Project 4: Decision Trees and k-Means Clustering

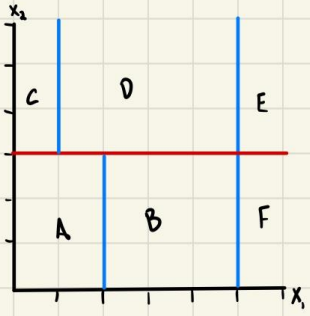
CS 434 -- Spring Quarter 2021

Connor W. Baldes

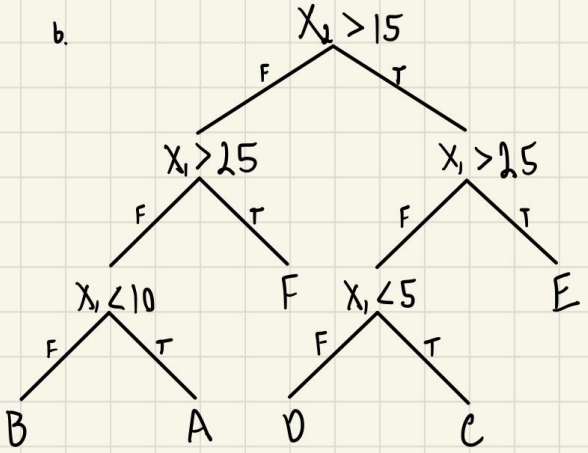
baldesc@oregonstate.edu

1.)

a.

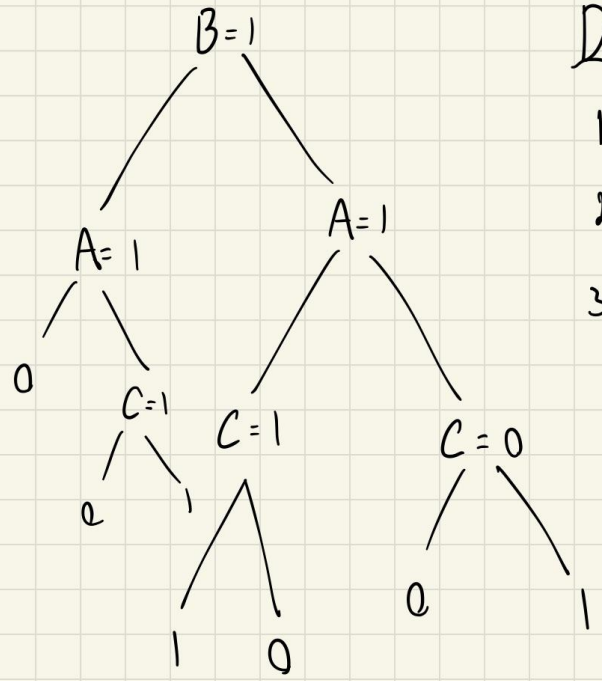


b.



C. It should make it easier to find an accurate tree because no one tree is right many different trees can yield accurrate results

2.)



Debrief:

1. 10
2. easy
3. alone
4. 80.1%

3.

