Water Contamination Coding Project Write-Up

DS 325, Section B

Quinn O’Connor & Charlie Kiddoo

Introduction/Abstract

Canada has more lakes and bodies of water within its borders than any other country, which unfortunately makes it quite vulnerable to aquatic pollution and water contamination. These bodies of water include rivers, lakes, coastal areas, and estuaries (called transitional in the dataset), which are all affected by runoff. Because each water body type gets affected differently, we thought it would be interesting to train our model to predict water body type based on the chemical composition and containment levels of a given water sample. For our dataset, we decided it was best to utilize a decision tree classifier with a one-hot encoder, an ordinal encoder, a label encoder, and a standard scaler. We chose a decision tree classifier as our dataset did not include linear relationships, and we could grade its accuracy via a confusion matrix. Our findings illustrate the potential of utilizing chemical datasets to screen aquatic environments and spearhead water management efforts in detecting water pollution in fragile ecosystems.

Methods

The dataset we used in our analysis provides information on the chemical composition, contaminants, temperature, water quality index, and waterbody type of water samples collected in Canada. Initially, we trained a decision tree model to predict the water quality index rating of each water sample based on the water’s chemical composition. After using this model, we thought it would be interesting to use a decision tree classifier to predict the waterbody type of each sample based on water quality index ratings and chemical composition.

For the preprocessing and cleaning of our data, we dropped features such as the date, area, and country, as we did not need them in our data analysis. For numerical features (such as water composition features), we used a standard scaler. We used a one-hot encoder for our water quality index ratings, using a column transformer to transform the data frame. We used a label

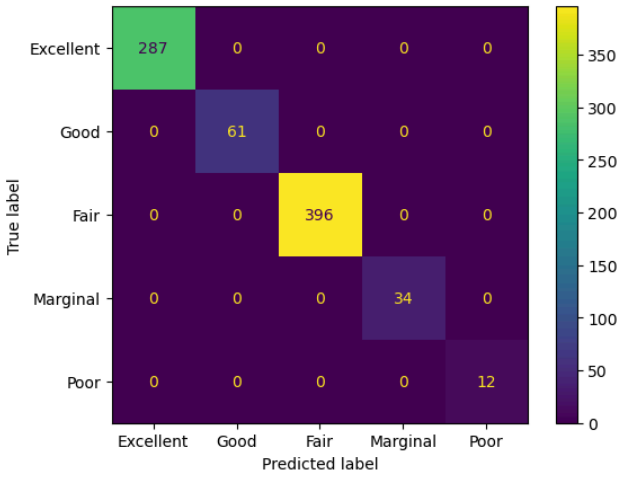
encoder for the waterbody type (our target data). We decided that a decision tree model would be a good choice for classifying our data.

After encoding our data, we split our data into training and testing sets, with a 20% test size. We then fit the decision tree model to the feature data in order to make predictions on the water quality index ratings of each sample. The decision tree then used the feature data to classify each sample as 'Excellent', 'Good', 'Fair', 'Marginal', or 'Poor'. We then changed our target data to the water body type and followed the same process in fitting a decision tree model, where each sample was classified as 'Coastal', 'Lake', 'River', or 'Transitional' bodies of water. To assess the accuracy of both decision tree models, we used confusion matrices to show whether samples were correctly or incorrectly classified.

Results

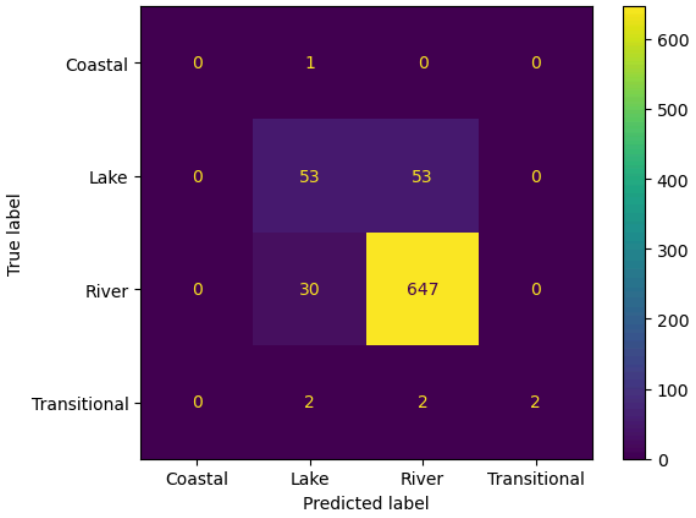
The performance of our decision tree was assessed by utilizing a confusion matrix, as shown in the figures below.

