

**II Trimester MSc (AI & ML) Advanced Machine Learning**

**Department of Computer Science**

# EARTH VITAL SIGNS FORECASTING: A MACHINE LEARNING APPROACH FOR PREDICTIVE ANALYSIS

by

Melvin Infant A(2348533) Marus Daniel P(2348532) Kripa Mary Jose(2348526)

January 2024



CERTIFICATE

*This is to certify that the report titled* ***Earth Vital Signs Forecasting: A Machine Learning Approach for Predictive Analysis*** *is a bona fide record of work done by* ***Melvin Infant A(2348533)****,****Marcus Daniel P(2348532), Kripa Mary Jose(2348526****), in partial fulfillment of the requirements of II Trimester of Msc Artificial Intelligence and Machine Learning during the year 2023-24.*

**Course Teacher**

Valued-by: (Evaluator Name & Signature) 1.

2.

Date of Exam:

# Table of Contents

**Page Number**

1. **Abstract**
2. **Introduction**
3. **Data Pre-processing and Exploration**
   1. Data understanding and exploration
   2. Data cleaning and handling missing values
   3. Data integration and feature engineering
4. **Algorithm Implementation**
   1. Algorithms implemented
      1. Linear Regression
      2. Random Forest
      3. Polynomial Regression
   2. Correct parameter tuning
   3. Efficient coding and algorithm execution
5. **Model Evaluation and Performance Analysis**
   1. Evaluation metrics and performance assessment
   2. Comparative analysis of different models
   3. Insightful interpretation of results
6. **References**

# Abstract

The profound impacts of climate change on our planet necessitate advanced predictive models to understand and foresee critical Earth vital signs. This project focuses on forecasting key environmental indicators, including global warming, seasonal variations in carbon dioxide levels, and methane concentrations. The data, meticulously curated from reliable sources such as NASA, encompasses variables like global warming trends, monthly average carbon dioxide levels, and methane mean concentrations.

The primary objective of this project is to employ machine learning techniques for accurate and insightful predictions. The chosen methods include Random Forest, Multiple Linear Regression, and Polynomial Regression. Each algorithm brings unique strengths to the table, offering a comprehensive analysis of the complex relationships within the Earth's vital signs dataset.

This project is paramount in our collective journey towards sustainable development and environmental stewardship. By leveraging advanced machine learning techniques, the research offers a forward-looking approach to predicting Earth's vital signs, encompassing critical factors like global warming, carbon dioxide dynamics, and methane concentrations. In the current climate emergency, where decisive action is imperative, the project is a crucial tool for policymakers, providing them with accurate insights to formulate effective climate strategies. Aligned with key Sustainable Development Goals, it addresses the pressing need for climate action, biodiversity preservation, and responsible resource management. The project not only contributes to the global scientific community but also holds the potential to raise public awareness and foster a sense of shared responsibility in the face of escalating environmental challenges. In essence, its significance lies in its ability to provide actionable information for informed decision-making and contribute meaningfully to the global pursuit of a sustainable and resilient future.

# Introduction

Earth's vital signs are key environmental indicators such as global temperature, atmospheric carbon dioxide, and methane concentrations. These metrics serve as crucial markers, offering insights into the planet's health and climate. Monitoring sea level variations and extreme weather events further aids in understanding environmental changes and potential impacts on ecosystems. These vital signs are essential for assessing the state of the Earth's environment and guiding efforts toward sustainable practices and climate resilience.

Life on our planet is clearly under siege. The statistical trends show deeply alarming patterns of climate-related variables and disasters. Earth’s “vital signs” are worse than at any time in human history, an international team of scientists has warned, meaning life on the planet is in peril.

Many climate records were broken by enormous margins in 2023, including global air temperature, ocean temperature and Antarctic sea ice extent, the researchers said. The highest monthly surface temperature recorded recently and probably the hottest the planet has been in 100,000 years. Moreover, these vital signs are directly correlated with natural disasters, such as hurricanes and wildfires, enabling the prediction and preparation for such events. Changes in sea levels and ocean health are critical aspects, impacting coastal regions and marine life. By comprehensively understanding Earth's vital signs, scientists, policymakers, and communities can develop informed strategies for sustainable development, balancing economic progress with environmental preservation.

