Valle, Cody J

Abstract

This paper outlines an approach to predict the temperature of the sea given air temperature, humidity, wind speeds, coordinates, and the date.

El Nino Dataset

What’s the temperature?

Introduction:

The dataset that I decided to use was called El Nino dataset. It contains information recorded from about 70 buoys spread out across the equatorial Pacific Ocean (see Figure 1). I attempted to classify the sea temperature of every row. I found that the sea temperature was not very related to the other data points in the dataset.

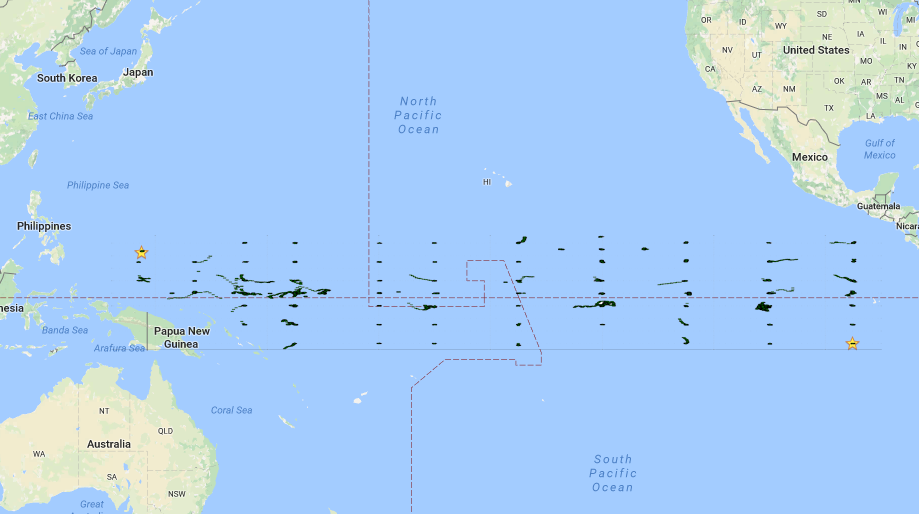


Figure 1: Real world locations of each instance from the data set

Data Analysis:

The attributes in the dataset are latitude and longitude coordinates, zonal and meridional wind speeds (east/west and north/south), air and sea temperatures, and the date the data was taken. The buoys have been taking this data since 1980, and the dataset has records up to 1998. There are over 175,000 rows in this dataset. All columns except the date are continuous. Some of the rows were missing data, and since there was so much, I decided to remove every row that was missing at least one data point. This left me with a little over 100,000 rows.

