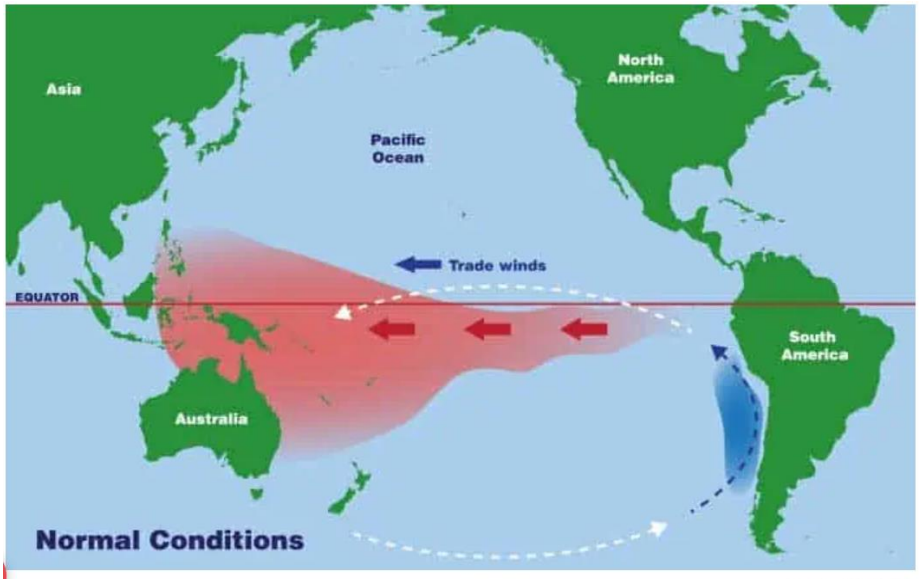
Data603 - Term Project

Factors that have greater effect on climate variations



Credit: NASA JPL

Project Report by,
Tejasri Pavuluri

Contents

1. Introduction

Motivation for the project includes an ongoing interest in utilizing scientific computing in both science, and industry. **El Nino** data analysis grabbed my attention because of the recent media on global warming; therefore the possible oceanic changes that might take place. This further strengthened my interest in climate data and to know the changes happening around different weather conditions. What factors cause climate variations? While analyzing these factors, I have also included a few additional questions that might strengthen my results.

2. Dataset description:

To analyze the factors that have an impact on climate change 'El nino' dataset is used. The dataset contains oceanographic and surface meteorological readings taken from a series of buoys positioned throughout the equatorial Pacific. The data was collected with the Tropical Atmosphere Ocean (TAO) array, which consists of nearly 70 moored buoys spanning the equatorial Pacific, measuring oceanographic and surface meteorological variables.

El Niño is an oscillation of the ocean-atmosphere system in the tropical Pacific which impacts weather patterns. It's also known as the El Niño-Southern Oscillation (ENSO) cycle.

The data consists of 178080 rows and 12 features taken from buoys as early as 1980-1990. The dataset can be downloaded from UCI Machine Learning Repository or Kaggle. For this project, the UCI Machine Learning Repository dataset was used, which consists of 2 downloads - tao-all2.data.gz(contains the data for 7 March 80 to 3 May 98) and elnino.gz(contains the data for 23 May 98 to 5 June 98). Below are the variable (or attributes) characteristics

1. **Observation** - number of observations in the dataset
2. **Year, Month, Day, Date** - these columns we can see the timeline of data that was collected at the same time of the day
3. **Latitude** - shows that the buoy moved around different locations. Latitude value stayed within a degree from approximate location
4. **Longitude** - longitude values were sometimes as far as five degrees off of the approximate location
5. **Zonal Winds** - winds fluctuate between -10 m/s and 10 m/s
6. **Meridional Winds** - winds fluctuated between -10 m/s and 10 m/s
7. **Humidity** - values in the tropical Pacific were typically between 70% and 90%
8. **Air Temperature** - temperature fluctuate between 20 and 30 degrees Celsius
9. **Sea Surface Temperature** - temperature fluctuate between 20 and 30 degrees Celsius

All the readings were taken at the same time of the day from the buoys. In this analysis, I used features to predict the target 'Sea Surface Temperature'. The small el nino dataset contains the buoy index and corresponding attributes from each buoy. Among the 59 buoys there is an average data count of 7 days per buoy, with a maximum of 14 days.

3. Details of dataset

3.1 Loading and Understanding dataset

After loading the dataset and looking at the information of the dataframe, we can see that all the columns have numerical continuous features and most of the columns have no NaN values (except for Zonal Winds, Meridional Winds and Humidity). A few of the columns have missing values and are denoted as periods(.) in this dataset. These were replaced as 'Nan' in the preprocessing step.

	Observation	Year	Month	Day	Date	Latitude	Longitude	Zonal Winds	Meridional Winds	Humidity	Air Temp	Sea Surface Temp
0	1	80	3	7	800307	-0.02	-109.46	-6.8	0.7	NaN	26.14	26.24
1	2	80	3	8	800308	-0.02	-109.46	-4.9	1.1	NaN	25.66	25.97
2	3	80	3	9	800309	-0.02	-109.46	-4.5	2.2	NaN	25.69	25.28
3	4	80	3	10	800310	-0.02	-109.46	-3.8	1.9	NaN	25.57	24.31
4	5	80	3	11	800311	-0.02	-109.46	-4.2	1.5	NaN	25.3	23.19
...	...	...	...	...	...	...	...	...	...	...	...	...
178075	178076	98	6	11	980611	8.96	-140.33	-5.1	-0.4	94.1	26.04	28.14
178076	178077	98	6	12	980612	8.96	-140.32	-4.3	-3.3	93.2	25.8	27.87
178077	178078	98	6	13	980613	8.95	-140.34	-6.1	-4.8	81.3	27.17	27.93
178078	178079	98	6	14	980614	8.96	-140.33	-4.9	-2.3	76.2	27.36	28.03
178079	178080	98	6	15	980615	8.95	-140.33	NaN	NaN	NaN	27.09	28.09

178080 rows × 12 columns

Below were few improvements done to the dataset:

- Naming features
- Changing missing values to NaN
- To compute statistics on data types, converted 'object' data type to 'Float64'

After preprocessing, the main numerical attributes were separated for further analysis. Below are the columns that are considered for analysis

```
['Latitude', 'Longitude', 'Zonal Winds', 'Meridional Winds',
'Humidity', 'Air Temp', 'Sea Surface Temp']
```