This project embarks on a crucial journey to forecast Earth's essential indicators, encompassing global warming trends, methane concentrations, and carbon dioxide dynamics. The urgency of such predictive modeling arises from the unprecedented changes witnessed in recent years, as evidenced by alarming shifts in global temperatures, increases in greenhouse gas levels, and the surge in extreme weather events worldwide. These changes, attributed to human activities and natural disasters, underscore the pressing need for accurate forecasting tools to guide mitigation and adaptation strategies. Recent data from NASA reveals substantial alterations in vital signs such as extreme temperature variations, heightened carbon dioxide levels, and the intensification of storms. Consequently, the project focuses on predicting key indicators, including methane mean concentrations, global warming trends without smoothing, and various carbon dioxide metrics.

The algorithms employed - Random Forest, Multiple Linear Regression, and Polynomial Regression - are chosen for their capacity to discern intricate patterns within the complex web of environmental data. This project stands as a critical response to the environmental crises of our time, aiming to empower decision-makers with accurate and timely information for sustainable climate action.This project aspires to contribute meaningfully to our understanding of Earth's vital signs, offering a glimpse into a future where proactive measures and informed choices pave the way for a more resilient and sustainable planet.

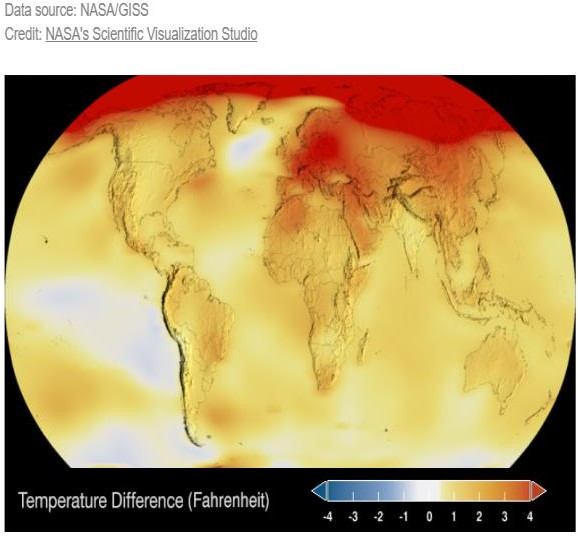
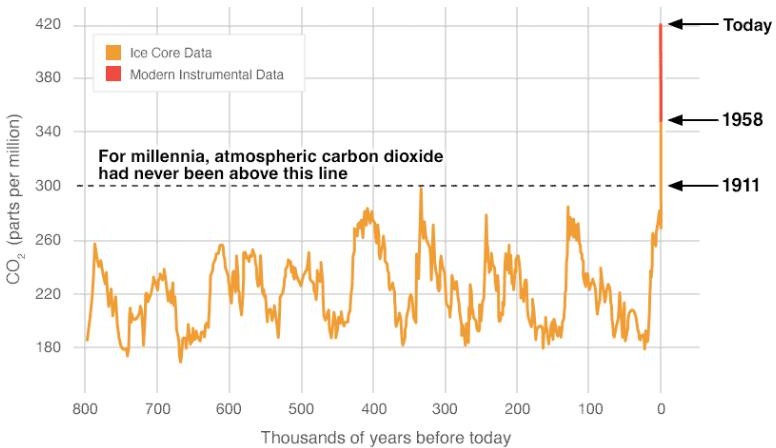


Fig a. Carbon dioxide trend from NASA Fig b: Global Temperature Trend from Nasa

# Data Pre-processing and Exploration

In this phase, we'll thoroughly analyze the dataset using descriptive statistics, data visualization, and exploratory methods. This aims to uncover its distribution, relationships, and characteristics, providing deep insights for informed decisions in subsequent stages of data pre-processing and model development.

