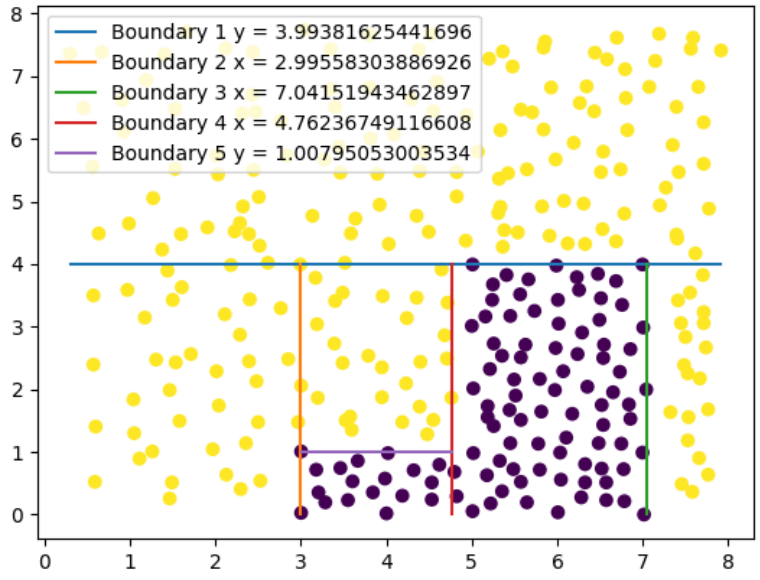
Decision Tree Construction

In this project we are assigned to implement decision tree classifier. We are given a dataset which has two features to classify. I used recursion to implement decision tree classifier. First, I load the data. Then I call find_split method two find best split point which has most information gain. So in find_split method I am calling two different functions to calculate information gain on x and y axis separately. I compared returned functions so that I can select which axis has bigger information gain. Based on the result I am partitioning this data two separate data and recursively sending these data to find_split method. As a base condition for recursion, I am using if information gain is zero than stop. While returning I am saving boundaries on each axis, min and max value on these axes, entropy values for both left and right branches, weighted averages, and information gain. Finally, I am plotting these boundaries on figure along with data points.

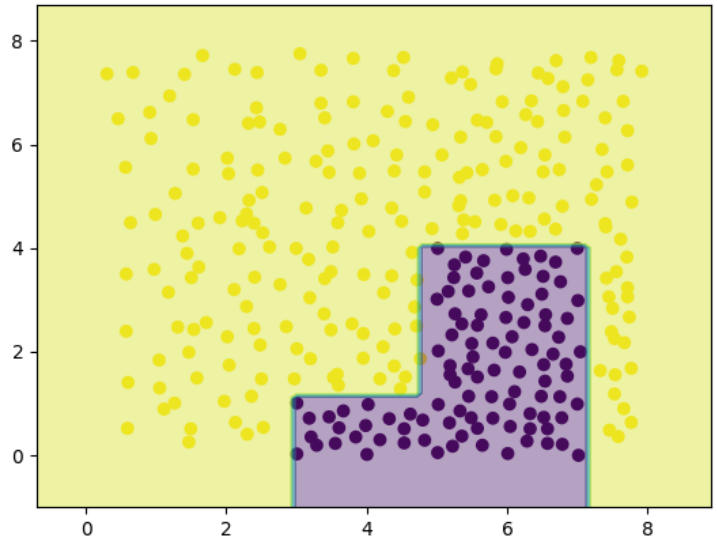
Comparison with Scikit Learn

There is a small difference between my output and scikit learn output, but I can confirm that I am classifying all datapoints correctly. I think the difference causes from selecting boundary point. I am selecting last point that matches the rule, but I think scikit learn starts the boundary from the next point. The difference between these boundaries is very small and there is no data point laying on these lines so there is no misclassifying in my algorithm.