Figure 1

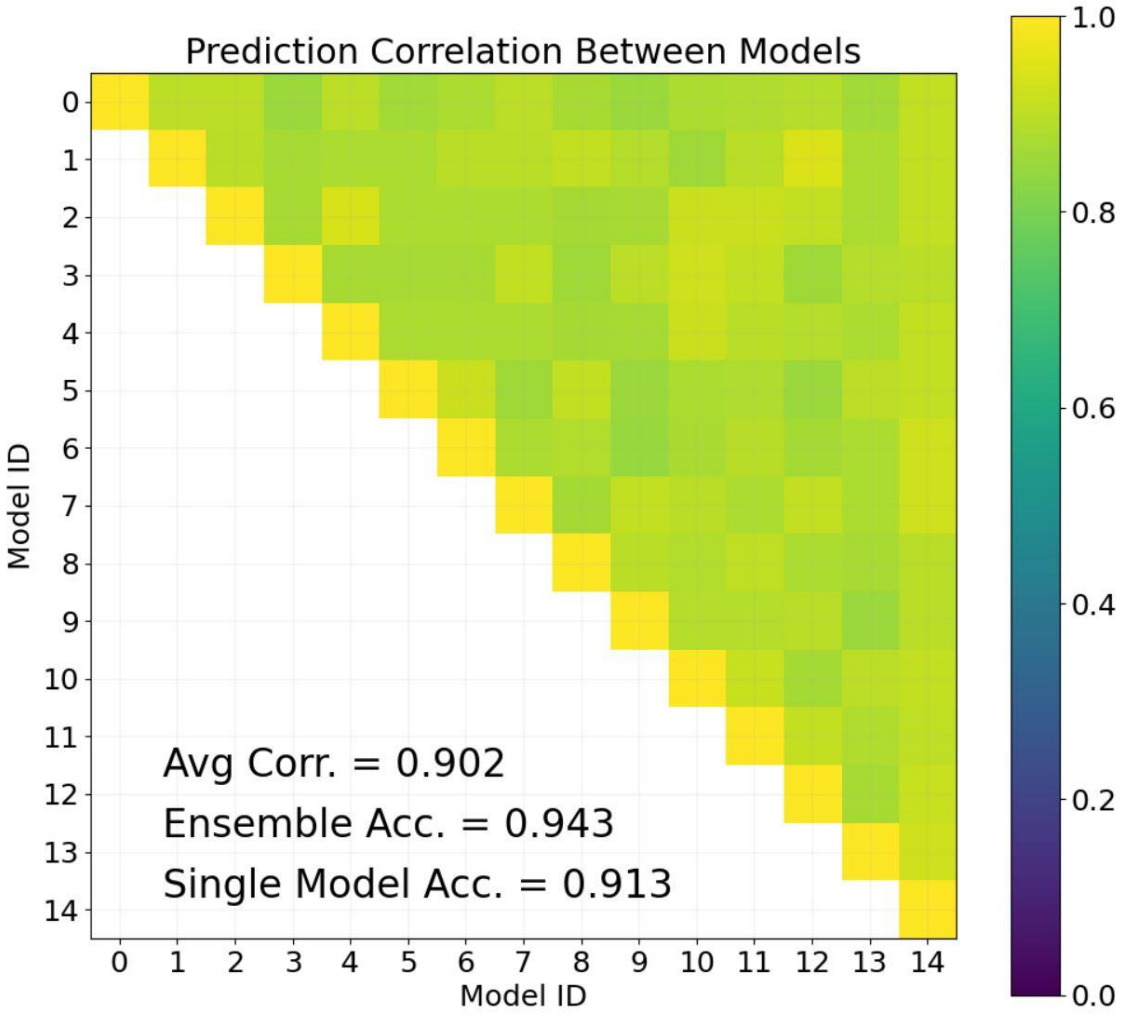