**Datasets:**

1. Extreme Temperature:

* Identify temperature records categorized as extreme. Explore variations, frequencies, and locations of extreme temperatures. Understand the temporal and spatial distribution patterns.

1. Carbon Dioxide (CO2):

* Examine CO2 levels in the atmosphere. Investigate trends, seasonal variations, and potential sources of increased concentrations. Assess the impact on climate and air quality.

1. Global Warming:

* Explore indicators of global warming such as temperature anomalies. Analyze historical trends and patterns to understand the pace and extent of global warming over time.

1. Methane:

* Investigate methane concentrations, a potent greenhouse gas. Examine sources and variations, considering both natural and anthropogenic contributions. Understand the role of methane in climate change.

1. Storm Data:

* Analyze storm-related data, including frequency, intensity, and geographical distribution. Explore the correlation between storm occurrences and other vital signs. Understand the implications for weather patterns and climate.

**3.2 Data Cleaning and Handling Missing Values**

In this phase, we identify and rectify missing or erroneous data points in the dataset using techniques like imputation, outlier removal, and data integrity validation. Fortunately, no missing values were discovered, and outliers were successfully identified and removed. For datasets covering the years 1980 to 2022, missing values were imputed. In cases where some datasets had missing values, we filled them with zeros to ensure comprehensive and reliable data for subsequent analysis.

**3.3 Data Integration and Feature Engineering**

Feature engineering involves the strategic selection and construction of pertinent features to enhance the predictive capabilities of our models. Techniques like encoding categorical variables, assessed Variance Inflation Factor (VIF) values to identify relevant features. This step is pivotal in developing predictive models that can effectively discern patterns and predict vital signs with high accuracy. These preprocessing stages are indispensable, serving as the foundation for the subsequent application of machine learning techniques to forecast Earth's vital signs accurately.

1. **Algorithm Implementation**
   1. **Algorithms implemented**
      1. **Linear Regression**

Linear Regression is a statistical method for modeling the relationship between a dependent variable (Y) and one or more independent variables (X). The formula for simple linear regression is :

Y = b0 + b1\*X

Y is the dependent variable, b0 is the intercept, b1 is the slope, and X is the independent variable. The goal is to find the values of b0 and b1 that minimize the difference between the observed and predicted values of Y.

The formula for multiple regression is:

Y = b0 + b1\*X1 + b1\*X2+..+bn\*Xn

Where Y is the independent variable, bo is the intercept, and b1,b2, and bn are the coefficients for each respective independent variable.

# Random Forest Regression

Random Forest Regression is an ensemble learning method that combines multiple decision trees to make predictions. The formula involves aggregating the predictions from each tree to produce the final result. The overall prediction is obtained by averaging or taking a weighted sum of the forecasts from individual trees. The algorithm is robust, less prone to overfitting, and suitable for both classification and regression tasks.

# Polynomial Regression

Polynomial Regression is a form of linear regression where the relationship between the independent variable (X) and the dependent variable (Y) is modeled as an nth-degree polynomial. The formula is :

Y = b0 + b1\*X + b2\*X^2 + ... + bn\*X^n

Y is the dependent variable, X is the independent variable, and b0, b1, ..., bn are the coefficients to be determined. This regression type accommodates non-linear relationships by introducing polynomial terms.

# Correct parameter tuning

Employ techniques like grid search to tune algorithm hyperparameters efficiently, enhancing model performance without manual trial and error. Grid Search is a hyperparameter tuning technique widely employed in machine learning to optimize the performance of a model. In essence, it systematically searches through a predefined grid of hyperparameter values to find the combination that yields the best model performance. The hyperparameters are crucial settings for a machine learning algorithm that are not learned from the data but are set prior to the training process.

