The dataset that I decided to use was called El Nino dataset. It contains information recorded from about 70 buoys spread out across the equator in the Pacific Ocean. This is a plot of the latitude/longitude pairs of every observation superimposed on the Pacific Ocean. The attributes in the dataset is latitude and longitude coordinates, zonal and meridional wind speeds (north/south and east/west), air and sea temperatures, humidity, and the date the data was taken. The buoys have been taking this data since 1989, but the dataset only has records up to 1998. There are over 175,000 rows in this dataset. So, I removed every row that was missing at least one point of data. This left me with a little over 100,000 complete rows of data. My goal was to build a classifier to predict the sea temperature given the rest of the attributes. Since the sea temperature was continuous, I decided to categorize it by rounding to the nearest whole number. This left me with 14 classes to predict. Next, I wanted to categorize the humidity attribute. I decided to use entropy to find the best split points. Performing the calculations on such a large dataset took a couple days, but it finally finished. I decided to use 12 categories, so I chose the most spread out split points with the highest information gain. For the zonal and meridional winds, I rounded them to the nearest integer, which gave me about 24 categories each. Then, I started plotting attributes against each other. I found that the sea temperature versus month had a sinusoidal relationship, as can be expected. Sea temperature versus year was linear, but with a slope of zero, so it was not helpful. Finally, sea temperature versus air temperature was linear and with a good slope. I used the non-categorized data to test a linear regression, and received an accuracy of 47%. After normalizing all rows, k-Nearest Neighbors was accurate 58% of the time. Using the categorized dataset, a decision tree was 12% accurate and a random forest was 30% accurate. So, of these four classifiers, my best was k-Nearest Neighbors. I measured its performance by splitting it into 10 stratified folds, and using each fold as test set against the rest as a training set. I believe the classifier could be improved if I had split the data into clusters, at most one for each buoy, since the weather near New Guinea probably doesn’t affect the weather near Mexico.