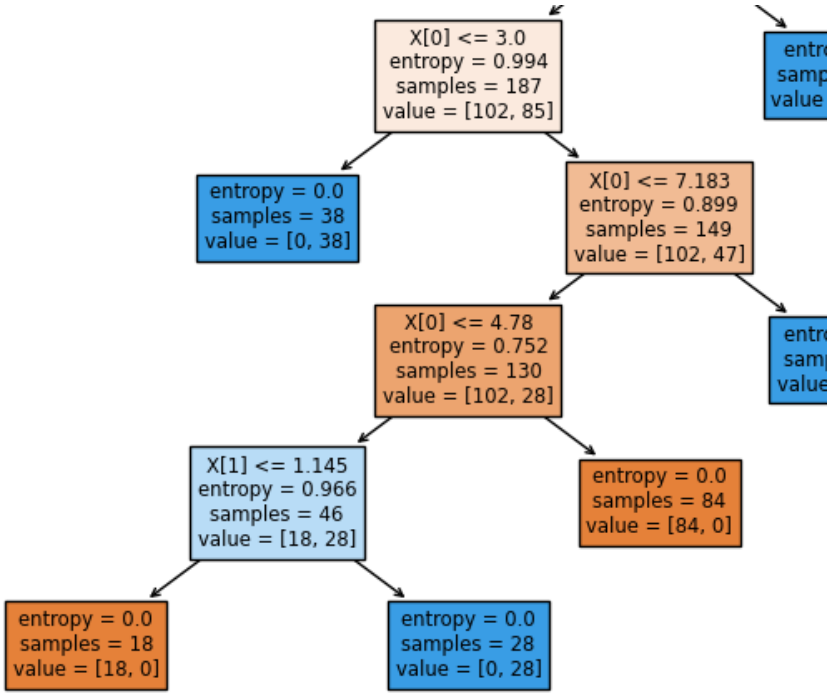
In my second dataset I have classified all datapoint successfully. Same difference between skicit learn and my algorithm continues. I think scikit learn takes a point between two different class points.