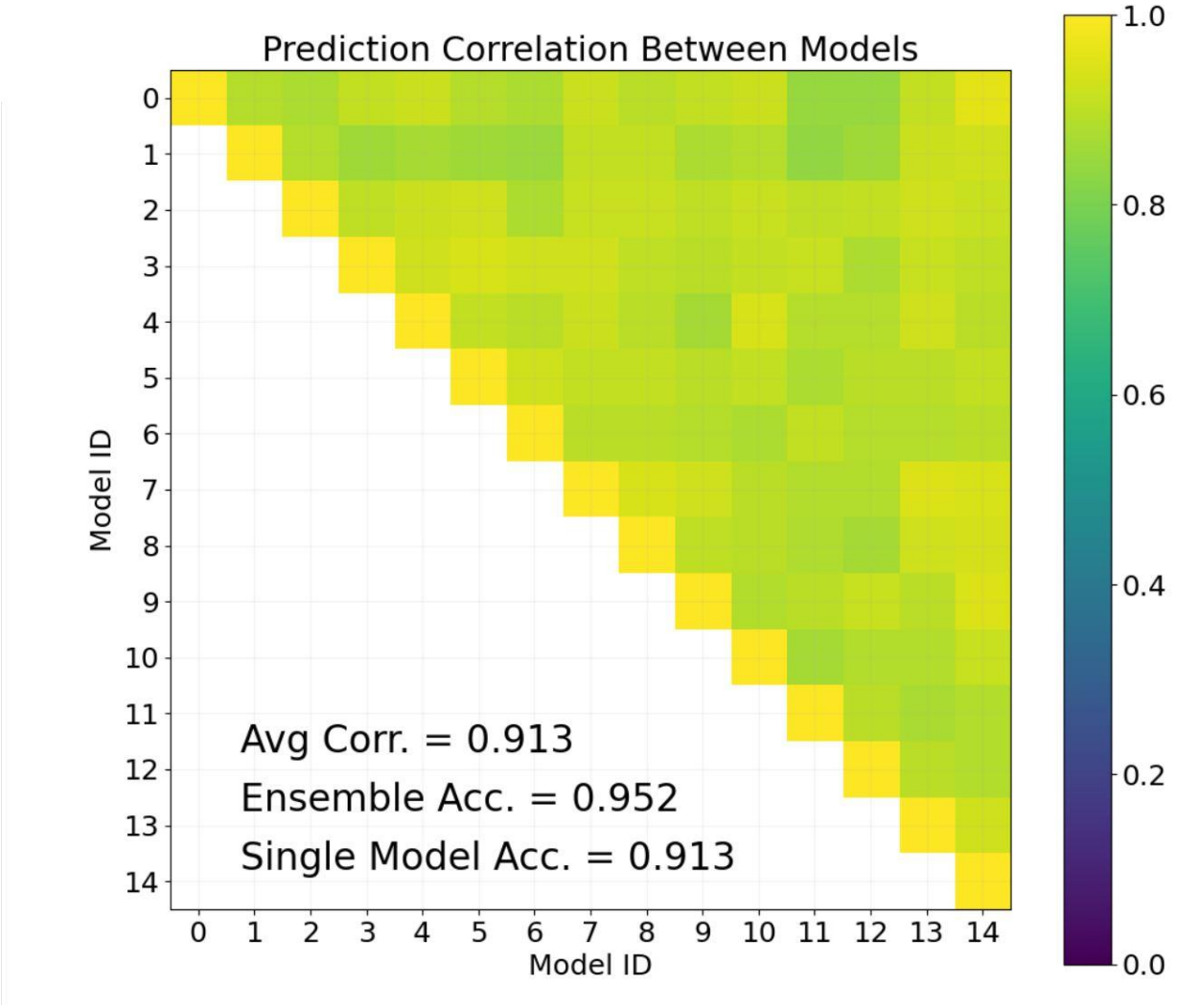
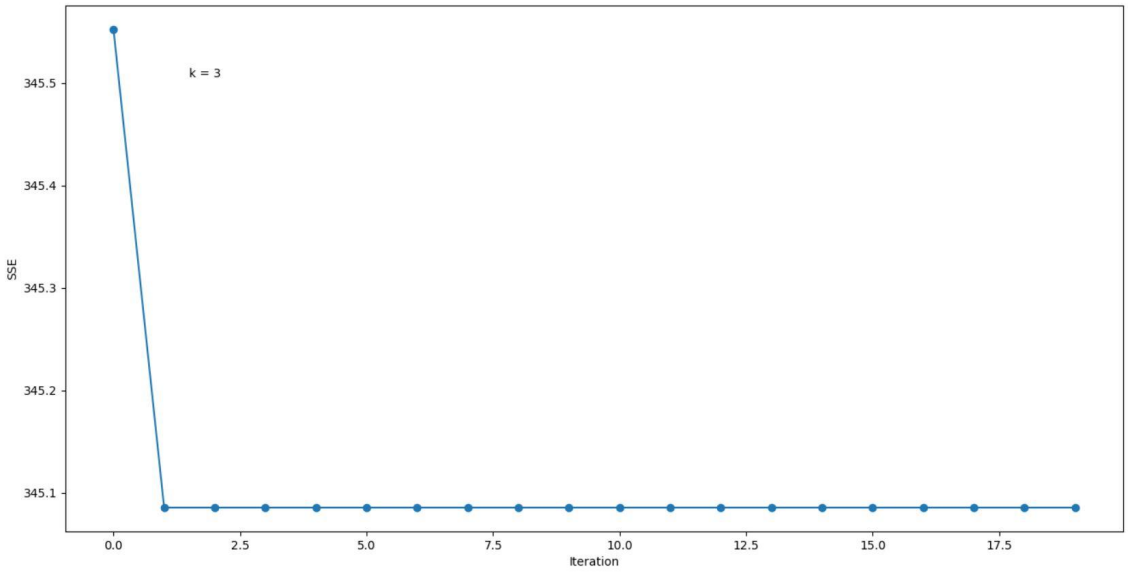