Water Quality Confusion Matrix



The confusion matrix demonstrates that our model allocated categories to certain water qualities perfectly, as each true label was predicted correctly with no misclassifications. Those categories are 'Excellent', 'Good', 'Fair', 'Marginal', and 'Poor', comprising a combined 790 data points. With an accuracy score of 100%, there were no instances of any false positives or false negatives. Meaning that our precision and recall scores are both 1.0, suggesting that the model has fully grasped the structure of the data and the ability to distinguish between water quality categories. As for our other confusion matrix, pictured below, it did not perform as well.

Water Body Type Confusion Matrix



We believe this to be a result of even data, where the vast majority of the data points were 'river' or 'lake, while only a combined seven data points made up both 'coastal' and 'transitional.'

Discussion

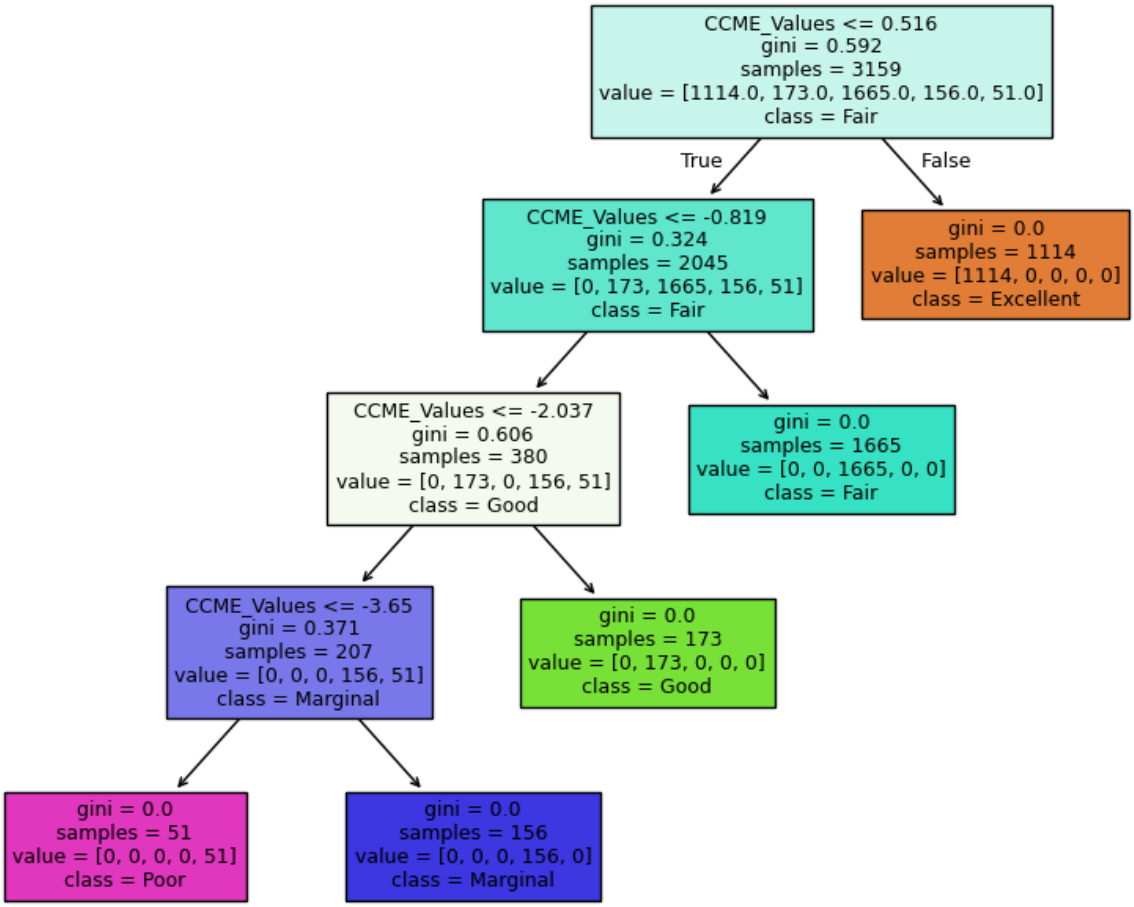
The decision tree model for water quality index ratings was 100% accurate and made no incorrect classifications. This is likely because the model relied mostly on the feature containing CCME values, which rate water quality levels on a scale of 0 to 100. Different ranges of CCME