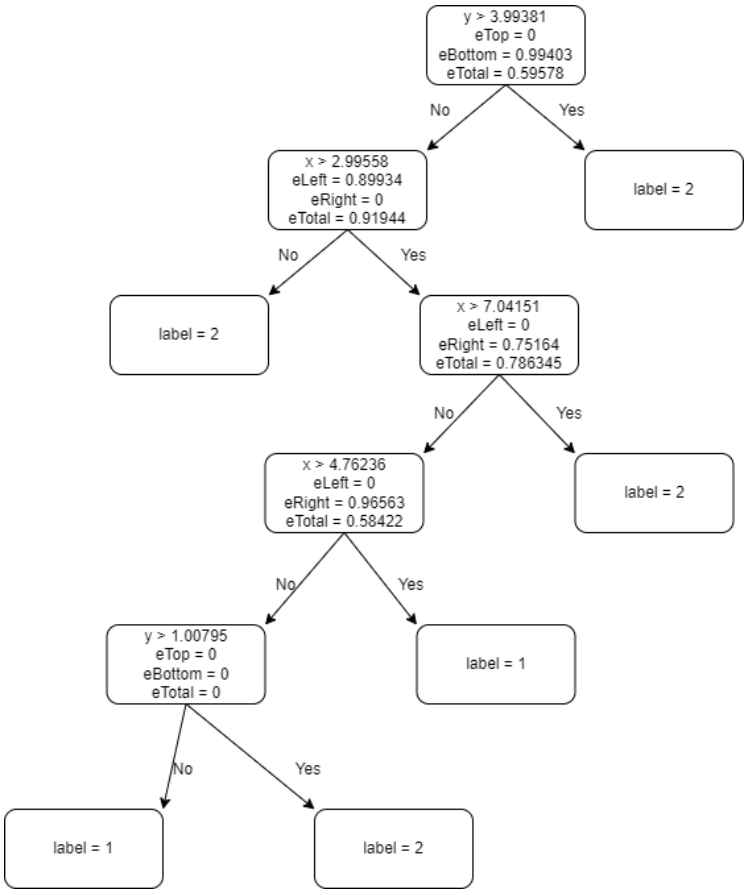
My Output 1



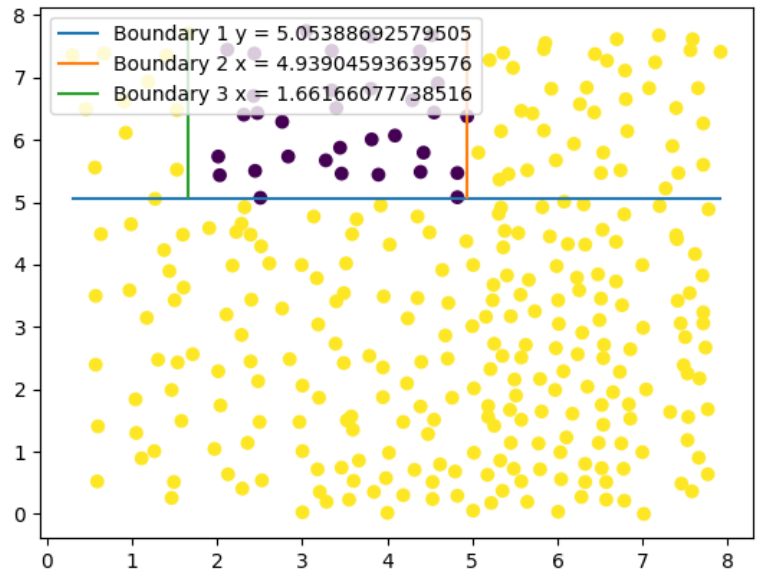
Scikit Learn Output 1



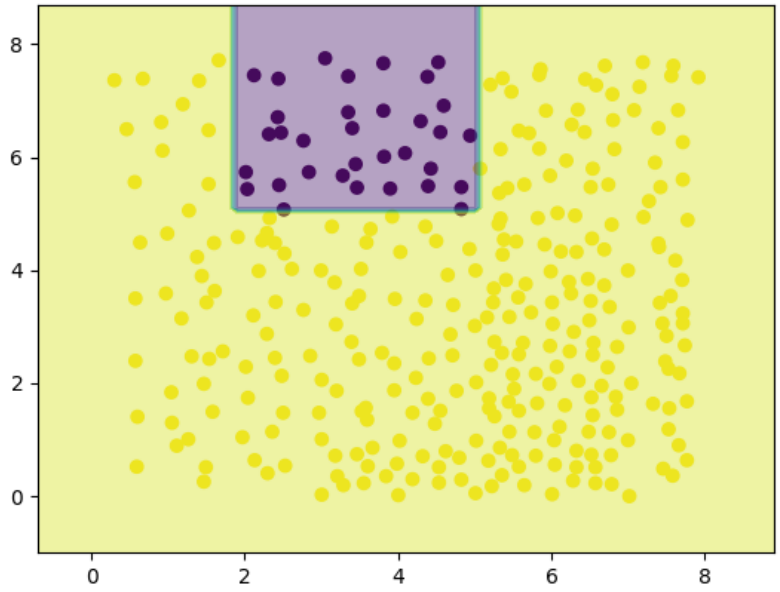
Scikit Learn Tree 1