Prediction Correlation Between Models
Model ID
Model ID
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Avg Corr. = 0.902
Ensemble Acc. = 0.943
Single Model Acc. = 0.913
1.0
0.8
0.6
0.4
0.2
0.0

Figure 1

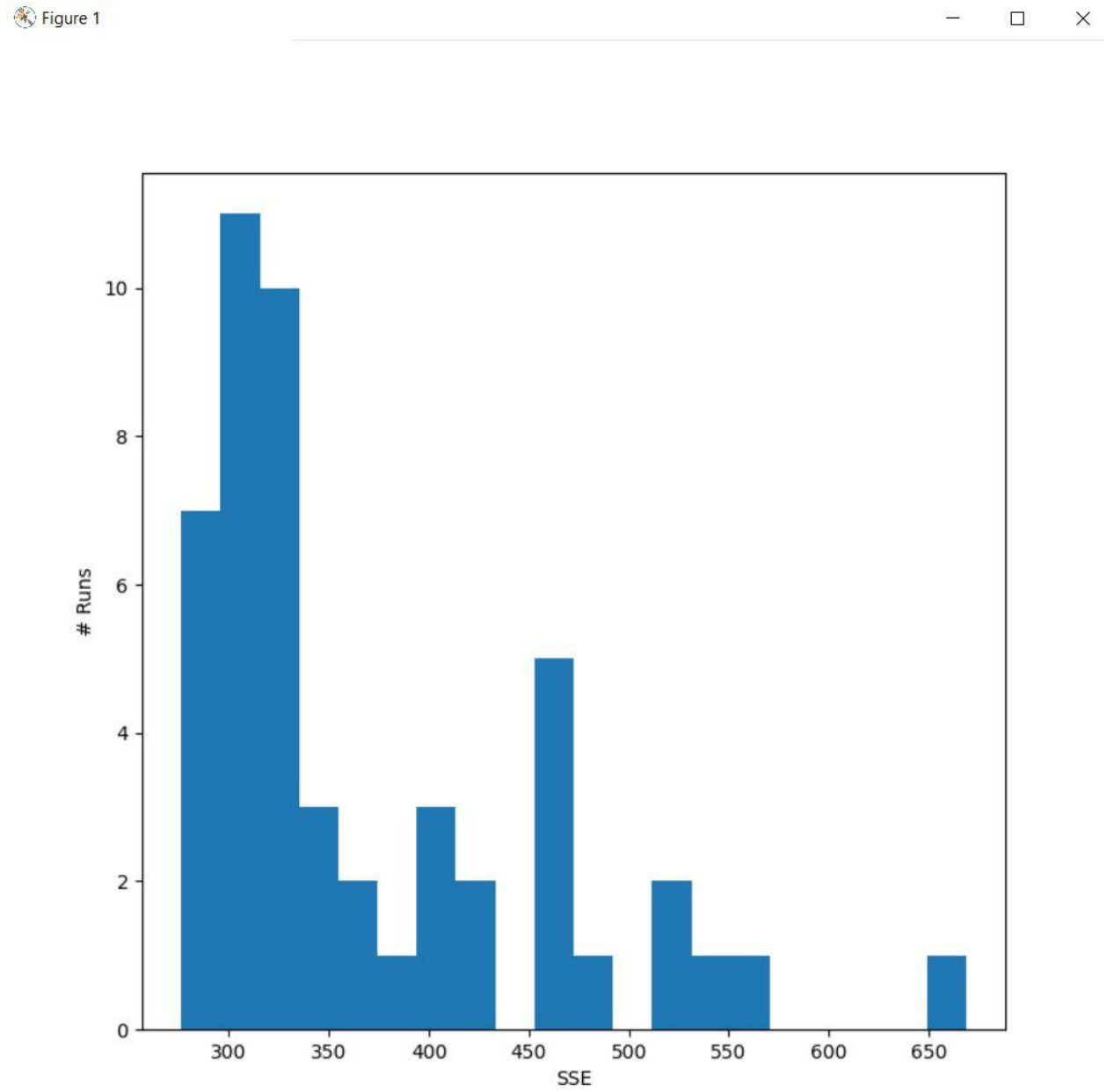

Prediction Correlation Between Models
Model ID
Model ID
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Avg Corr. = 0.913
Ensemble Acc. = 0.952
Single Model Acc. = 0.913
1.0
0.8
0.6
0.4
0.2
0.0

4.