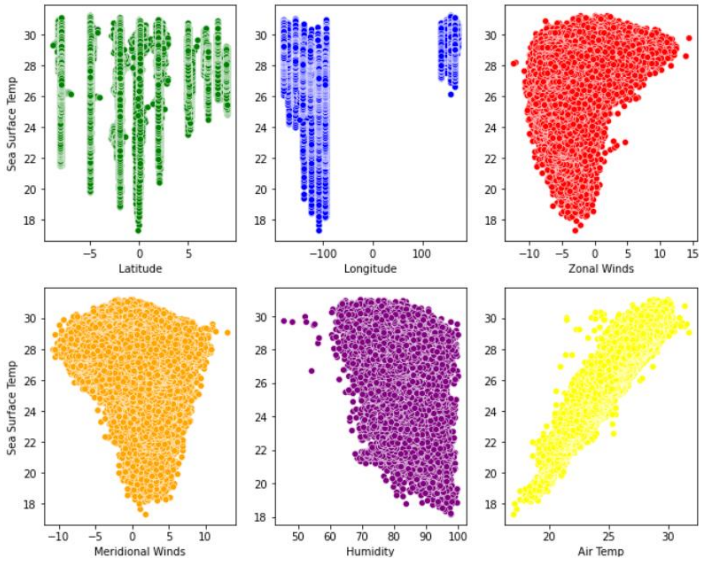
Machine Learning basic Libraries required to load and perform data analysis:

- pandas
- numpy
- matplotlib
- geopandas
- shapely.geometry
- seaborn

Note that, as the project analysis continues, libraries will be added as required.

3.2 Primary data analysis

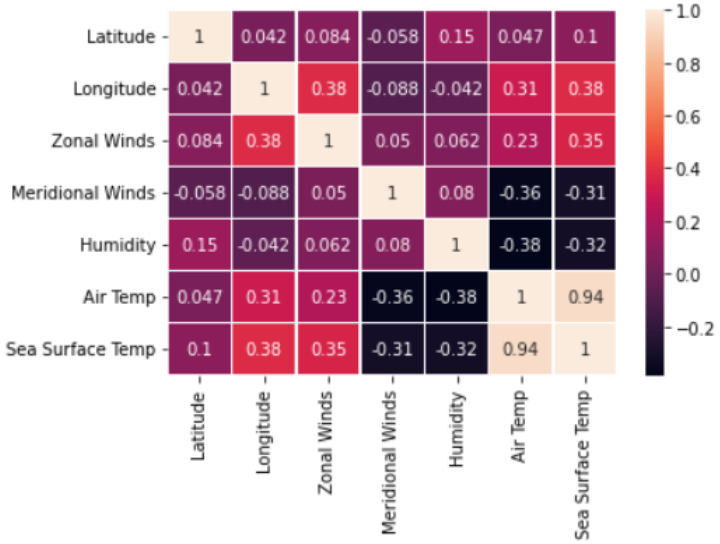
Initially, an investigation was conducted to understand the relationship between the features with Sea Surface Temperature. The best way to see the relationships between numerical features and the target at a glance is to draw scatter plots. Let's see the regular scatter plots to show the relationships of the numerical features with Sea Surface Temperature.



Few Research Questions that can be answered from the analysis are:

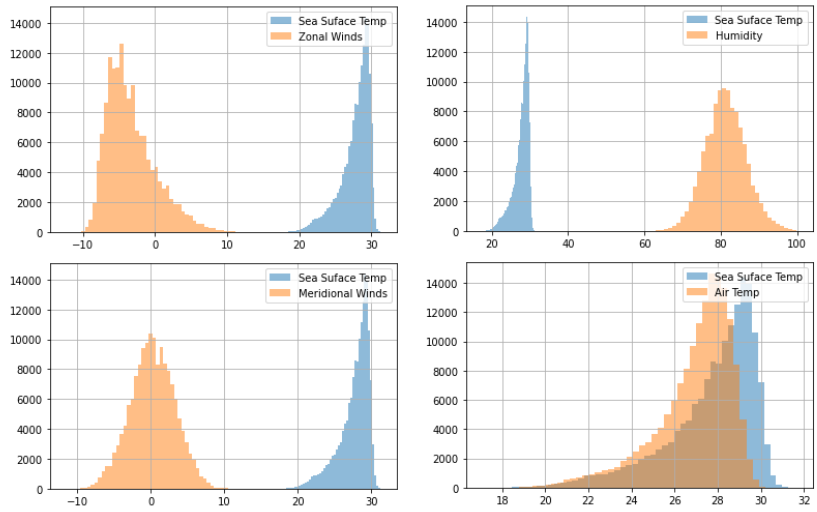
- How do the variables relate to each other?

In the plot above, we do not have an easy way of understanding the relationship between the features and the target. We can hardly see the relationships as there were many overlapped data points in the graph. To check those relationships in numbers, a correlation matrix was used among the numerical features and Sea Surface Temperature. From the correlation values, we can see that there is positive and strong correlation between Air Temperature and Sea Surface Temperature.



- Which variables have a greater effect on the climate variations?

Given that this dataset is already analyzed for normality, additional statistical tests and plots were used to understand the variables' distribution and the effect on the target.



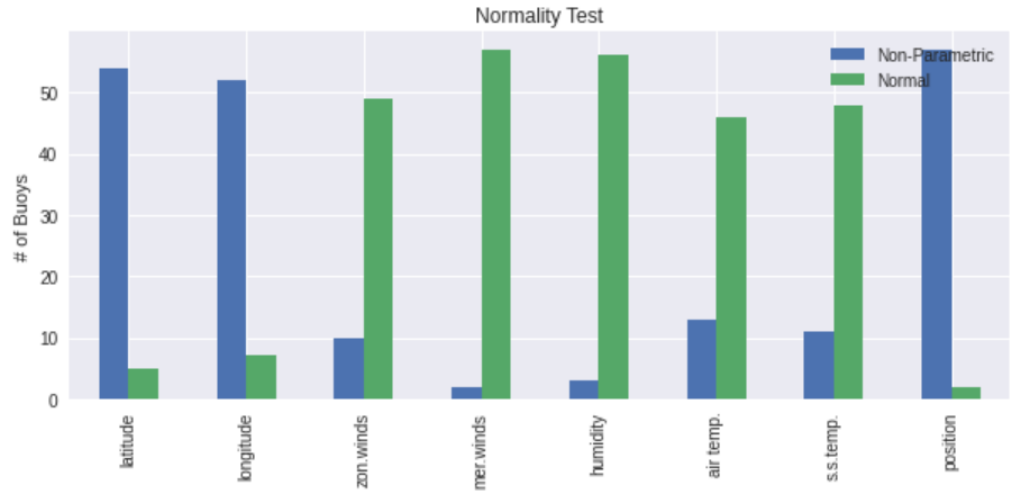
The statistical distribution of different variables along with sea surface temperature indicate that 'Air Temperature' and 'Sea Surface Temperature' are left skewed and the average temperatures are around 28 degrees Celsius.
The Humidity and Air temperature had the strongest relationship, where the negative value indicates just the direction of the linearity. Humidity seemed to have a secondary effect on Sea Surface Temperature. The positional data had little effect on Humidity, and Air temperature had increasing strengths of relationship in a linear sense.