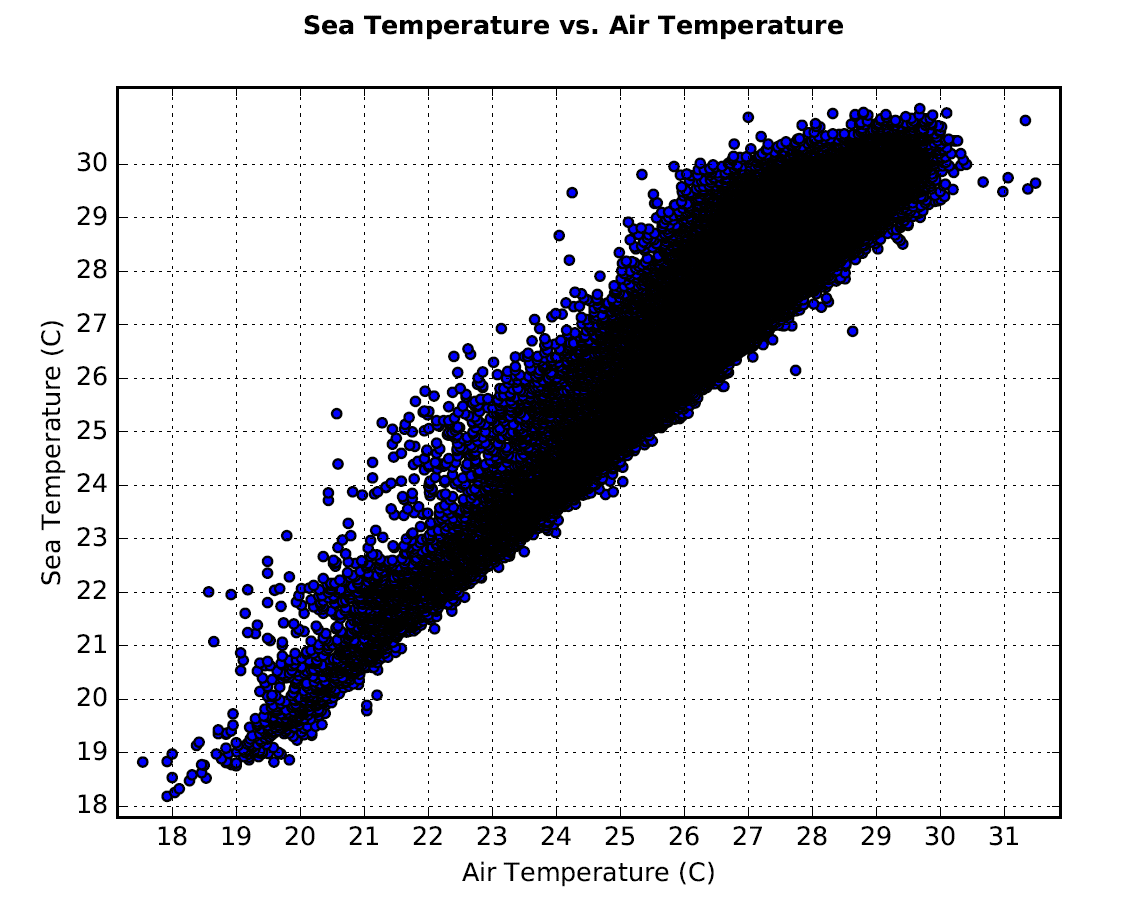
Since the data was continuous, some type of regression could be used. After some data exploration, I found that there were two possible options for regression classifiers. The sea temperature was a linear function of the air temperature (see Figure 2), and sea temperature was a sinusoidal function of the month (see Figure 3).

Figure 2: This shows the linear relationship between the sea temperature and the air temperature.