values correspond with the actual classifications of water quality levels, which allowed our model to be 100% accurate. Perhaps it would be more interesting to use the decision tree without the CCME value feature, so the model could make more use of the other features.

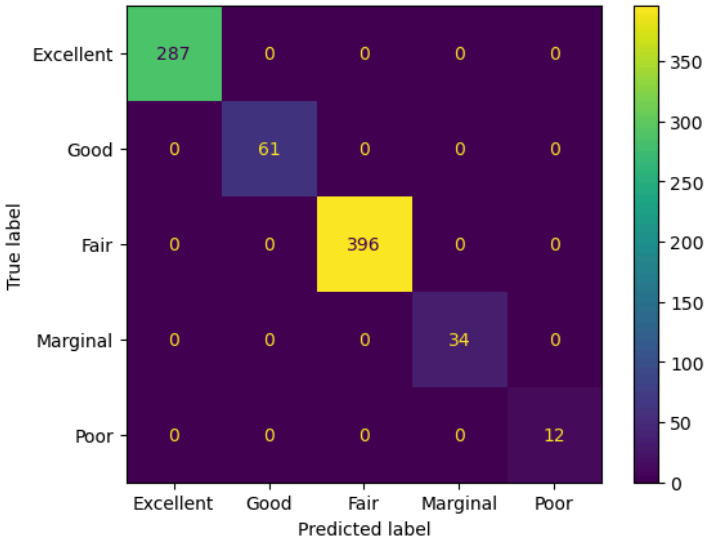
The results of the first decision tree gave us the idea to use a decision tree model to predict the water body type of each sample. This second decision tree did a very good job at correctly classifying rivers, however, there was some misclassification regarding the other water body types. This is likely due to a very uneven distribution of water body type data. The dataset had 697 river samples, 106 lake samples, 6 transitional samples, and only one coastal sample. If the dataset had more evenly distributed water body type samples, then perhaps the decision tree model would be better at making predictions with more diverse training data. Perhaps using a gradient boosted tree model would be more effective in making correct classifications. Overall, we believe that the model was mostly effective considering our very uneven distribution of water body type data.

In the future, it would be interesting to explore other features within this dataset and to use alternative models such as a linear regression model. Using machine learning to perform analyses on water quality data is very interesting and could be useful in a variety of contexts, such as water contamination, water conservation, or many other contexts important in today’s world.

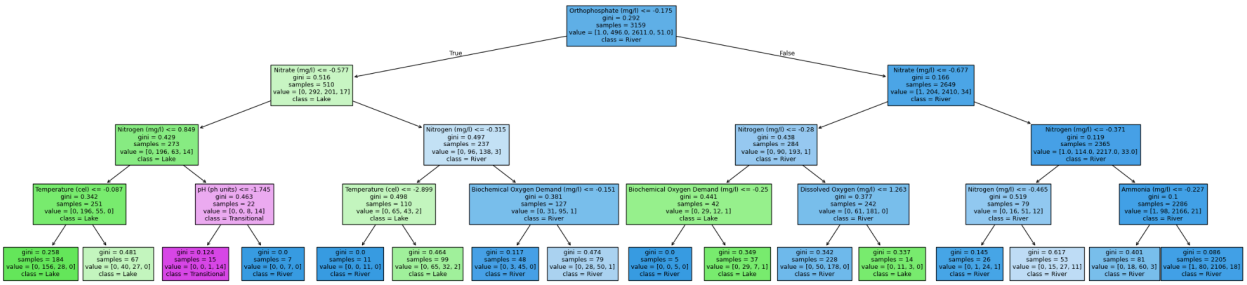
Water Quality Decision Tree



Water Quality Confusion Matrix



Water Body Type Decision Tree



Water Body Type Confusion Matrix