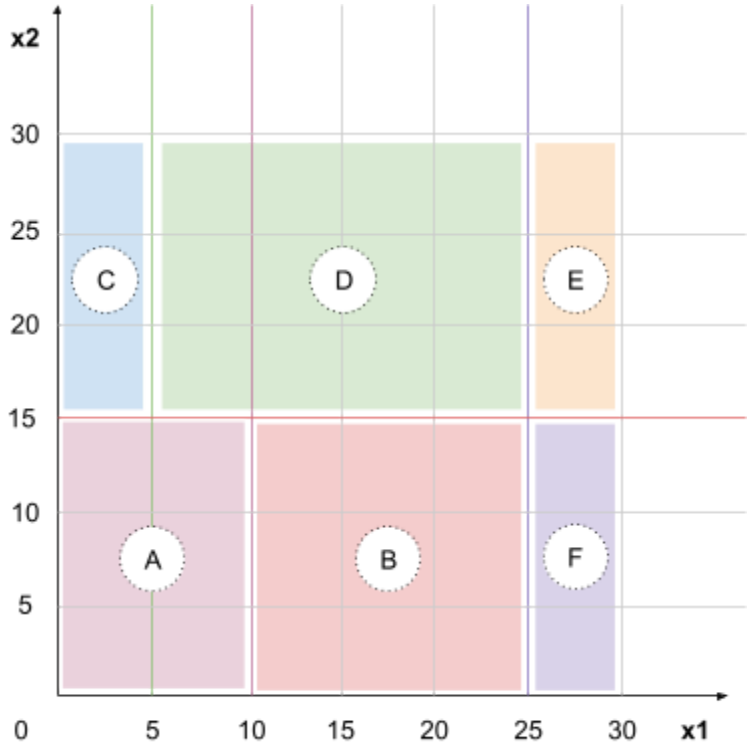
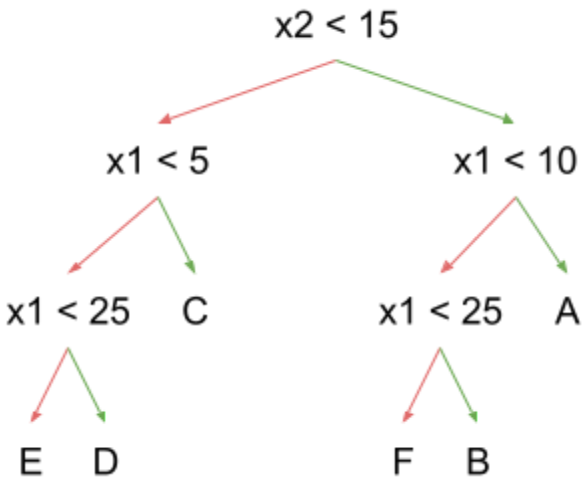
CS434 - HW 4

Decision Trees and Ensembles

1. Drawing Decision Tree Predictions
    a. Made using Google Drawings:



    b. Made using Google Drawings (assumes the interval as the previous problem):



    c. In this case, the redundancy allows for the learning algorithm to reach the accurate tree in multiple different ways. This means that the algorithm will not be restricted to one tree, rather a family of trees, all of which are correct.
2. Manually Learning a Decision Tree

Calculate Information Gain:

$$H(Y) = -\sum_{i=1}^{k} P(Y = y_i) log_2 P(Y = y_i) = -\frac{3}{6} log_2 \frac{3}{6} - \frac{3}{6} log_2 \frac{3}{6} = \frac{1}{2} + \frac{1}{2} = 1$$

$$IG(A) = H(Y) - H(Y|A)$$
$$H(Y|A) = -\frac{3}{6}\left(\frac{2}{3} log_2 \frac{2}{3} + \frac{1}{3} log_2 \frac{1}{3}\right) - \frac{3}{6}\left(\frac{1}{3} log_2 \frac{1}{3} + \frac{2}{3} log_2 \frac{2}{3}\right) = 0.918296$$

