**3 Approach**

**3.2 Decision Tree**

A decision tree is an important approach to perform predictive analysis to help with making decisions. The biggest advantage of decision tree is it can represent the process and results visually and explicitly. It is a graphically method that makes use of branching methodology to exemplify all possible outcomes based on certain conditions.

The setup of decision tree needs to convert our prediction as binary. In this model, we define microplastic content per litter less than 2/L as 0 (no/few microplastic contamination) and more than 2/L as 1 (more microplastic contamination).

The next step is to choose appropriate feature to construct predictive decision tree. In the microplastic prediction model, we built our decision trees based on following four features: sampling date, sampling latitude, sampling longitude and the water temperature around sampling area. We calculated the entropy before/after split and the information gain for each split. The feature with the best gain of information is chosen as prioritized feature to perform split. The prevent the decision tree overgrow, we need to set up stopping criterial initially. In the microplastic prediction model, we discontinued datasets split when less than 5% of data was unused. After the decision tree is built, we performed 10fold cross-validation to test the accuracy of our model.

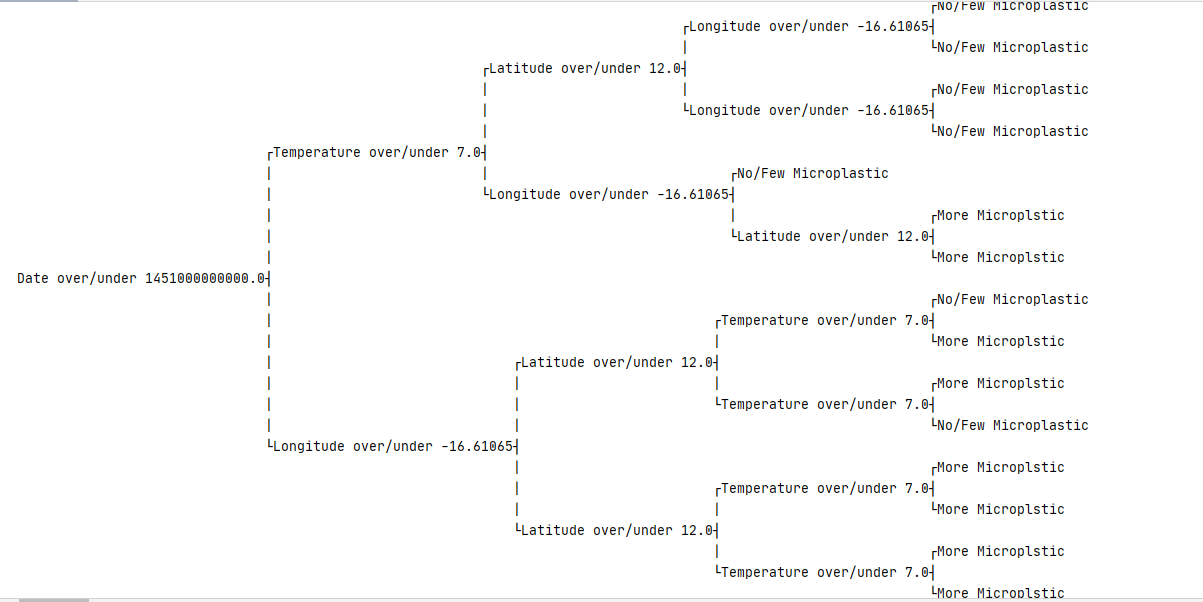
To further expand our decision tree model, we build our random forest based on two different strategies. One is to use a random subset with 80% of the samples to build each tree, and the other is to exclude one feature for each tree. When the random forest is generated, a majority vote method will be applied for our classification prediction.

**4 Results**

**4.2 Decision Tree For Microplastic Prediction**

**4.2.1 Microplastic Decision Tree**

We generate our decision tree with the best information gain method as described in our approach. The code we used is: tree = initial\_tree('Marine\_Clean.csv', printTree = True). We can see under different conditions (date, temperature, latitude or longitude), we would get different microplastic content. This decision tree model can be used to facilitate our prediction.



**Figure 4.2.1 Decision Tree to predict microplastic.** The tree was built based on Date, Latitude, Longitude and Temperature. The No/Few microplastic is defined as less than 2/L and more microplastic is defined as more or equal to 2/L in that area.

**4.2.2 K fold Cross-Validation of Decision Tree**

In our case, we performed 10 fold cross-validation to verify our decision tree model. The accuracy based on 10-fold verification is 77.03 percent. The reason of the model is not very accurate might be due to our early termination criteria, which is essential to trim the tree and prevent the over grow of our tree.

**4.2.3 Microplastic Random Forest Model**

As we described, we can generate our random forest based on two different approached, either by 80% of data or by excluding one feature. We can do predictions of the existence of microplastic based on above random forest.