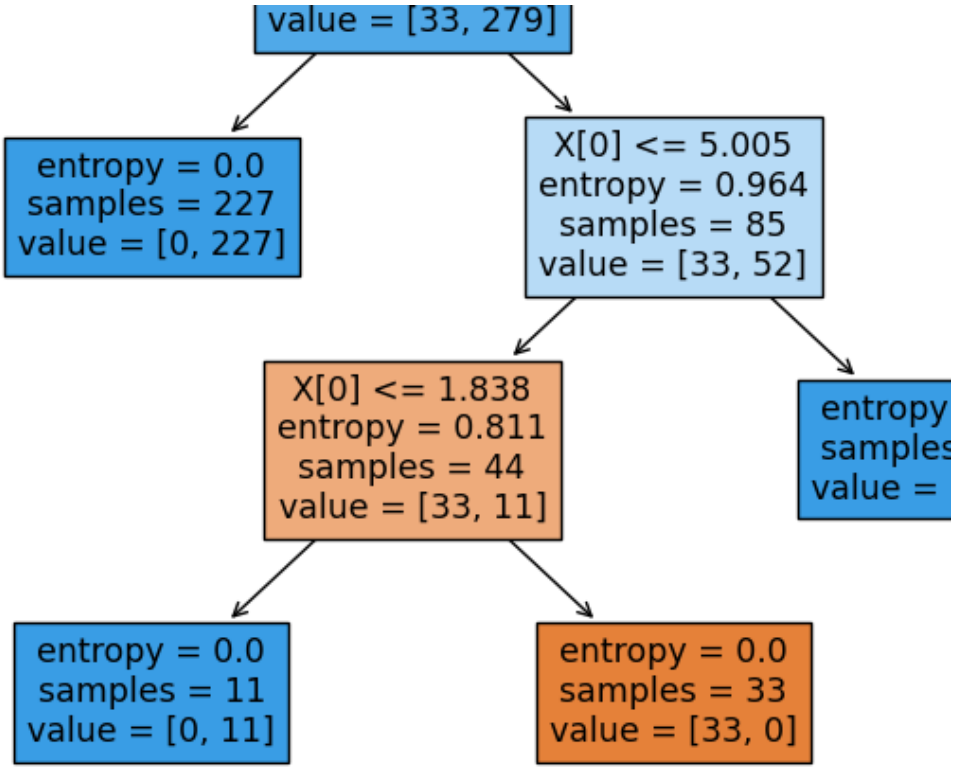
My Tree 1



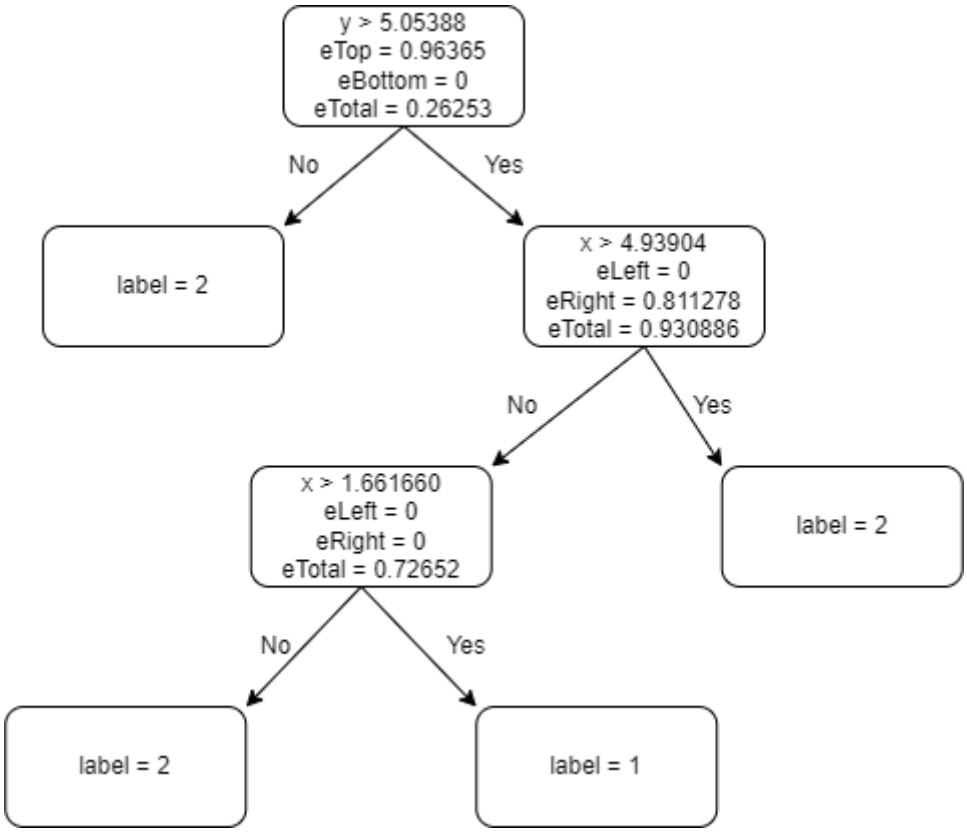
Second Data My Plot 1



Second Data Scikit Learn 1



Second Data Scikit Tree 1



Second Data My Tree 1