Daniel Nisnevich {207447970}

5.2:

k	l	accuracy
1	l1	0.967044
1	l2	0.966711
10	l1	0.961718
10	l2	0.957723
100	l1	0.923103
100	l2	0.920107
1000	l1	0.745007
1000	l2	0.741678
3000	l1	0.401798
3000	l2	0.398136

1.We can note that there is a slight difference (10^{-3}) when we change the distance function.

However, when changing the K to a number over 10 , for our specific data the accuracy drop significantly.

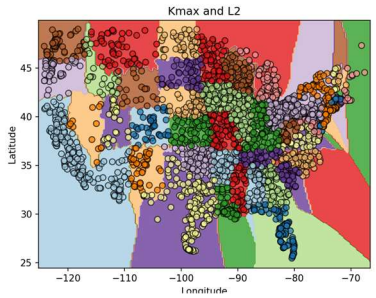
Moreover, in the web I saw that to avoid ties we want the 'K' to be odd but looking at our specific data I see an opposite result :

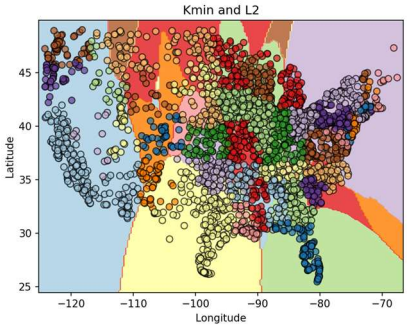
```
   k   l  accuracy
0  2  l1  0.958056
1  2  l2  0.960719
```

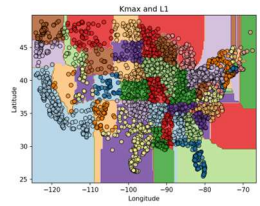
```
    k   l  accuracy
0  11  l2  0.961052
1  11  l1  0.965379
```

k	l	accuracy
1	l1	0.967044
1	l2	0.966711
10	l1	0.961718
10	l2	0.957723

$$K_{max} = 0.966711, K_{min} = 0.398136$$



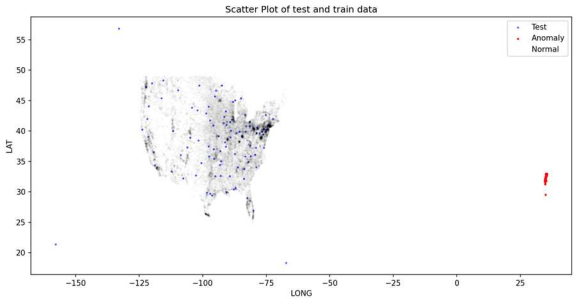




a) The difference between (i) and (ii) is that (i) differentiates between data sets in a better way. We can see significantly more dividing of the space, $K_{\max}$ has better results because (ii) tens to show more overfitting behavior and preforms poorly on data that is new like we can see in the up- right corner.

   $K_{\max}$ has better generalization to unseen data because it considers a smaller amount of neighbors. We can see that the choice of 3000 neighbors looks a lot more random and the distribution is inconsistent.

b) The choice of distance metric didn't change the outcome of our model. I can't deduce that the change of distance function doesn't matter because it might have more impact on different data during training or different data type.



5.4

The anomalies seem to differ significantly from the data set. Looks like an error or a someone tampered with the data and added coordinates in Tel Aviv to a dataset of coordinates in the U.S. This experiment makes me dwell whether Israel is the 51th state.

6.1

Looking at the results, the highest accuracy score is of the tree with Max_Depth of 20 and Max leaf nodes of 1000.

It looks like when the depth is too deep there seems to be overfitting as expected but for some reason the data set is probably big enough that though the delta of accuracy is of 10^{-3}.

2.

{'max_depth': 20, 'max_leaf_nodes': 1000}	1.0000	0.9732	0.9742

The tree shows excellent performance in all the three checkups, it is sufficient to capture the full complexity of the data and generalize for given test data .It is without a doubt the best tree..

3.

No, 50 leaf nodes are not enough to capture the complexity of the data and be able to handle new data.The perfect accuracy was not achieved in any of the trees I can deduce that we can have overlapping of decisions for the same state and it deeply depends on our

data.As we can see with the smaller trees the it does depend on the max depth too but the accuracy is significantly lower by 10^{-1} which means for every 10th coordinate our model will be wrong and that is with more than 50 leafs. We can see in the images below the difference.