# Efficient coding and algorithm execution

To ensure effective coding and streamlined algorithm execution in machine learning projects, we considered the following best practices:

* **Data Preprocessing**: Efficiently clean and preprocess data by addressing missing values and outliers and encoding categorical variables. Utilize libraries like pandas and sci-kit-learn for proficient data manipulation.
* **Feature Selection**: Opt for relevant features that significantly contribute to the prediction task, reducing data dimensionality and enhancing efficiency. Techniques such as correlation analysis, feature importance from tree-based models, or model-based feature selection can be applied.
* **Algorithm Selection:** Choose algorithms appropriate for the specific problem, weighing the trade-offs between complexity, accuracy, and computational efficiency. Ensemble methods like or Random Forest can be efficient and effective for large datasets.
* **Hyperparameter Tuning**: Employ techniques like grid search to tune algorithm hyperparameters efficiently, enhancing model performance without manual trial and error.
* **Algorithm Implementation:** Implement algorithms using optimized libraries like Tree and Sklearn in Python, designed for efficient numerical computations and machine learning tasks.
* **Code Optimization:** Write clean, modular, and optimized code. Leverage vectorized operations and minimize unnecessary loops for improved computational efficiency.

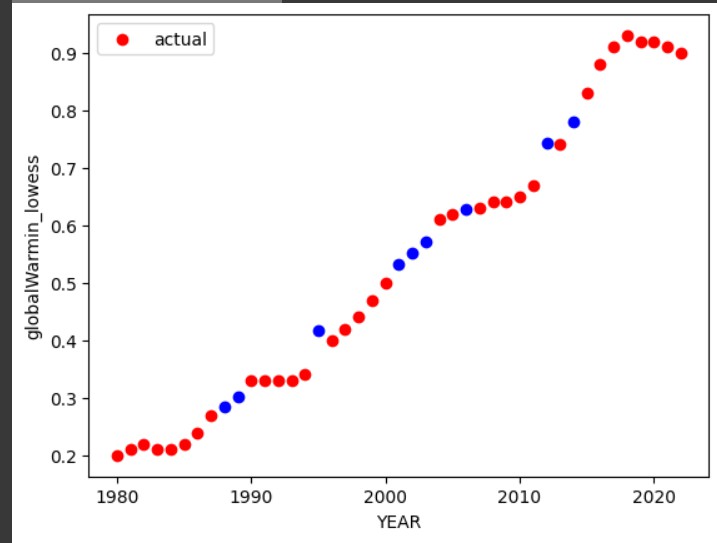
# Model Evaluation and Performance Analysis

* 1. **Evaluation Metrics and Performance Assessment:**

We have used MAE, MSE, and R Score to understand the algorithm’s efficiency. Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors, providing a straightforward measure of model accuracy. Mean Squared Error (MSE) squares the errors, emphasizing more significant discrepancies, and is helpful in understanding the overall model performance. R2 Score assesses the proportion of variance explained by the model, indicating the goodness of fit and how well predictions align with the actual data. Together, these metrics offer a comprehensive evaluation of predictive model performance.

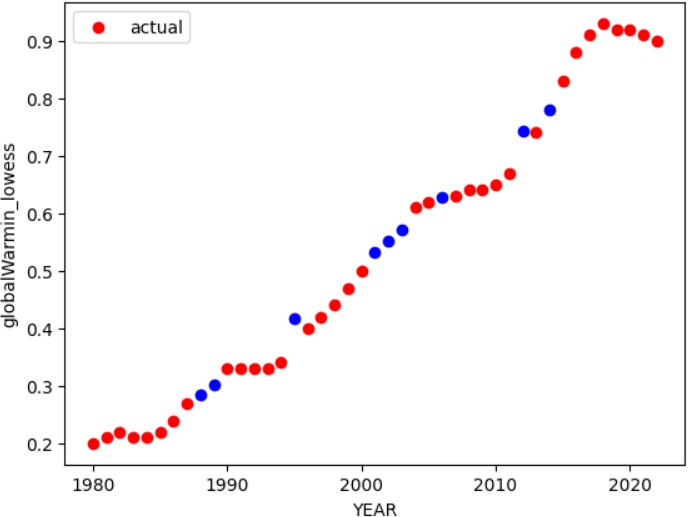
# Linear Regression:

Mean Absolute Error: 0.020383260073923597 Mean Squared Error: 0.0006519102205942983 R2 Score: 0.9735235018711702



# Random Forest Regression:

Mean Absolute Error: 0.020222222222222277 Mean Squared Error: 0.0005396800000000046 R2 Score 0.9780815884476532

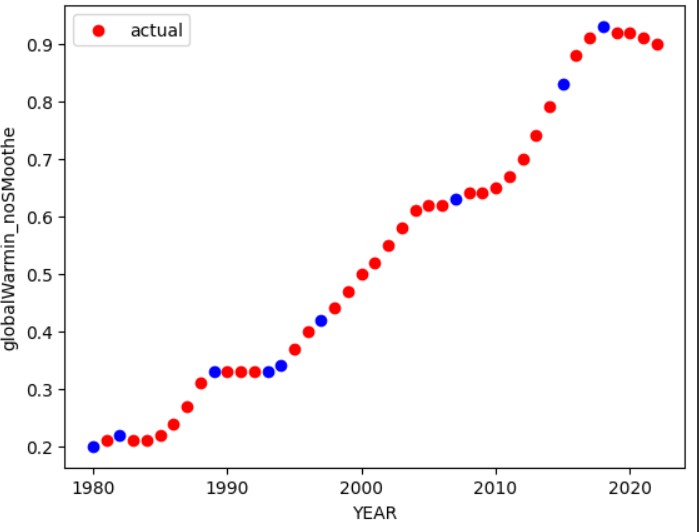


# Polynomial Regression:

CO-EFFICIENT: [ 0.00000000e+00 -4.24953878e-01 1.10910537e-04] INTERCEPT: 406.75885642487157

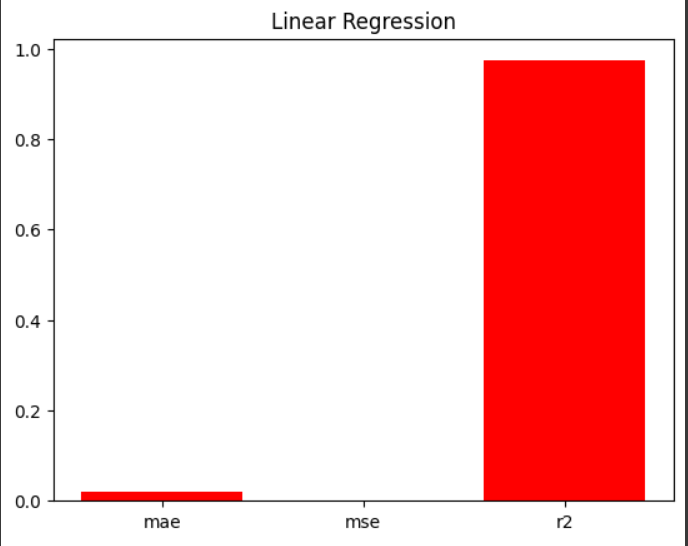
R-squared (R2): 0.9789

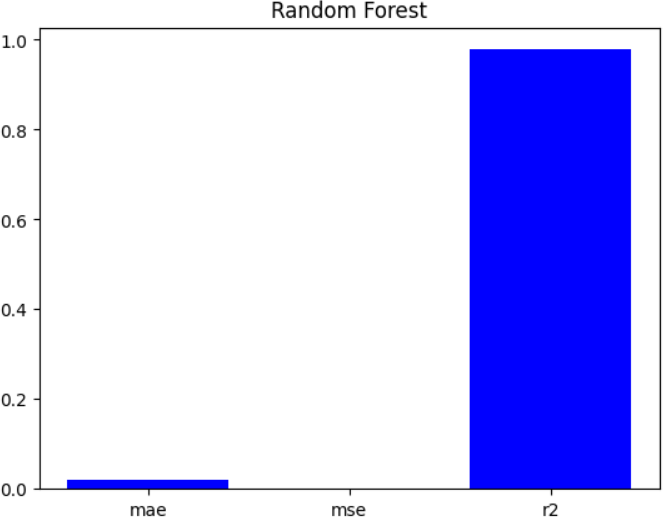
Mean Squared Error (MSE): 0.0013 Mean Absolute Error (MAE): 0.0322

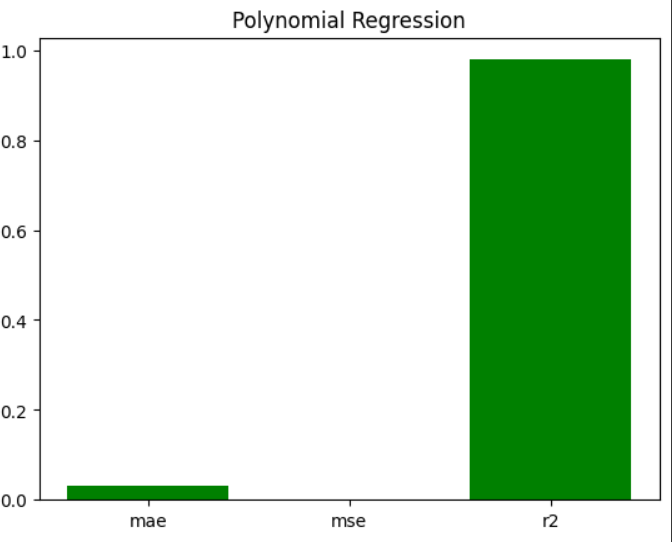


# Comparative Analysis of Different Models:

A comparative analysis of algorithms is crucial for selecting the most effective model based on metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 Score. MSE measures the average squared differences between predicted and actual values, providing insight into prediction accuracy. By comparing these metrics across algorithms, one can identify the model that not only minimizes errors but also maximizes the explained variance, ensuring a comprehensive evaluation and informed selection of the best-performing algorithm for a given task.







Among the three algorithms evaluated, the linear regression model demonstrated the lowest Mean Squared Error (MSE), indicating its superior performance in capturing the underlying patterns in the data. The linear regression analysis yields a predictive line with minimized MSE, effectively representing the relationship between the observed features and outcomes.' This line is characterized by its ability to closely fit the data points, offering a reliable model for forecasting and understanding the trends in the given dataset.

# Insightful Interpretation of Results

The results of the model evaluation provide valuable insights into the effectiveness of machine learning techniques in predicting the vital signs. The interpretation of these results should consider the models, potential overfitting, and their capacity to generalize to new data. It is essential to analyze the implications of each model's performance in context to understand how these models could potentially contribute to early detection and intervention.

From the above project it is found that:

The linear regression model demonstrates good R2 showing its ability to correctly predict cases in unknown data samples.The random forest exhibits lowesr mse and mae which suggesting robust performance in predicting of known data but doesnt perform very well on unknown data.The polynomial regression model shows good r2 fit but comparatively not performs well on unknown data models also indicating potential limitations in its predictive performance for vital signs which maybe due to lack of data as the proper data is available from the year 1980

# Conclusion

The provided data reveals a clear correlation among the various features, with each feature exhibiting a connection to one or more others. Notably, there is a discernible upward trend over the years, evident not only in the overall global warming metrics such as globalWarming\_no\_smoothing and globalWarming\_lowess but also extending to interconnected features like carbondioxide\_MonthyAverage, carbondioxide\_DeSeasonalized, and carbondioxide\_UncOfMonth. This cumulative effect further influences variables like methane\_mean and ocean\_temperature.

Upon closer examination through scatter plots, it becomes apparent that this upward trajectory follows a linear pattern. As the years progress, a consistent increase is observed in each feature, indicating a potential cause for concern. The interdependence of these features suggests a complex network of influences, wherein the rise in global warming metrics triggers a chain reaction affecting carbon dioxide levels, which, in turn, impacts methane levels and ocean temperatures.

Given this steady increase across multiple indicators, there arises a compelling prediction that if the current trend persists, the Earth may soon reach an alarming state. With predictions extended over the next 10 years, it becomes imperative to closely monitor and address these trends to mitigate potential adverse effects on the planet's overall health and sustainability.

