Itamar Epstein 203253145

Machine Learning Methods

Exercise 2

5.1 Task

Question 1 (k ∈ {1, 10, 100, 1000, 3000} and distance metrics in {L1,L2}):

	1	10	100	1000	3000
l1	0.967044	0.961718	0.923103	0.745007	0.401798
l2	0.966378	0.957723	0.920107	0.741678	0.398136

5.2 Questions:

Question 1: Based on the 5 × 2 table of results from task 3:

What is the trend you see when the number of k increases?

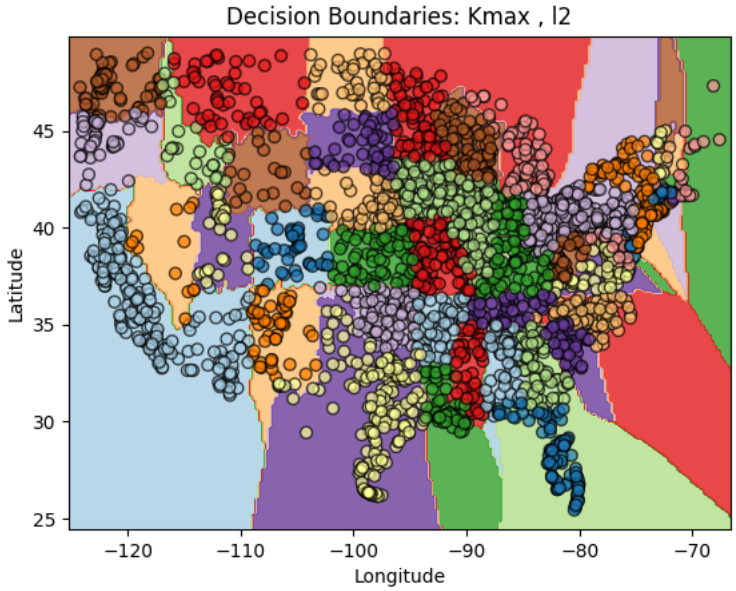
The trend observed is a decrease in test accuracy as the number of neighbors (k) increases. As k grows from 1 to 3000, there is a consistent decline in accuracy for both L1 and L2 distance metrics.
Explanation: The decrease in accuracy with increasing k suggests that considering a larger number of neighbors leads to a less generalized model, potentially overfitting to the training data.

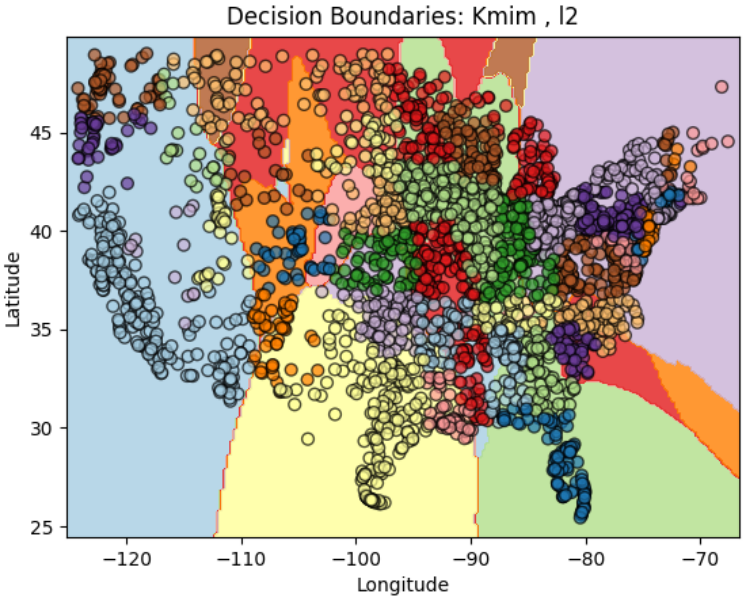
Does it changes between different distance metrics?

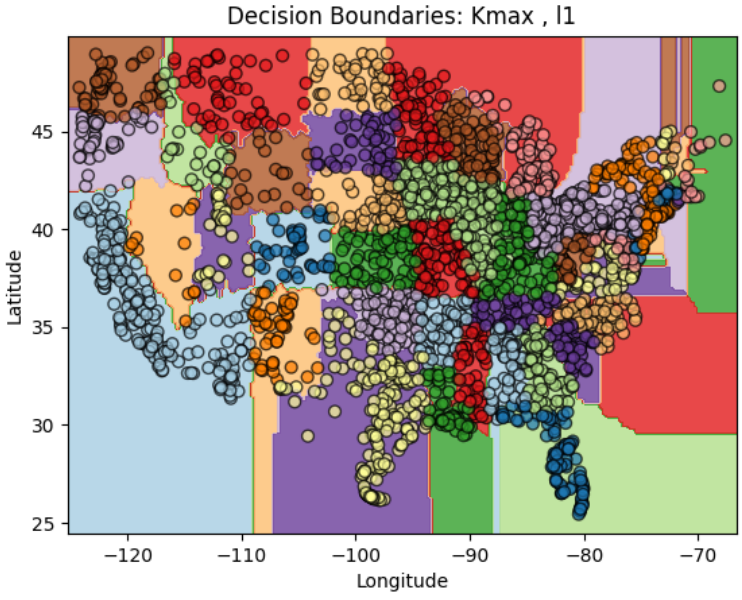
Yes, there is a notable difference in performance between the L1 and L2 distance metrics. the L1 distance metric has better results then the L2 metric for all values of k.
Explanation : The difference in performance between L1 and L2 suggests that the choice of distance metric has an impact on the model's ability to classify cities based on geographical coordinates , Indicating the need of a good distance function.

Question 2:

visualize the differences between k values and distance metrics:







(a) **:** Based on the plots of the (i) { kmax with L2 } and (ii){ kmin with L2 }:

What is different between the way each one divides the space?

The main difference is that the decision boundaries formed by kmax with L2 are highly accurate and have many detailed divisions. However in decision boundaries formed by kmin with L2 show intricate patterns but also have significant gaps of color.

Explanation : The decision boundaries in plot (i) are more accurate and detailed because the model considers only one nearest neighbor (small k) It captures better the differences between data points, leading to many detailed divisions. As well high accuracy is shown because the model is classifying the data points based on their immediate neighbors. And the second plot may overfit to irrelevant data, resulting in big gaps in decision boundaries. It becomes less sensitive to local patterns, potentially indicating insensitivity to small, localized data variations.

Why does kmax results in better accuracy?

Kmax results in better accuracy because a small k value (1 in this case) allows the model to be highly sensitive to local patterns and variations. The decision boundaries is because the model is classifying the data points on its immediate neighbors. In contrast Kmin (k= 3000) leads to less sensitivity to local patterns causing a decrease in overall accuracy.

(b) **:** Based on the plots of the at the kmax with L2 distance metric and kmax with L1 distance metric.

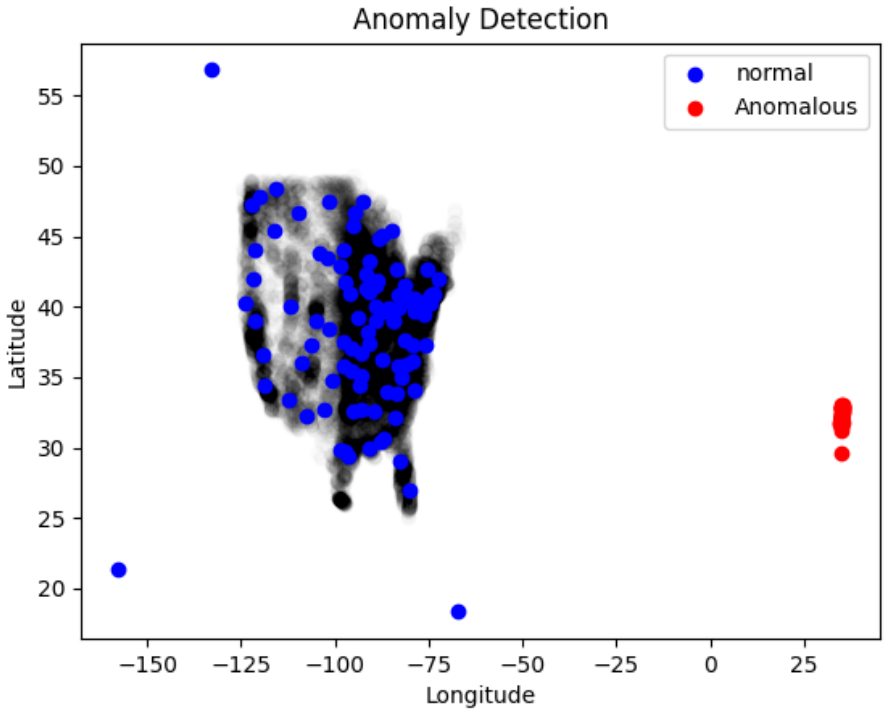
How does the choice of distance metric affect the classification space?

The main difference is that the decision boundaries formed by kmax with L2 are More [1]smooth decision boundaries. However in decision boundaries formed by kmax with L1 show Less smooth boundaries, more flicks, but still very accurate.

Explanation : The choice of distance metric affects the classification space in terms of the smoothness of decision boundaries. L2 distance metric results in smoother boundaries indicating that it is less sensitive to increasing differences of the latitude and longitude in out data set, while L1 distance metric produces less smooth boundaries with more abrupt changes. however, both achieve very high accuracy in classifying the data points. The difference in the use of L2 or L1 distance metric depends desired classification space.

[1]smooth = more flexible and allows fewer straight borders

5.3 Plot Anomaly Detection Using kNN:



5.4 Question Anomaly Detection Using kNN:

Question 1:

What can you tell about the anomalies your model found? How are they different from the normal data? Explain.

The main ideas and diffrenedses that can be taken from the anomalies the model found to the normal data is:

- The anomalies data has a significantly different longitudinal value compared to all the train data and normal data.
- The normal data is shattered across a broader geographic area, representing the typical distribution of the training set. While the anomalies data is concentrated in a specific, smaller geographic area compared to the normal data

  Explanation : due to the significantly different longitudinal values, that have a direct influence on the distinct features. Causing the anomaly scores (Sum of the 5 distances to the nearest neighbors) to stand out from the normal data . In addition the concentrated and isolated geographic distribution makes the model refer all this data points together as the same anomalies.

6 Decision Trees

6.2 Questions:

Question 1 +2 :

Choose the tree with the best validation accuracy. What is its test accuracy? Report the training accuracy, and test accuracy of the tree from Q1

Tree(max_depth=100, max_leaf_nodes=1000, accuracy_training=1.0, accuracy_validation=0.9806924101198402, accuracy_test=0.9770306258322237)

Is this tree successful in generalizing from its training examples?

Yes, Altho the training accuracy is 1.0, indicating that the tree correctly classifies all the training examples. While perfect training accuracy could sometimes indicate overfitting. But in this case The validation accuracy and the test accuracy are high (approximately 98.04% and 97.84%), suggesting that the model performs well on data it hasn't seen during training. And more so This indicates successful generalization to new, unseen examples and the ability to generalize to an independent dataset not used during training or validation.

Look at the test accuracies of the other trees. Is our validation set sufficient to choose the best tree? Explain your answers.

Yes , because based on the the other trees test accuracies. The training accuracy values span over a large range, suggesting diversity in model performance. Notting that not every model did good on the validation set. And more so the large range indicates that the validation set provides a meaningful evaluation of the model's performance because all the models were trained on the same training set.
However, the validation accuracy for the best-performing tree is relatively high, indicating that this particular tree generalizes well to unseen data, and the model chosen tree is likely a good candidate.

Question 3:

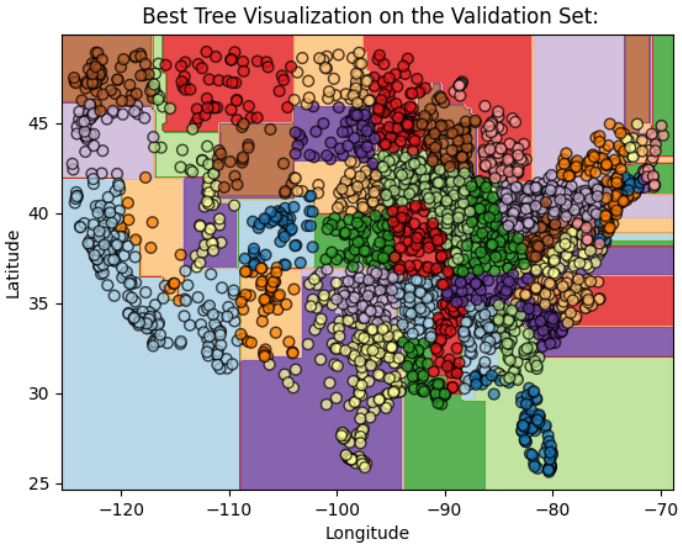
Are 50 nodes enough? There are 50 states in the USA, and a leaf node predicts a single state each time. Are 50 leaf nodes enough to achieve perfect accuracy?

No , The results of the experiment demonstrate that having 50 nodes is not sufficient to achieve perfect accuracy in predicting the states based on the longitude and latitude features. The complexity of the geographical distribution of U.S. states and the fact that there are 50 states, indicate that a decision tree with only 50 leaves cannot achieve perfect accuracy. This is because the decision tree draws a straight line on every decision and after 50 nodes (decision) there will be a case that the lines came to a rectangle and the state borders are not perfect lines that are vertical to the axis of longitude-latitude.
In summary, more then 50 leaf nodes in a decision tree are necessary to better capture the complex and non-linear relationships present in the geographical coordinates of U.S. states, allowing for improved accuracy and generalization of the complex axis of longitude-latitude of the border.

Question 4 - How trees see the world:

What is the shape it creates for each class?

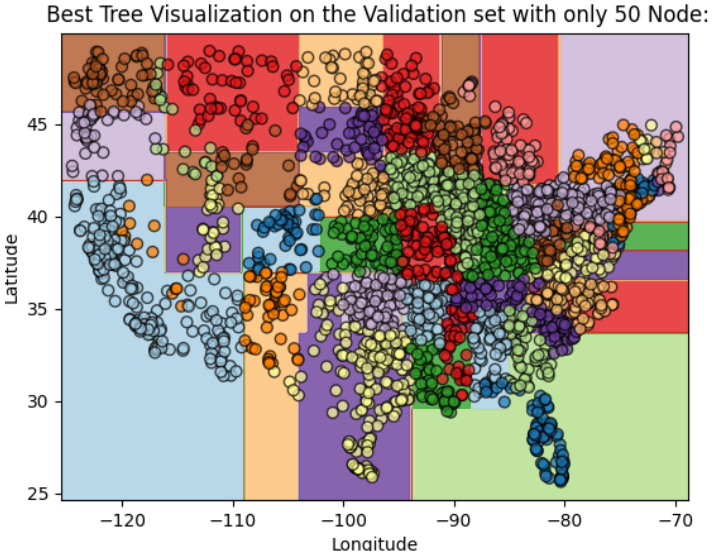
The shape for each class is a rectangle (square) shape , indicate that casing proses spoken about in question 3.



Question 5 - Restricted leaves:

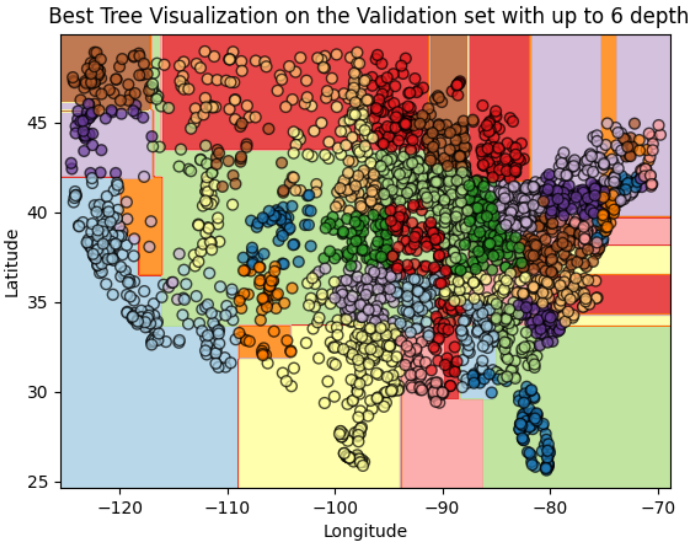
How does the prediction of the 50 compare to the overall best from Q1?

This tree, when compared to the tree from Q1, has simpler square shapes of borders of the states, due to its fewer decision nodes. The reduced number nodes in decision-making limits its accuracy as it fails to separate different grouped data.



Question 6 - Restricted depth:

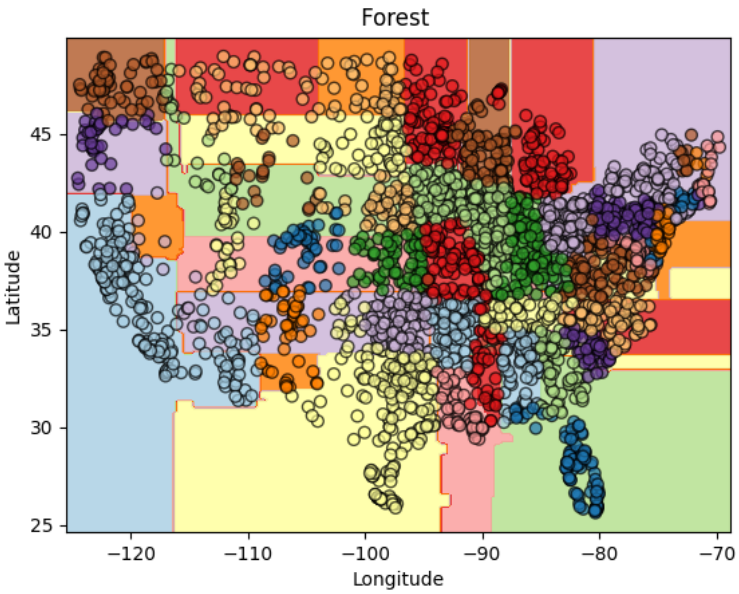
What has changed in the way the space is divided compared to Q4? Explain.



This tree, when compared to the tree from Q4, the lines of the class that indicates the borders of the states, are larger spaces in some cases the in Q4. And in addition a lot of cases that the lines are correct , the color is wrong indicates a wrong class label. This two changes are a straight cause of the “Restricted depth” because it means the tree can have less decision nodes, that makes the model be more specific

Question 7 - Random Forest:

Is this model more expressive than the one from Q1? How this visualization helped reach that conclusion? Explain.

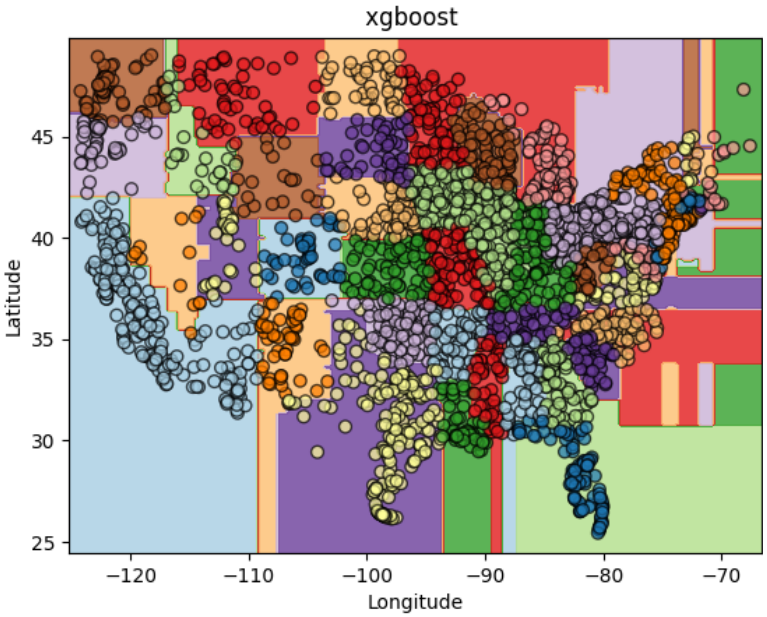


Comparing the visual plot of the tree in Q1 with the visualization of the forest, it can be seen that the tree from Q1 is more expressive. Because the borders in this plot are more of rectangular shapes, indicating a smaller number of classes for the model. This outcome is due because all the trees in the forest have a maximal depth parameter. “Restricted depth”

In conclusion despite generating 300 different trees within these limitations, the forest's combined expressive capacity does not match a model with a significantly higher maximal depth. The difference in the visualization of the two questions contribute to reaching the conclusion that a model with its higher maximal depth, is more expressive than then a forest generated under the specified limitations.

Question 8. Experimenting with XGBoost (Bonus 5 pts).

How are the predictions of XGBoost different from the ones of the random forest? Which algorithm is more successful in this task? Explain.



xgboost test accuracy
0.9653794940079894

The predictions generated by the XGBoost forest have a higher accuracy compared to the predictions of the random forest. In the visualization it can be seen that the XGBoost model is capable of classifying more precise and well-fitted rectangular areas. The high accuracy performance of the XGBoost model is because of its way of choosing each tree in the forest for particular types of data, particularly in this case the longitude and latitude borders of U.S. states.

The XGBoost algorithm achieves better results due to its unique approach compared to the random forest. In XGBoost, each tree is created by specializes in a distinct area of the U.S. map, aligning with the specific geographical features of the dataset. However, the random forest generates trees without explicitly dividing the data, leading to less specialized trees and ultimately resulting in lower accuracy.