4. Decision trees see the space as divided to rectangles which represent a decision boundary.As deep the tree goes it allows for another deviation and allows the subgroup the data more accurately. When visualizing the data it can be seen as a summation of many different rectangles that should somewhat capture the different states in the US (if the tree is deep enough to do so). There is a straight correlation between depth of the tree and the dividing of the rectangles and the classification of the data.

6.5

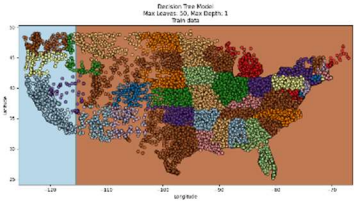
I would expect it to preform worse but the accuracy tends to be higher than 80% which is quite ok. I think that it highly depends on the data we are training on and the problem we are trying to solve. Still if we like a model that works well we would choose the model from Q1.

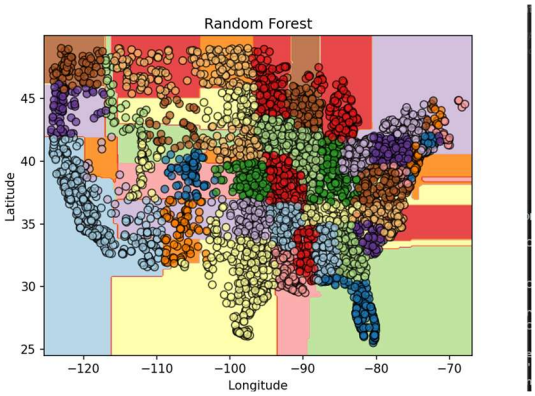
6.

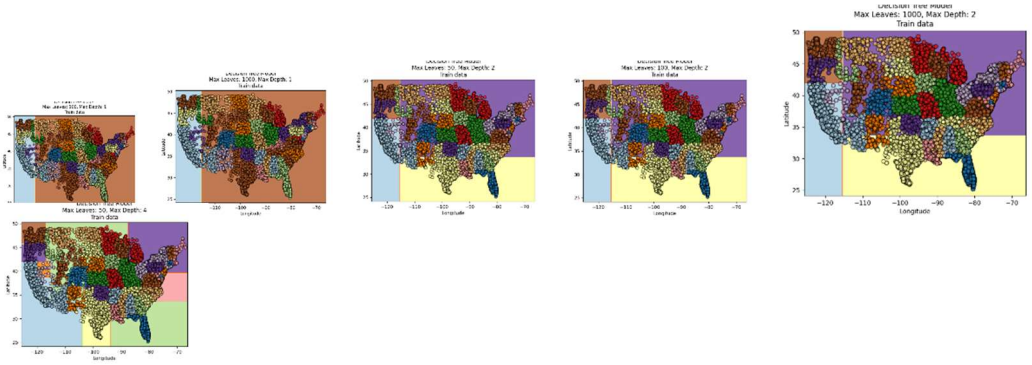
Here we can see that the accuracy drops very highly and looks to stay in the region of 50%.

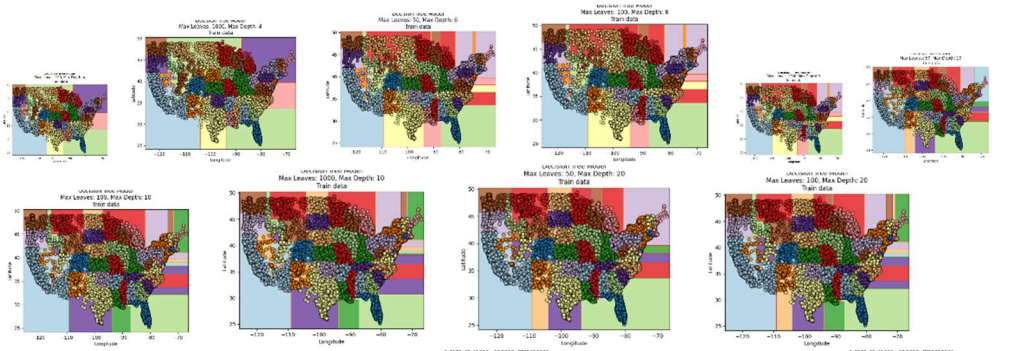
The decision trees split less and they cluster different labels as the same. Looks like the depth matters a lot to the accuracy and the max leaves affect it less(when talking about improving from some point of course when we have a small amount of leaves it still drops the accuracy).