# Reference

1. NASA. "Vital Signs of the Planet: Global Climate Change and Global Warming." NASA, [Online]. Available: https:/[/www](http://www.nasa.gov/earth/climate-change/vital-signs/).[nasa.gov/earth/climate-change/vital-signs/](http://www.nasa.gov/earth/climate-change/vital-signs/)
2. NASA. "Climate Change: Vital Signs of the Planet." NASA, [Online]. Available: https://climate.nasa.gov/effects/
3. ABC News. "Climate emergency: 16 vital signs of the planet's health are now at record extremes." ABC News, 2021. [Online]. Available: https:/[/www](http://www.abc.net.au/news/science/2021-07-).[abc.net.au/news/science/2021-07-](http://www.abc.net.au/news/science/2021-07-) 28/climate-emergency-pandemic-planet-vital-signs-worsening/100329094
4. New Scientist. "16 vital signs of the planet's health are now at record extremes." New Scientist, [Online]. Available: https:/[/www](http://www.newscientist.com/article/2344150-16-vital-signs-of-the-planets-).[newscientist.com/article/2344150-16-vital-signs-of-the-planets-](http://www.newscientist.com/article/2344150-16-vital-signs-of-the-planets-) health-are-now-at-record-extremes/
5. Scientific American. "Earth's Latest Vital Signs Show the Planet Is in Crisis." Scientific American, [Online]. Available: https:/[/www](http://www.scientificamerican.com/article/earths-latest-vital-).[scientificamerican.com/article/earths-latest-vital-](http://www.scientificamerican.com/article/earths-latest-vital-) signs-show-the-planet-is-in-crisis/
6. GeeksforGeeks. "ML | Linear Regression." GeeksforGeeks, [Online]. Available: https:/[/www](http://www.geeksforgeeks.org/ml-linear-regression/).[geeksforgeeks.org/ml-linear-regression/](http://www.geeksforgeeks.org/ml-linear-regression/)
7. Bing Search. "Multiple Linear Regression." Bing, [Online]. Available: https:/[/www](http://www.bing.com/search?q=multiple%2Blinear%2Brregression&cvid=c09c0e4dc2f844fca0245bdb).[bing.com/search?q=multiple+linear+rregression&cvid=c09c0e4dc2f844fca0245bdb](http://www.bing.com/search?q=multiple%2Blinear%2Brregression&cvid=c09c0e4dc2f844fca0245bdb) ef7ce1b1&gs\_lcrp=EgZjaHJvbWUyBggAEEUYOTIGCAEQIxgnMgYIAhAjGCcyBggDECMY JzIGCAQQIxgnMgYIBRAjGCcyBggGEAAYQDIGCAcQABhAMgYICBAAGEDSAQg1NTk 5ajBqOagCALACAA&FORM=ANAB01&PC=DCTS
8. GeeksforGeeks. "Random Forest Regression in Python." GeeksforGeeks, [Online]. Available: https:/[/www](http://www.geeksforgeeks.org/random-forest-regression-in-python/).[geeksforgeeks.org/random-forest-regression-in-python/](http://www.geeksforgeeks.org/random-forest-regression-in-python/)
9. GeeksforGeeks. "Python Implementation of Polynomial Regression." GeeksforGeeks, [Online]. Available: https:/[/www](http://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/).[geeksforgeeks.org/python-implementation-of-polynomial-regression/](http://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/)

**Team Details**

|  |  |  |
| --- | --- | --- |
| **Reg. no** | **Name** | **Summary of tasks performed** |
| **2348533** | **Melvin Infant A** | Collected datasets from NASA Global Climate Change, handled data preprocessing tasks, explored model1's dataset for insights, identified vital features, implemented model, and rigorously evaluated performance using relevant metrics. |
| **2348532** | **Marcus Daniel P** | Managed data preprocessing, examined insights from model2's dataset, pinpointed crucial features, implemented the model, and thoroughly assessed its performance using pertinent metrics, documentation |
| **2348526** | **Kripa Mary Jose** | Conducted data preprocessing tasks, explored insights within model3's dataset, identified essential features, implemented the model, and comprehensively assessed its performance using relevant metrics, documentation |