- Is there any statistical linear relationship between the attributes?

For continuous numeric data, testing of normality is very important because based on the normality status, measures of central tendency, dispersion, and selection of parametric/nonparametric tests are decided. The Shapiro–Wilk test should be used as it has more power to detect the nonnormality and this is the most popular and widely used method. Initial null hypothesis test on the target variable indicated that the sea surface temperature data is normally distributed.

To analyze the numerical correlation coefficient or the presence of non-parametric data the actual UCI El nino dataset that contains buoy data was used. For statistical analysis it is usually important to identify non-parametric data sets with a normality test because the datasets for each buoy are less than or equal to 14 days of captured data; Shapiro Normality tests will be utilized to determine normality for each buoy's captured attributes.

```
Text(0, 0.5, '# of Buoys')
<Figure size 432x288 with 0 Axes>
```



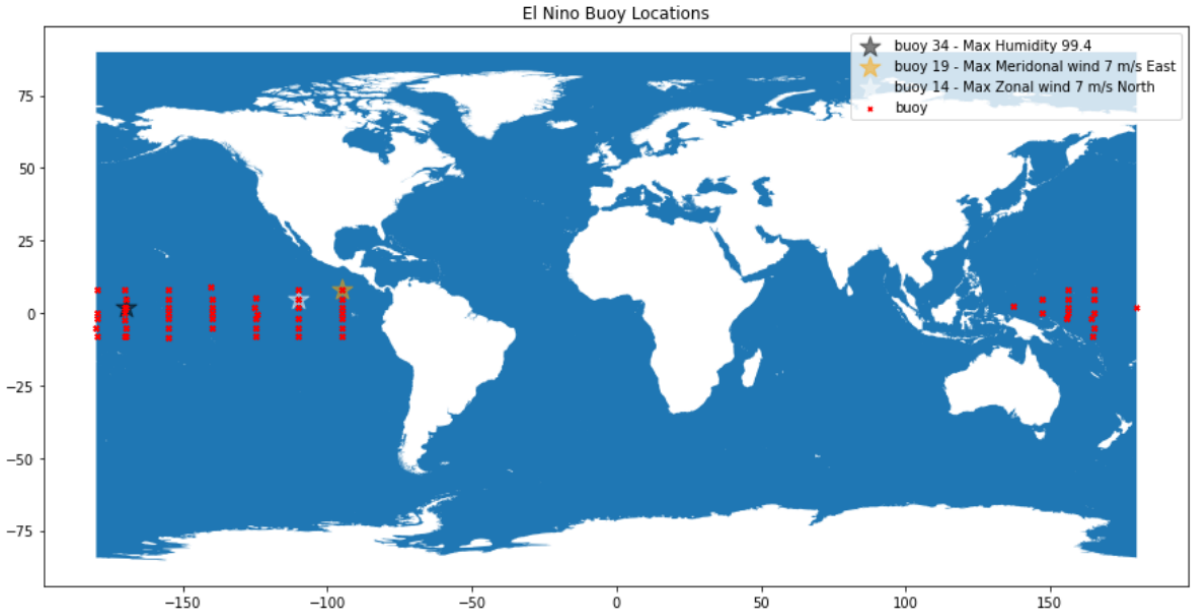
Each day range per buoy was analyzed under said Shapiro normality test, per each attribute. The Latitude and Longitude attributes were the only attributes with large non parametric consistencies, which could be expected as the data values are the result of scientific placement, not natural phenomena. While natural in terms of data, it is still interesting that all the attributes of Zonal Winds, Meridional Winds, Humidity, Air Temperature, and Sea Surface Temperature buoys were reversed with large portions of the data being normally distributed.

4. Other Insights and Found Dependencies

To help understand how the buoys are spatially located, the **Latitude** and **Longitude** data was used from the smaller dataset. This section addresses few research questions

- Where are the buoys located geographically?
- What were the maximum and average for each attribute (Zonal Winds, Meridional Winds and Humidity)?
- Does the amount of movement of the buoy affect the reliability of the data?

Using the UCI El Nino dataset that contains buoy data, a visual 2 dimensional world map plot was created that shows the buoys in the Pacific ocean with the maximum value of an attribute retrieved from a particular buoy.

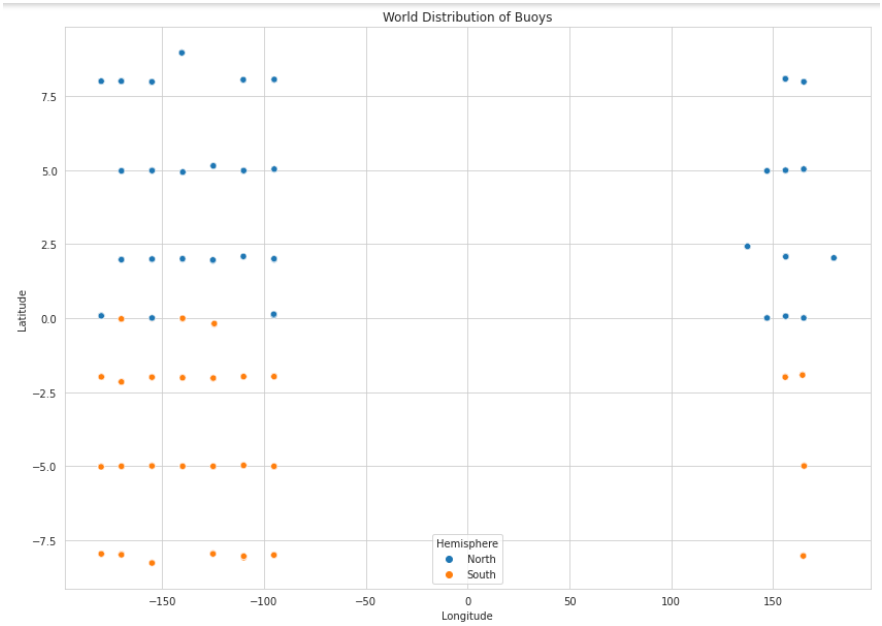


Exploratory analysis indicates Buoy#34 had the highest humidity of 99.4 degrees, where the group's average was 84.5 degrees. Buoy#19 had a zonal wind of 7 meters per second in an eastern direction, where the average wind was 3.9 m/s in a western direction. Buoy#14 had a zonal wind of 7.1 m/s north, and the average was .6 m/s south.

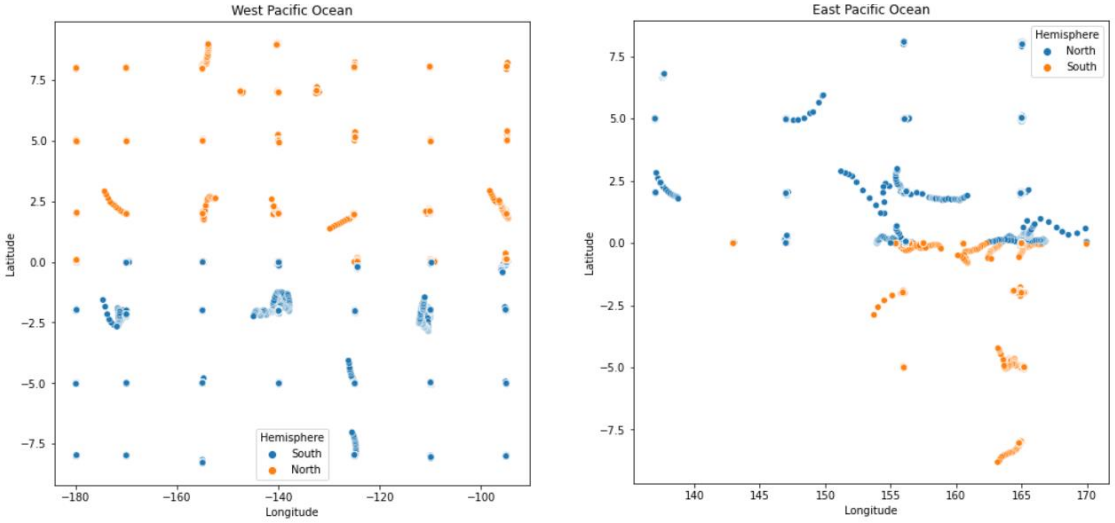
- Does the amount of movement of the buoy affect the reliability of the data?

To help understand the movement of buoy and the reliability of the data collected, the spatial grid is divided into two parts.

1. Longitude > 0 is the West Pacific Ocean and longitude < 0 is the East Pacific Ocean.
2. Latitude halves were divide as Hemispheres (North and South)



For a better understanding of the plot, the Hemisphere was highlighted in different colors. Here, we can note that the West Pacific Ocean apparently has a more equal spaced grid and more buoys than the East Pacific Ocean. Further we can look at the scatter plot for each Ocean individually.



From above scatter plot graphs we see that there is movement in the buoy and to note that the 'Latitude' axis is equal for both oceans. Buoy variation is observed between latitude 2.5 and -2.5 in both oceans.

A few more interpretations for West Pacific and East Pacific oceans are;

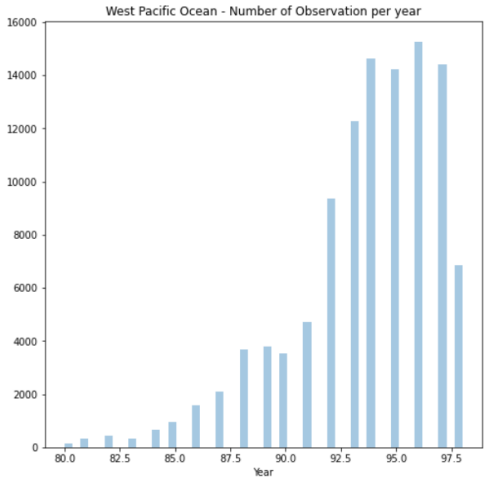
West Pacific Ocean:

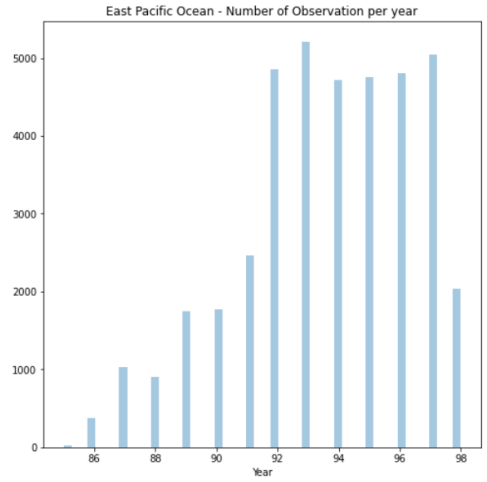
- More equal and larger spatial distributed of buoys which means more observations
- Movement of buoys seems less
- Longitude Variance about 90 degrees

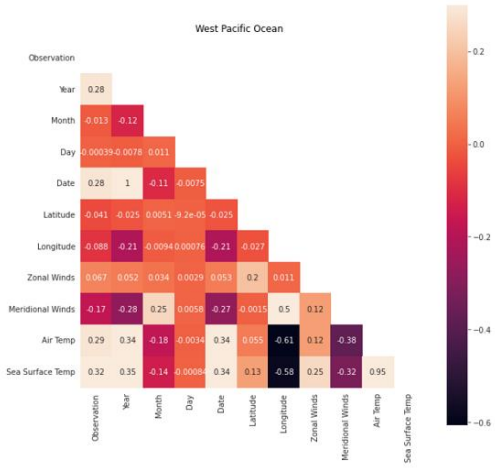
East Pacific Ocean:

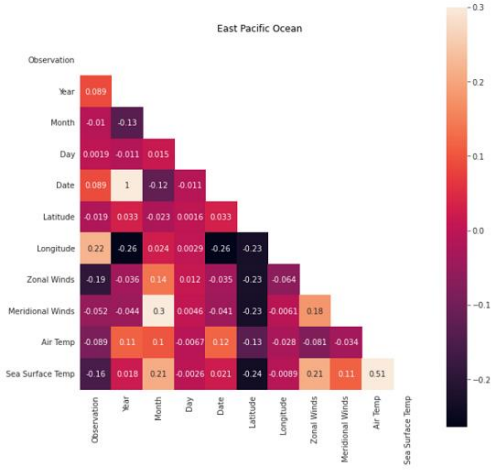
- Fewer grid of buoys and less equal spatially distributed
- Movement of the buoys seems more
- Longitude Variance about 40 degrees

By looking at the distribution plots, correlation of features, pairplots and the two linearly positively correlated attributes (Air Temperature and Sea Surface Temperature) from both the regions, we can analyze and understand if the buoy movement has any effect and is responsible for missing attributes. After collecting the missing attributes from each region, we can further conclude the data's reliability from the buoys' movement.









To sum up, below are few more conclusions after looking at the distribution and correlation plots

West Pacific Ocean:

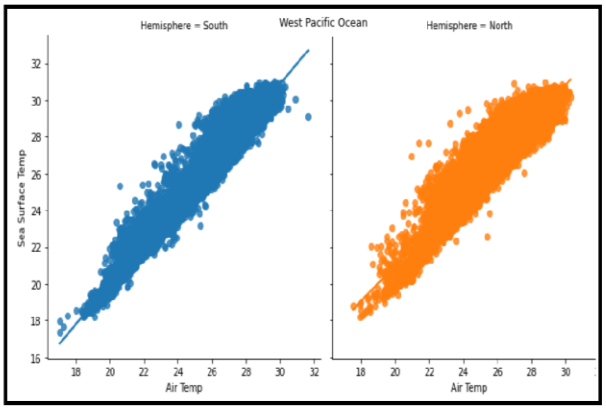
- There are 109,212 observations
- Zonal Winds and Meridional Winds have about 10% of missing data. We have to make note here that the movement of buoy was less compared to the East Pacific ocean.
- Positive correlation is observed between 'Air Temp' and 'Sea Surface Temp', with 0.95.
- Positive correlation(0.25) between 'Zonal Winds' and 'Sea Surface Temp'

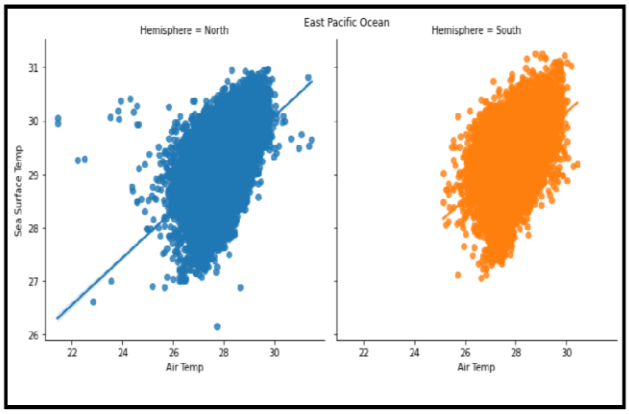
- Latitude have a positive correlation of 0.2 with Zonal Winds and 0.13 with Sea Surface Temp. There greater correlation for Latitude with Zonal Winds, Meridional Winds, Air Temp and Sea Surface Temp
- Longitude have a correlation of 0.5 for Meridional Winds which is positive but negative correlation of -0.61 for Air Temp, -0.58 for Sea Surface Temp

East Pacific Ocean:

- There are 39744 observations
- Zonal and Meridional winds missing values are about 20% of observations.
- Positive correlation is observed between 'Air Temp' and 'Sea Surface Temp', with 0.51.
- Positive correlation (0.21) is observed between 'Zonal Wind' and 'Sea Surface Temp' which is slightly less than the West Pacific value.
- Latitude have a correlation of -0.23 with Zonal Winds and others. Here we can see much greater negative correlation for Latitude.
- Similarly, Longitude has a correlation of -0.028 for Air Temp, -0.0089 for Sea Surface Temp and -0.0061 for Meridional Winds which indicates negative correlation with the winds and temperatures.

There is a good explanation that movement of buoy has a significant effect on the reliability of the variables. Below plot between the positively strongly correlated variables would also show the difference between the data collected from the West Pacific Ocean and East Pacific Ocean. We can see that there the data is linear and has less outlier than the data from East Pacific Ocean (were the buoy movement is more)





5. Predictive Modeling

5.1 Model Selection

From the EDA, we gained some insight of how each feature is associated with the target, **Sea Surface Temperature**. In this section, several models were built to predict the temperature rise in the Pacific Ocean. The regression models used in this analysis are **Linear Regression, K-Nearest Neighbors**, and **Decision Tree**.

Using train_test_split from sklearn the data was split into 70% train and the testing data is set to be 30% out of the entire cleaned data. Test set will be used to measure the performance of predictions for each model.

Below are few Machine Learning Libraries required to perform modeling:

- sklearn.model_selection train test split
- cross validation
- sklearn metrics
- linear regression
- KneighborsRegressor
- DecisionTreeRegressor
- GridSearchCV

5.2 Metrics Selection

Two performance metrics were used;

The first one is the **coefficient of determination**, which is usually expressed as **R^2**. The coefficient of determination is the ratio of the variance of a target explained or predicted by a model over the total variance of the target. It ranges from 0 to 1, and as the value is closer to 1, the model explains or predicts the variance of the target better.

The second metric is the **Mean Squared Error(MSE)**. The MSE is the average of the squared difference between the estimated or predicted values and the actual values of a target. This is always greater than zero. A lower value of the MSE indicates higher accuracy of predictions of a model. In this analysis, I use the square root of this metric(RMSE).

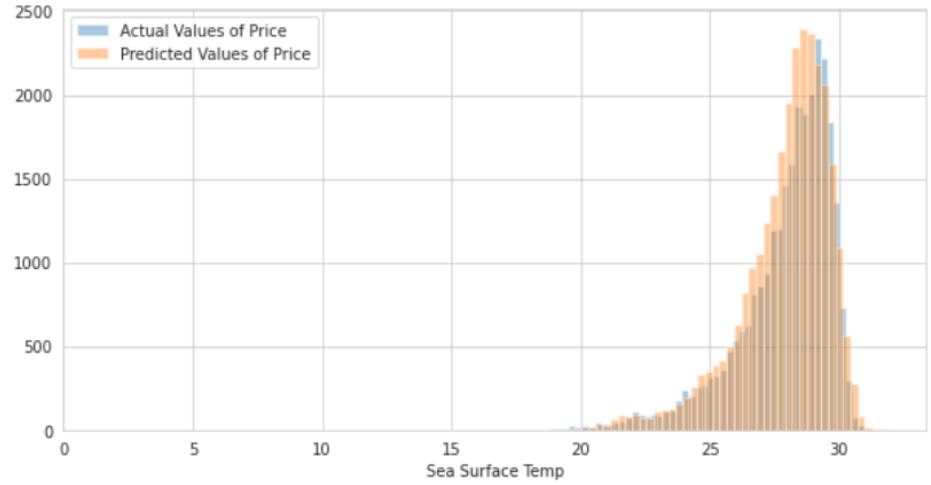
6. Basic Predictions

6.1 Linear Regression

A linear regression model is used to find a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares.

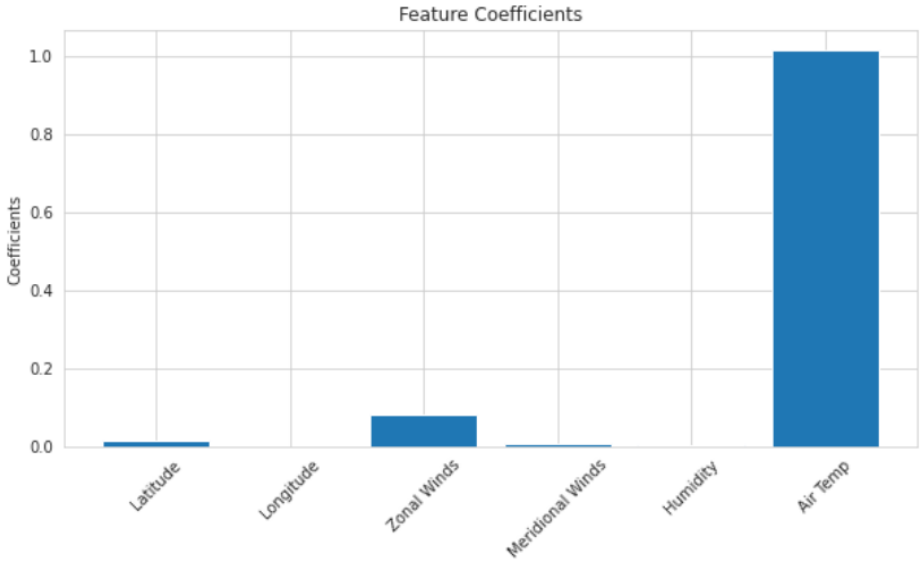
The Linear regression results for our dataset

```
R_squared: 0.9120690361149137
Square Root of MSE: 0.5524855619133298
```



- R^2 is 0.91 which means 91% of the sea surface temp can be predicted by the model.
- The RMSE is 0.55. This means that for all the predictions for the testing set, the average difference for each prediction is 0.55.

Each coefficient of the linear regression represents the magnitude of the impact of the feature on the target. The bar chart below shows the magnitudes for each coefficient in the model.



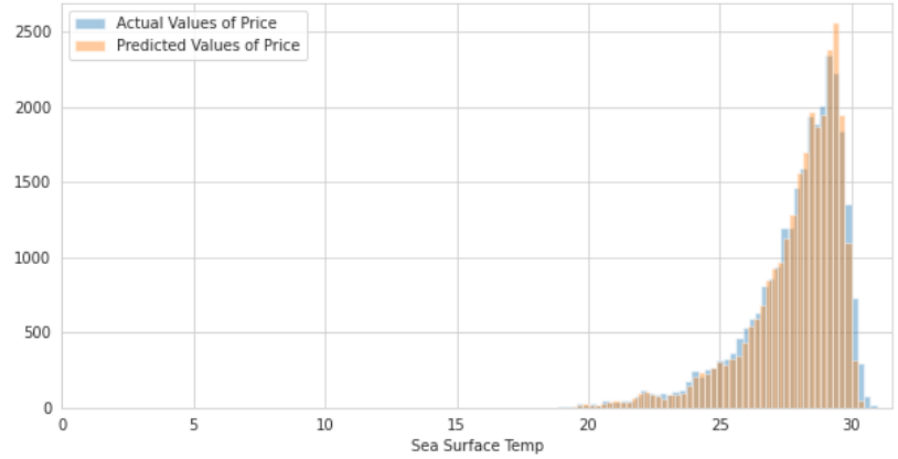
The coefficients for - longitude, Meridional Winds and Humidity in the linear regression model have least impact on the sea surface temperature. Zonal Winds and Air temp have an impact in predicting sea surface temp.

6.2 K-Nearest Neighbors Regression

KNN regression is a non-parametric method that approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood. The size of the neighborhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimizes the mean-squared error.

The kNN regression results for our dataset

```
R_squared: 0.9358772601107423
Square Root of MSE: 0.4717979523392589
```



The KNN model makes predictions for the target using the target values of K-nearest neighbors. The result above is obtained with the number of the neighbors equal to 5. The R^2 of this model is 0.935, and the RMSE is 0.47.

Compared to the distributions of the predicted values from linear and KNN, the model seems to predict the sea surface temp better.

6.3 Decision Tree Regression

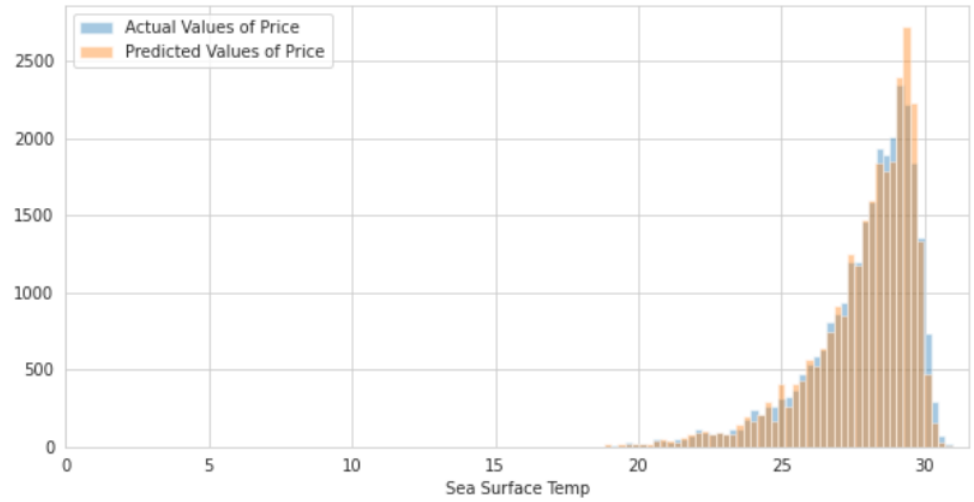
Decision Tree Regression model forms a tree structure that breaks down the dataset into smaller and smaller subsets while at the same time an association decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

In the decision tree model, the **max depth** is one of the factors to prevent the over-fitting issue of the model. As the depth of the tree is greater, the tree has more branches and becomes bigger. As the tree has more branches, the prediction for the training set can be more accurate.

However, there is a bigger variance in predicting the testing set. Therefore, setting the max depth optimally is important in order to avoid the over-fitting issue.

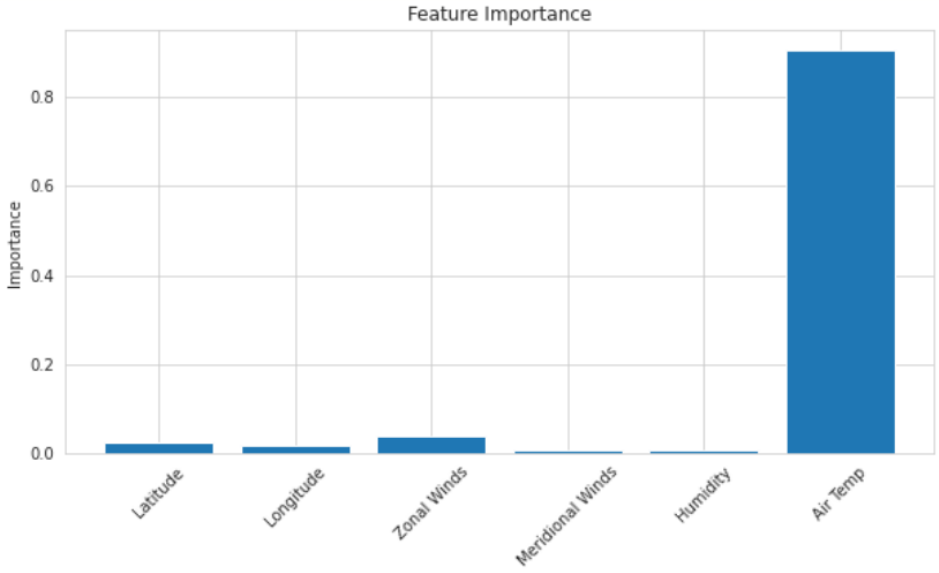
The kNN regression results for our dataset;

```
R_squared: 0.9398155233088793
Square Root of MSE: 0.4570800439883308
```



In the example above, max depth is set to 15. The R^2 of this model is 0.9398, and the RMSE is 0.457.

The decision tree does not create any coefficients for the features used in the prediction. However, by calculating how much the mean squared error decreases by selecting a feature for splits, we can measure the importance of each feature in the decision tree model.



In our dataset, air temp still appears to be most important in the predictions, and also noticed few other features like Longitude, Meridional Winds and Humidity to have a minute effect on sea surface temp.

6.4 Performance Summary

In the table below, the **Decision Tree** seems to be the optimal model to predict the Sea Surface Temp in the Pacific Ocean. However, it is too early to make a conclusion since there are more things to be considered.

Summary of Basic Results:

	R squared	RMSE
Linear Regression	0.912069	0.552486
KNN	0.935877	0.471798
Decision Tree	0.939816	0.457080

The first is that we use only one specific selection of training and testing sets, and the second is that for each model we choose one specific value for each hyper parameter. Since the values were chosen arbitrarily for the hyper parameters, the results can vary according to what values we choose for those parameters.

To get a robust result covering these issues, we need to go through the *cross validation and grid search* process as well.

7. Cross-Validation and adjustments of model hyper-parameters using Grid Search

Cross-Validation(CV) is a resampling procedure used to evaluate the models on unseen data with a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

CV randomly splits the entire data into K-folds, fit a model using (K-1) folds, validates the model using the remaining fold, and then evaluates the performance through metrics. After this, CV repeats this whole process until every K-fold is used as the testing set. The average of the K-number of scores of a metric is the final performance score for the model.

Grid-search is used to find the optimal hyperparameters of a model which results in the most 'accurate' predictions. It is an exhaustive search that is performed on the specific parameter values of a model. The model is also known as an estimator. A model **hyperparameter** is a characteristic of a model that is external to the model and whose value cannot be estimated from data. The value of the hyperparameter has to be set before the learning process begins. For

example, c in Support Vector Machines, k in k-Nearest Neighbors, max_depth in Decision tree, the number of hidden layers in Neural Networks.

7.1 Linear Regression

Since the linear regression does not have any hyper parameter in our analysis, only CV is performed here. The number of the folds in the CV is set to be 5.

```
CrossVal(LinearRegression())
```

```
R_squared: 0.8991446822431355
Square Root of MSE: 0.5682419569232426
```

The average of the R^2 is 0.89 and the RMSE is 0.

7.2 KNN Regression

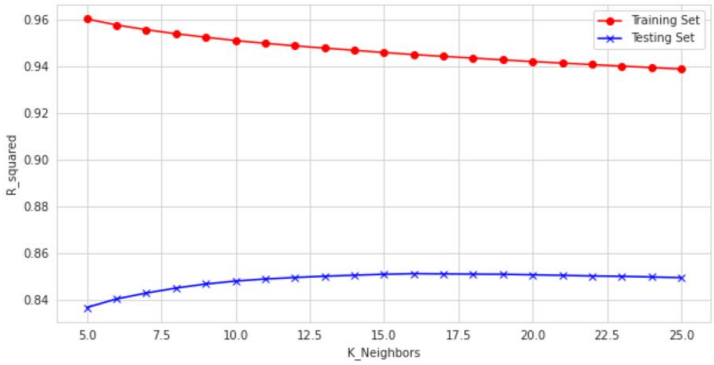
The hyper-parameter for the KNN that is used in the analysis is the number of nearest neighbors (n_neighbors). The range for the gris is the integers from 5 to 25.

```
param_grid = dict(n_neighbors=np.arange(5,26))

GridSearch(KNeighborsRegressor(), X, y, param_grid)
```

```
R_squared
The Best Parameter: {'n_neighbors': 16}
The Score: 0.8510733582554536

Square Root of MSE
The Best Parameter: {'n_neighbors': 16}
The Score: 0.6009227105104002
```

The optimal number of n_neighbors is 16. The R^2 is 0.851 and the RMSE is 0.6.

We can see how 16 is the optimal value for n_neighbors in our analysis by looking at the below Validation Curve.



7.3 Decision Tree Regression

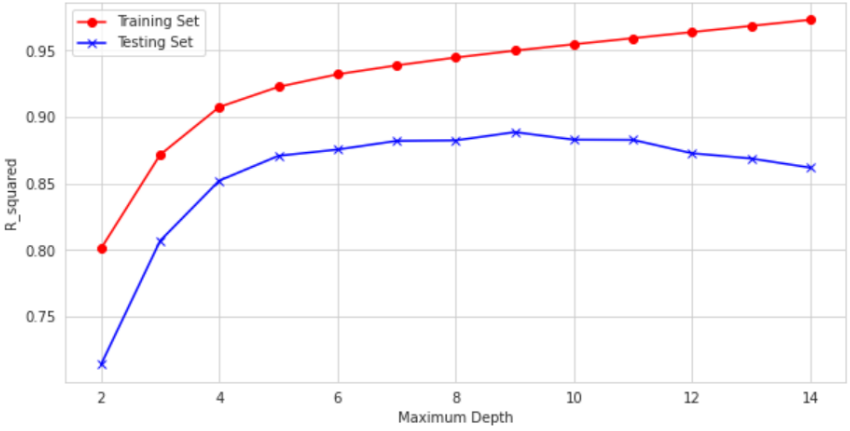
In the decision tree model, there can be several hyper parameters to be considered. In our analysis, only the max_depth is the option for the hyper parameter. The range of the max_depth to be checked is the integers from 2 to 15.

```
param_grid=dict(max_depth=np.arange(2,15))

GridSearch(DecisionTreeRegressor(random_state=0), X, y, param_grid)
```

```
R_squared
The Best Parameter: {'max_depth': 9}
The Score: 0.8887238489659964

Square Root of MSE
The Best Parameter: {'max_depth': 9}
The Score: 0.5163982264777722
```

The result indicates that the optimal value for the max_depth is 9. The R^2 is 0.88 and the RMSE is 0.51 under max_depth=9.

This is confirmed in the below validation curve as well.

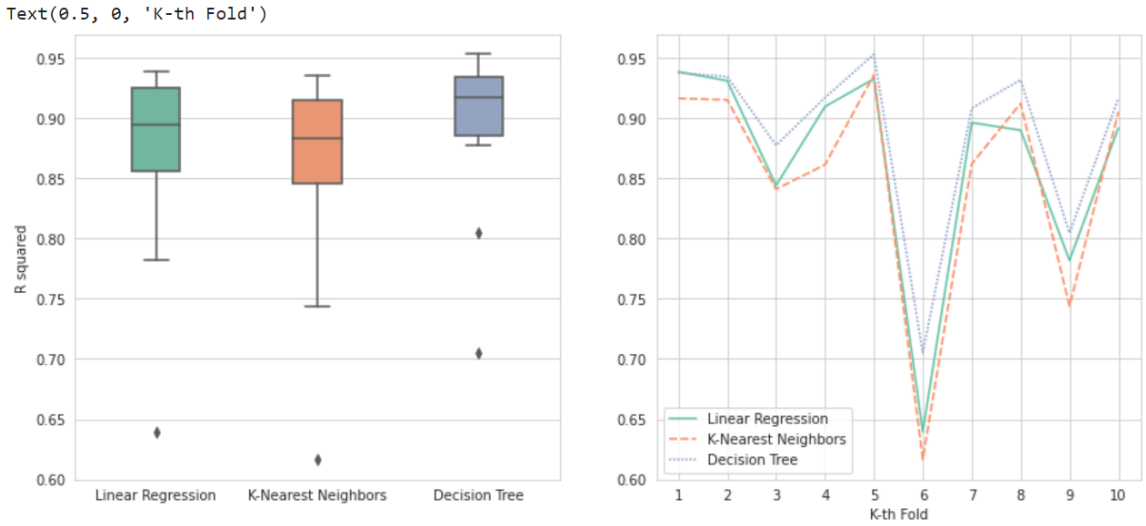


8. Plotting Cross-Validation Summary

Now that we have the hyperparameters for KNNand decision tree we can summarize the cross validation table. The table and the graphs below show the scores of the R^2 for each round of testing in CV. Since cv is set to be 10, we have 10 rounds of testing.

	Linear Regression	K-Nearest Neighbors	Decision Tree
1	0.938738	0.916592	0.937776
2	0.931256	0.915346	0.934679
3	0.844145	0.841111	0.877514
4	0.909687	0.861453	0.917183
5	0.932827	0.935987	0.953179
6	0.639539	0.616568	0.705168
7	0.896386	0.861955	0.908387
8	0.890099	0.912355	0.931939
9	0.781909	0.743892	0.804713
10	0.892075	0.905474	0.916700
Mean	0.865666	0.851073	0.888724

According to the result in the table, the best machine learning model in our analysis is the **DECISION TREE** since the mean of the scores for each round is the highest.



The box and line plots above show the distributions and the changes of the scores for each model. The **decision tree** model as well as Linear Regression shows a good performance in this analysis. KNN did not show a significant difference in their performance.

9. Conclusions

The data used in this analysis is the Elnino in UCI Machine Learning data set. The features selected for the prediction of the Sea Surface Temperature are 'Latitude', 'Longitude', 'Zonal Winds', 'Meridional Winds', 'Humidity', 'Air Temp'. For each numerical feature, the data observations beyond 1.5 times of the interquartile range are excluded as outliers. The models used for the predictions are Linear Regressions, K-Nearest Neighbors, and Decision Tree. The model showing the best performance for the prediction is **Decision Tree**, and the optimal value for the maximum depth is 9.

9.1 Additional Comments

Computational Complexity: The complexity of the algorithms are always expressed using Big O notation, which defines an upper bound of the algorithm. In this dataset, we have

n = number of training examples(178079),
d= number of dimensions of the data(9).

1. The complexity of K Nearest Neighbors to find k closest neighbor is O(knd), where k = number of neighbors.
2. The complexity of Logistic regression is O(nd)
3. The complexity of Decision Tree is O(n * log(n) * d)

This is a dataset with a minimum number of dimensions and training examples which doesn't need much computational time or complexity.

Lessons: This is a dataset that I would like to work on which contain real-time values and most importantly a dataset that I am most interested in to learn more about the variations in the climate. From this project I learned about the climate dataset, the weather conditions that cause global changes and the variables that affect the most in these climate changes. This is a unique experience as I have done an official regression analysis or even implemented the first machine learning model in python.

This is a very interesting project that explains to readers the key El nino conditions like

- The above-average Sea surface temperature that refers to warmer weather pattern in pacific ocean
- The data collected from different buoys cannot be reliable and depends on the location of a buoy, and that the buoys can record insufficient data
- All the regression methods yield good results and set examples for real life use cases. The dataset can be used for a lot more analysis considering more time.

10. References

United States El Niño Impacts | NOAA Climate.gov

Shape file downloaded from
Natural Earth » Downloads - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales (naturalearthdata.com)

GitHub - nvkelso/natural-earth-vector: A global, public domain map dataset available at three scales and featuring tightly integrated vector and raster data.