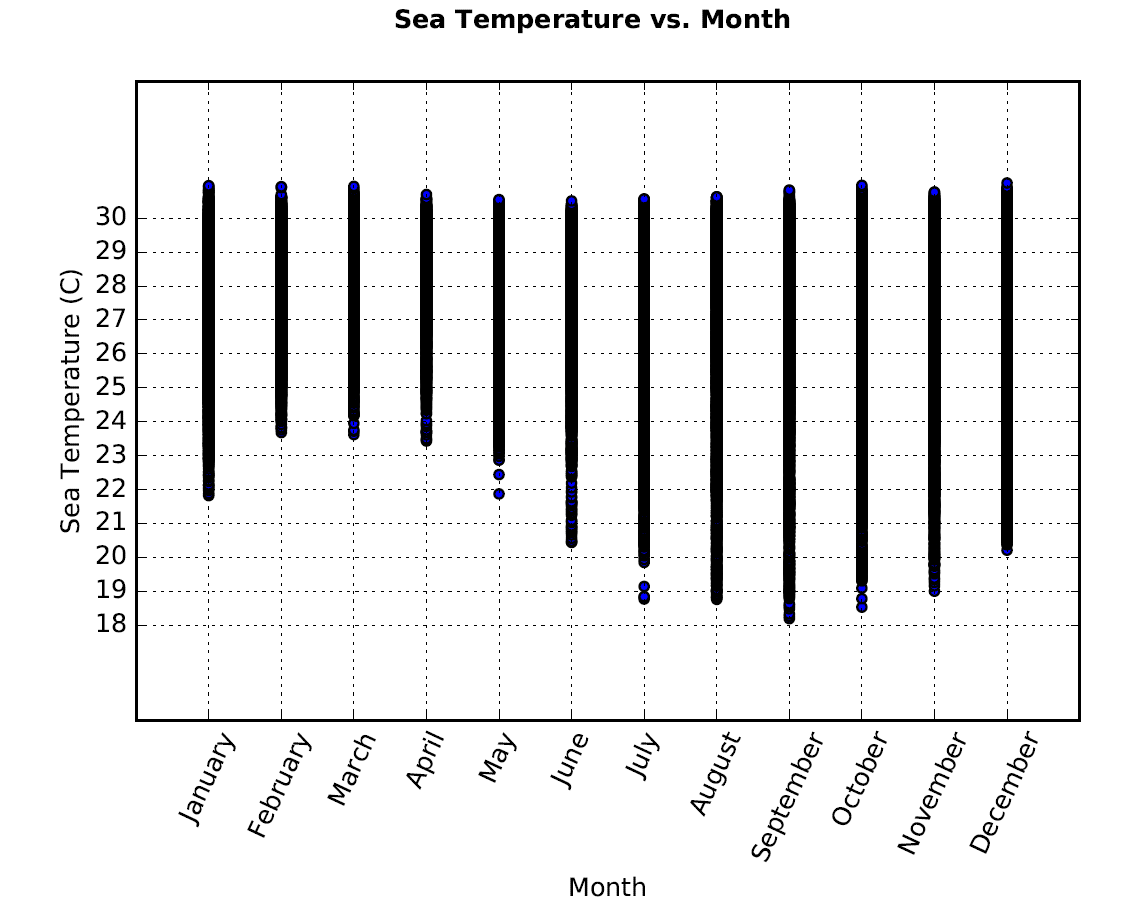


Figure 3: This shows that the average temperature rose and fell with the months.

Other types of classifiers required categorical data, so I decided to categorize the sea temperatures by rounding them to the nearest integer (or whole degree). This left me with 14 possible labels for the sea temperature.

In order to create a k-Nearest Neighbor classifier, I needed to normalize all of the rows. This was a simple number crunching task. I decided to use a k value of 5.

In order to create a decision tree and random forest, I needed to categorize all of the rows. I decided to use entropy to calculate the best split points on humidity. Calculating the entropy at every split point took a couple days, but it finally finished. With some threads, the time was reduced to 30 minutes. I decided to use 12 categories for the humidity, and hard coded the categorization. For the rest of the continuous rows, I decided to just round each value to the nearest integer.

Some interesting statistics about this dataset include that humidity was never below 52%, and was an average of 80%.

Classification Results:

The Linear Regression classifier that used the air temperature was accurate 47% of the time. Testing it was pretty straightforward. Create the equation based on the data, then use it to predict the sea temperature.

The k-Nearest Neighbor classifier was accurate 58% of the time. In order to test the classifier, I split the dataset into 10 stratified folds, and used each fold as a test set with the rest as the training set.

The Decision Tree classifier was accurate 12% of the time. A Random Forest classifier was accurate 30% of the time. To test the decision tree and random forest classifiers, I split the dataset into 10 folds and created a test and training set the same way as for the k-Nearest Neighbor classifier. I used 21, 120, and 3 for M, N, and F respectively. I did not test to find a best combination of M, N, and F values.

All of these results are found in Figure 4.

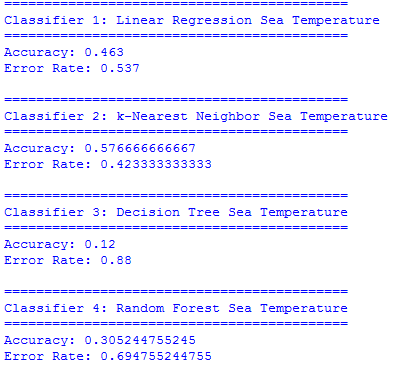


Figure 4: Results of each classifier.

Conclusion:

The sea temperature is not very predictable when given surrounding factors. The best classifier was k-Nearest Neighbors, but the accuracy of the Linear Regression model promotes the idea that the best determining factor is the air temperature. This does make sense, yet I though humidity might have more to do with it as well. The biggest challenge of working with this dataset was its size. It forced me to look at the most efficient ways to calculate a value or distance. I could have improved my classifiers by splitting the buoys into clusters. It’s very possible and highly likely that different locations had varying temperatures.