Figure 1

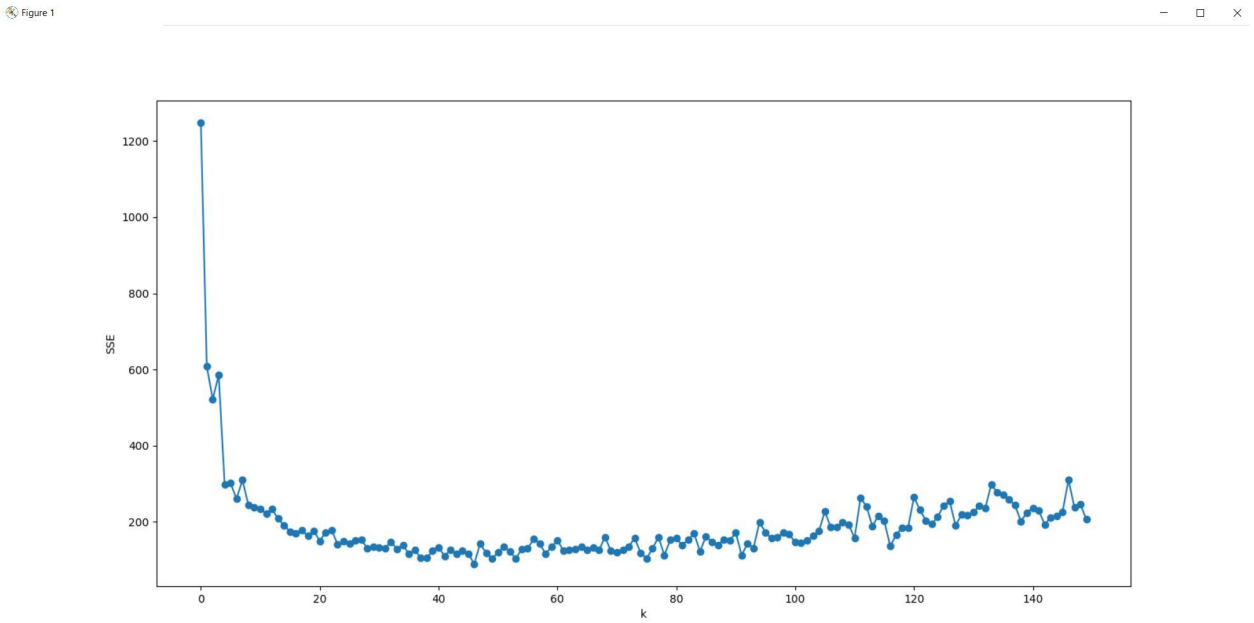

k = 3
345.5
345.4
345.3
345.2
345.1
SSE
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Iteration

5.



It is clear that on average the iterations with a higher number of runs had a lower SSE, this tells me that when I apply k-means to real data I should run it multiple times because that will help decrease the SSE caused by where the centroids are initialized.

6.



My plot shows that as k increases the SSE decreases, I believe there must have been some issue in my implementation that caused the SSE to stop decreasing after k was larger than 50 but in general I can still see what the trend was. Choosing k based on SSE doesn't make sense because SSE in k-means clustering is based on each point in a cluster's distance to that cluster's centroid and therefore as k gets larger there are less points in each cluster making the clusters smaller and closer together reducing SSE. Looking at the resulting graphed clusters, however, shows as k gets larger the distinction between clusters becomes less and less, meaning that model really isn't predicting actual clusters in the data and is therefore useless.

7. A. I believe that k is fine as it is. I tried values lower than 10 but the clusters I got had more images that were not similar to each other so I decided to stick with 10.

B.

C. For this I tried comparing k=25 and k=10 as before SSE was smaller when k=25 but the clusters I got split groups of images that should have been grouped together and therefore lowered the overall quality of the clusters.

8. A. - Trees/nature
- Plain roads
- Skyscrapers
- Rodes with other stuff around them
- Unique buildings

- More roads
- More roads
- Multiple buildings
- Empty roads
- Multiple buildings/nature

B. - 36/40
- 29/32
- 45/45
- 35/38
- 30/50
- 21/27
- 6/10
- 35/50
- 25/26
- 43/50