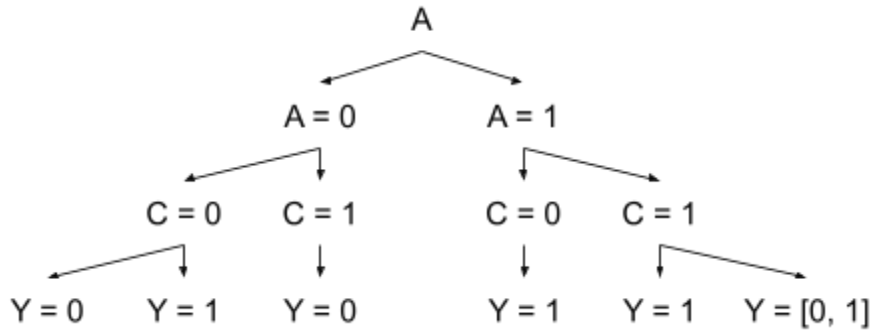
$$IG(B) = H(Y) - H(Y|B)$$
$$H(Y|B) = -\frac{2}{6}\left(\frac{1}{2} log_2 \frac{1}{2} + \frac{1}{2} log_2 \frac{1}{2}\right) - \frac{4}{6}\left(\frac{2}{4} log_2 \frac{2}{4} + \frac{2}{4} log_2 \frac{2}{4}\right) = 1$$

$$IG(C) = H(Y) - H(Y|C)$$
$$H(Y|A) = -\frac{3}{6}\left(\frac{2}{3} log_2 \frac{2}{3} + \frac{1}{3} log_2 \frac{1}{3}\right) - \frac{3}{6}\left(\frac{1}{3} log_2 \frac{1}{3} + \frac{2}{3} log_2 \frac{2}{3}\right) = 0.918296$$

$$IG(A) = 1 - 0.918296 = 0.081704$$
$$IG(B) = 1 - 1 = 0$$
$$IG(C) = 1 - 0.918296 = 0.081704$$

Both A or C will work for the attribute split as they both have the highest information gain, and those reduce uncertainty by the most.

Tree made using Google Drawings:

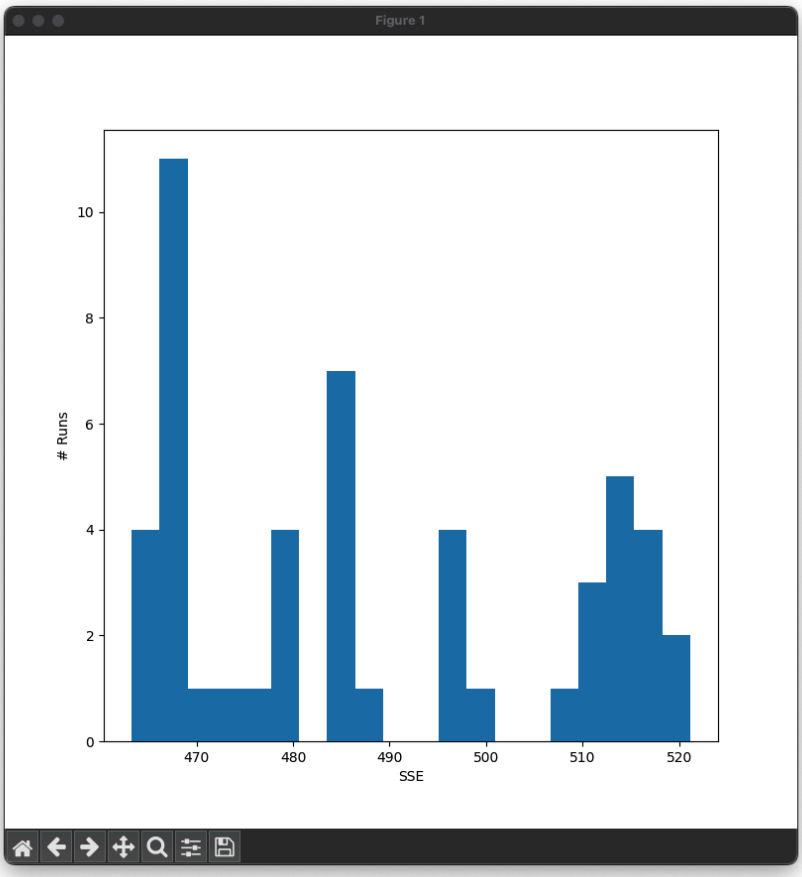


3. Measuring Correlation in Random Forests
   ???

k-Means Clustering

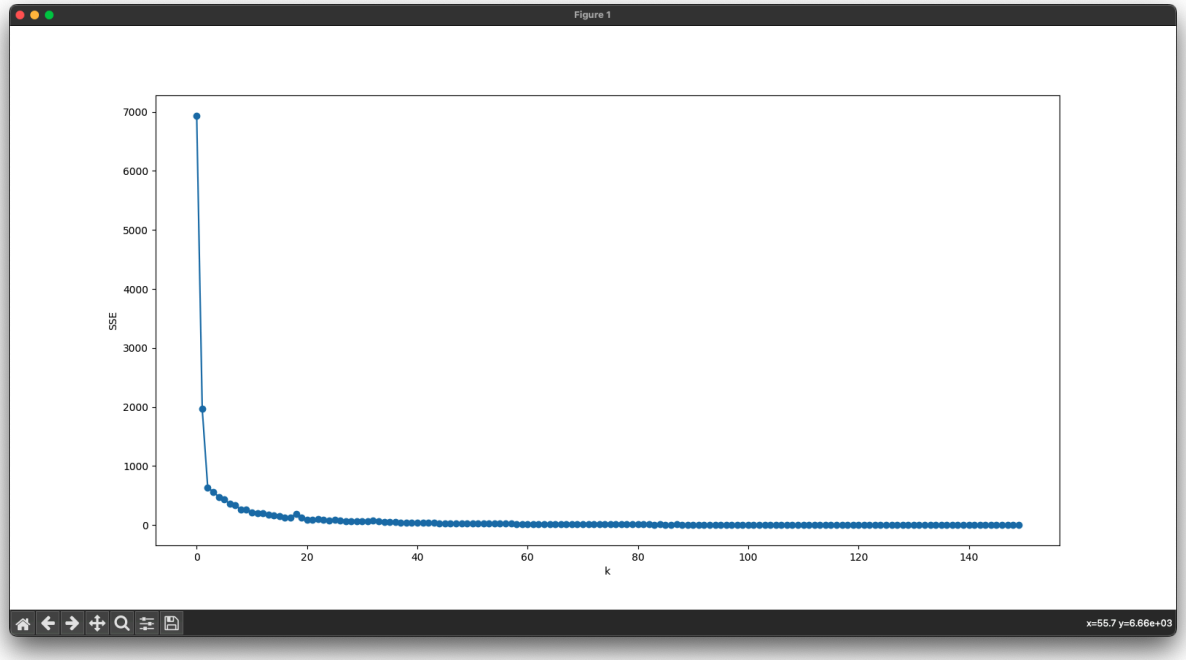
4. Implement k-Means
   See code

5. Randomness in Clustering



Above, you can see the resulting plot that I got after running the algorithm fifty times with five clusters. For a large majority of the time, runs resulted in an SSE of around 470, with smaller peaks around 485 and 515. For a real world dataset, if the initialization of the centroids are of poor quality, it could lead to poor clustering, which would hurt downstream tasks.

6. Error Vs K



Above you can see the plot that results from running with k = 1 through 150. While the plot looks good -- the SSE getting closer and closer to 0 -- this is actually not a good thing. The more and more centroids that you add, the more that they will match up with specific points, rather than groups of points. This will lead to overfitting your data and making it useless in the real world.

7. Clustering Images
    a. Looking at what images are in the dataset, it seems that images of nature (trees), buildings, highways, and a panda waving hello. Images were only returned for four of the nine clusters, leading me to believe that the k value is not very well set. As there are only really three different images, they should be placed in three clusters (unless there are more pandas). This might end up restricting the algorithm too much, meaning that it will more easily make mistakes.
    b. Below is k = 25. Some clusters were empty, so they are not included.

Skyscrapers: 49/50 = 98%	Open Roads: 49/50 = 98%

	
Skyscrapers: 17/19 = 89.47%	Forests: 41/50 = 82%
	
Skyscrapers: 49/50 = 98%	Skyscrapers: 32/32 = 100%
	
Skyscrapers: 36/38 = 97.74%	Open Roads: 44/50 = 88%
	

Forests: 49/50 = 98%	Open Roads: 28/29 = 96.55%
	
Skyscrapers: 13/13 = 100%	
	

c. The SSE for *k = 25 is 2355.3757355034377* compared to the SSE for *k = 10 being 2492.8149326129387*. I think in some regards, this is a good indicator of clustering quality, but it can't be the only indicator used, unless you will end up overfitting the training data causing it to be useless.

8. Evaluating Clustering as Classification
   See above for labels and accuracy. Overall, the accuracy wasn't too bad. Not good enough for image recognition, but it is able to accurately group certain pictures. When a picture included multiple different labels, such as a tree taking up half of the frame with a building taking up the other half, it was prone to mistakes.

Debriefing

1. How many hours were spent on this assignment?
   ~5 hours
2. Rating?
   6/10
3. Alone or with others?
   Alone.
4. Understanding?
   Pretty good understanding.

5. Comments
   No.