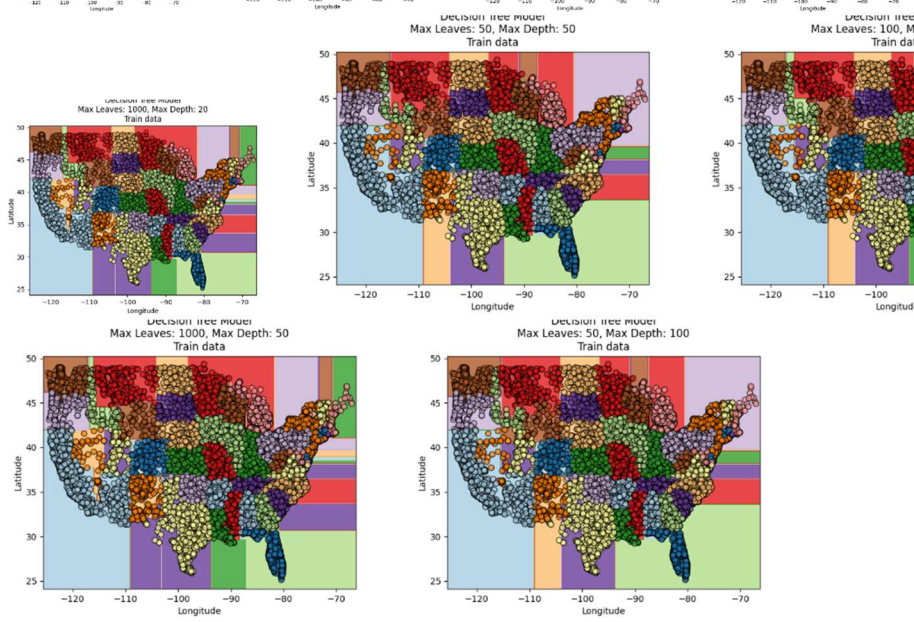
7.The random forest implementation doesn't look as good as the implementation inQ1, we can see a drop in accuracy it is more expressive than the option of the same trees with depth 6 so it does improve the quality of tree for the given depth. We can observe that it did a well enough job to get the to higher than 80% accuracy.

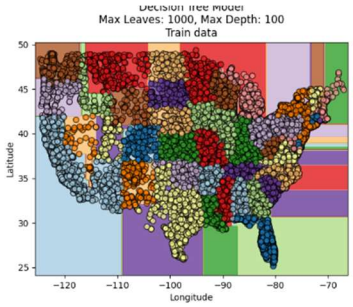


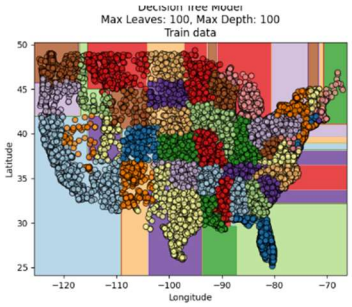


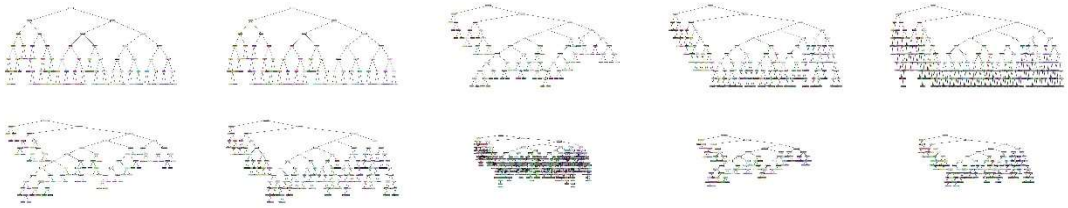





Max Leaves: 50, Max Depth: 50
Train data
Max Leaves: 100, Max Depth: 50
Train data
Max Leaves: 1000, Max Depth: 20
Train data
Max Leaves: 1000, Max Depth: 50
Train data
Max Leaves: 50, Max Depth: 100
Train data
Latitude
Longitude


Decision Tree Model
Max Leaves: 1000, Max Depth: 100
Train data
Latitude
Longitude
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70


Decision Tree Model
Max Leaves: 100, Max Depth: 100
Train data
Latitude
Longitude
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70



Random tree: