Project Milestone

Project Name: Predicting impact of CO2 emission, temperature, windspeed and other parameter on sea level

DCS 550

By: Aniruddha Joshi

Date: Mar 01, 2023

# **Introduction**

Rising sea levels pose a significant threat to many costal countries around the world, with potential consequences including erosion, and salinization of land. Understanding the factors that influence sea level rise is crucial for predicting future impacts and developing effective mitigation strategies.

This project investigates the relationship between various environmental parameters and sea level. The project aims to develop a model that can predict how sea level changes in response to factors such as:

1. Temperature : the temperature recorded
2. CO2 Emissions : the CO2 emissions recorded
3. Precipitation : the precipitation recorded
4. Humidity : the humidity recorded
5. Wind Speed : the wind speed recorded

The outcome variable

* Sea Level Rise : the sea level rise recorded

**Impact of Sea Level Rise:**

1. Sea level rise is a global issue with far-reaching consequences. It not only threatens coastal communities but also poses risks to biodiversity, ecosystems. Understanding the drivers behind sea level changes is crucial for effective mitigation strategies.
2. Millions of people inhabit coastal regions, and sea level rise directly impacts their livelihoods, homes, and well-being. Predictive models can help governments and communities plan for the future, minimizing the potential humanitarian crises associated with rising sea levels.
3. Coastal infrastructure, including ports and urban centers, faces significant risks from sea level rise.
4. Predictive models that consider various environmental factors can contribute valuable information for the formulation of evidence-based policies to address and alleviate the impacts of sea level rise.

As we face the environmental change challenge, the need for a comprehensive understanding of the factors driving sea level rise has never been more critical. This project aims to identify the relationship between the CO2 emissions, temperature shifts, windspeed variations, and other parameters, providing a roadmap for predicting sea level changes.

By analyzing historical data and employing machine learning techniques, this project aims to:

* Quantify the combined effects of CO2 emissions, temperature, windspeed, humidity and other relevant parameters on sea level rise/fall.
* Develop a predictive model that can forecast future sea level rise under different scenarios.
* The findings of this project can contribute to informed decision-making regarding coastal management, adaptation strategies, and climate change mitigation efforts.

## **Data Sets**

Following datasets are being used:

Refer to <https://www.kaggle.com/datasets/goyaladi/climate-insights-dataset/data>

The data is obtained from Kaggle website.

## **Summary of Milestones 1-3**

## EDA:

Lets look at the data.

A screenshot of a computer

Description automatically generated

It consists of below columns:

1. Date : the date of the record
2. Location : the location of the record
3. Country : the country of the record
4. Temperature : the temperature recorded
5. CO2 Emissions : the CO2 emissions recorded
6. Sea Level Rise : the sea level rise recorded
7. Precipitation : the precipitation recorded
8. Humidity : the humidity recorded
9. Wind Speed : the wind speed recorded

The data type of each column is as shown below.

A screenshot of a computer program

Description automatically generated

Now plotting the relationship between the variables

|  |  |
| --- | --- |
| A graph of a graph showing the temperature and temperature  Description automatically generated with medium confidence | A screenshot of a graph  Description automatically generated |

From the diagram above

* Emissions and Sea Level shows a very small positive correlation (0.16 factor).
* There is no correlation between temperature and Humidity
* Overall it seems there is no correlation between Precipitation and Humidity with other parameters

## Data Preparation

**Data Cleaning:**

Data often contains errors, missing values, and outliers. Data cleaning involves identifying and rectifying such issues to enhance the quality of the dataset. Imputing missing values, correcting inaccuracies, and handling outliers are essential tasks to eliminate noise and inconsistencies that could skew analysis results.

Refer below image of violine plot to identify if there are any outliers. If yes then outliers beyond 3 standard deviations are removed.

A diagram of a variable distribution

Description automatically generated

Following steps followed during data cleaning:

* Removing rows with 50% missing data
* Filling up missing data with column averages

**Feature Engineering:**

In this process created new features such as converting seal level from scaler value to categorical value such as Increase or decrease. There are approximately 50% values are above 0 indicating the sea level increased.

**Data Scaling:**

Scaling numeric features ensures that variables with different scales contribute equally to the model. Common scaling methods include Min-Max scaling and Z-score normalization, which standardize the range of values for numerical variables.

For this project Min-Max scaling has been used.

# Model building and evaluation:

Initial evaluation:

In the initial phase of our analysis, a detailed evaluation was conducted through regression analysis, employing various methodologies to identify the relationship between the selected variables (CO2 emission, temperature, windspeed, etc.) and the target variable, sea level. The outcomes of this analysis, summarized in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **SR. No** | **Method** | **MSE – Mean Squared Error** | **R-Squared** |
|  | Linear Regression | 0.01464 | -0.00178 |
|  | Random Forest Regressor | 0.0153 | -0.0514 |
|  | Decision Tree Regressor | 0.02918 | -0.9957 |
|  | SVR | 0.0152 | -0.0442 |

The R-squared values, a key metric reflecting the proportion of variance explained by the model, reveal a low explanatory power across all regression methodologies. With R-squared values almost near zero, it is evident that the selected regression models struggle to capture the relationships within the dataset. The models, as indicated by the Mean Squared Error (MSE), failed to effectively minimize the discrepancies between predicted and actual sea level values.

In response to above problem, a decision was made to reframe the problem. The sea level, originally treated as a continuous variable, was transformed into a categorical variable, reflecting whether it was increasing or decreasing. This strategic shift allowed to explore the data from a different perspective, leveraging classification models to understand the underlying patterns in sea level trends.

Subsequent to this transformation, Decision Tree Classifier, Random Forest Classifier, and K Neighbors Classifier, was employed. The results of this categorical approach were excellent, with all models exhibiting an accuracy, recall, f1-score are close to 100%. This indicates a substantial improvement in predictive performance compared to the initial regression attempts.

# Conclusion

The analysis and model building undertaken in this project provide insights into the relationship between various environmental parameters (CO2 emission, temperature, windspeed, humidity etc.) and sea level. The initial regression analysis, despite its thorough exploration using different methodologies, revealed a significant challenge: the models struggled to explain the variance in sea level adequately. This was evidenced by low R-squared values.

However, the decision to move from regression to a classification approach by categorizing sea level changes as either increasing or decreasing proved to be a good turning point. The subsequent implementation of classification models, including Decision Tree Classifier, Random Forest Classifier, and K Neighbors Classifier, yielded highly accurate results. This shift indicates that the dataset contains patterns that are more apparent when approached from a categorical perspective.

The base model, suggest that there is a relationship between the sea level and other environmental parameters. While achieving high accuracy in classification is a positive indicator, several considerations must be addressed before deploying the model. The next level is to add the location co-ordinates, other natural/human made factors.

Potential challenges:

Data Quality and Quantity: Ensuring data representativeness, addressing missing values, and obtaining more granular data could enhance the model's predictive capabilities.

Complexity of Environmental Factors: The model may benefit from incorporating additional environmental factors that influence sea levels, such as ocean currents, ice melt, and geological factors. Expanding the feature set could improve the model's accuracy.

Climate Change Dynamics: The project might consider incorporating more comprehensive climate change models to better understand the long-term impact on sea level.

In conclusion, the analysis in this project helped predicting sea level changes by employing a diverse range of models. The project highlights the urgency of comprehending the relationships between key parameters and sea level changes. By shedding light on the factors influencing sea level fluctuations, this work contributes to our understanding of environmental dynamics, providing a foundation for informed decision-making aimed at mitigating the effects of climate change.