Celebrate life
RATHINAM

RATHINAM
TECHNICAL CAMPUS
(AUTONOMOUS)

A+ ACCREDITED BY NAAC

CD2L

GLOBAL CLIMATE CHANGE

KARTHICK BALASHANMUGAM	**– 2 YR - CSE**
AHAMED ISMAIL	**– 2 YR - CSE**
SHOBIKA	**– 2 YR - CSE**
VIJAYALAKSHMI	**– 2 YR - CSE**

RAISE CENTRE OF EXCELLENCE

MAY 2023

OVERVIEW:

Climate change is a major problem in the modern world and there are still some people around the world who do not believe in climate change or global warming and presenting definite proof can help spread more awareness around.

The main motivation for this project is to automate the climate change data for the machine to understand the pattern in the phenomenon and be able to predict future changes in the mean surface temperature of the earth with the pattern in the data. This will allow us to choose the measures to be taken, the magnitude of their significance, and which places are more in need of maintenance. Firstly, the data for this project was acquired from an online source and visualized the data to see if there are any existing patterns after cleaning the data. then, tests were done with various models through a simple trial and error method to deduce which model would be best effective in properly understanding the trends in the data and compared them to figure out the best-suited model for this project, then the parameters of the model were configured to increase its prediction accuracy and finally, recorded its progress.

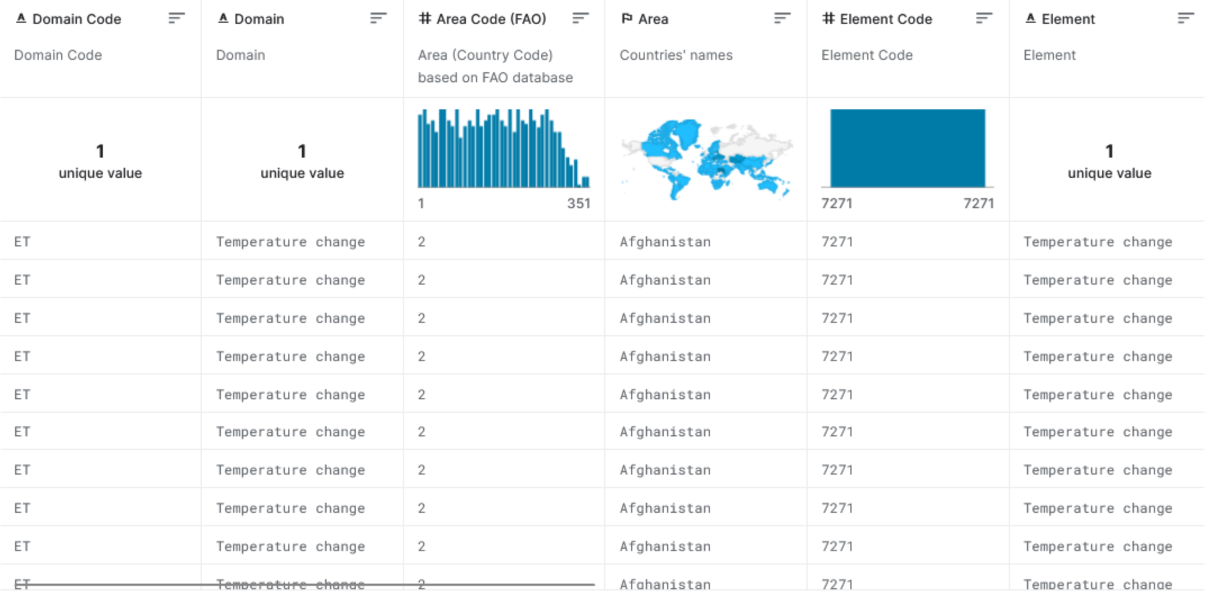
DATA ACQUISITION:

The dataset used for this project was acquired from Kaggle and uploaded by the user SEVGI SY
This dataset consists of 229926 rows and 14 columns.
Here is some information about the dataset provided by the publisher.

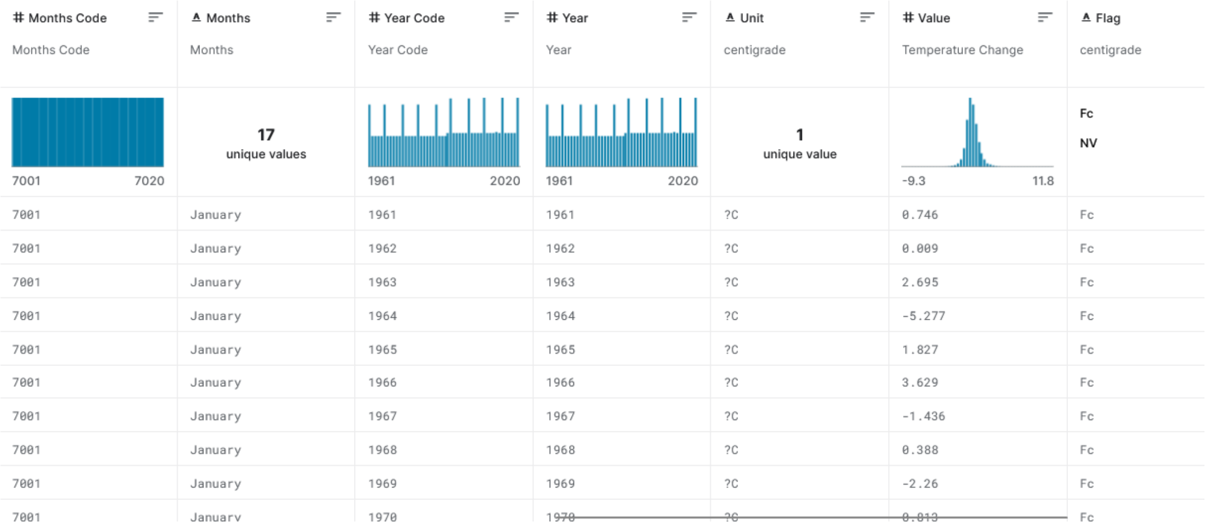
Data description

The FAOSTAT Temperature Change domain disseminates statistics of mean surface temperature change by country, with annual updates. The current dissemination covers the period 1961–2019. Statistics are available for monthly, seasonal, and annual mean temperature anomalies, i.e., temperature change with respect to a baseline climatology, corresponding to the period 1951–1980. The standard deviation of the temperature change of the baseline methodology is also available. Data are based on the publicly available GISTEMP data, the Global Surface Temperature Change data distributed by the National Aeronautics and Space Administration Goddard Institute for Space Studies (NASA-GISS).

METADATA:



Domain Code	Domain	Area Code (FAO)	Area	Element Code	Element
Domain Code	Domain	Area (Country Code) based on FAO database	Countries' names	Element Code	Element
1 unique value	1 unique value	1 351		7271 7271	1 unique value
ET	Temperature change	2	Afghanistan	7271	Temperature change
ET	Temperature change	2	Afghanistan	7271	Temperature change
ET	Temperature change	2	Afghanistan	7271	Temperature change
ET	Temperature change	2	Afghanistan	7271	Temperature change
ET	Temperature change	2	Afghanistan	7271	Temperature change
ET	Temperature change	2	Afghanistan	7271	Temperature change
ET	Temperature change	2	Afghanistan	7271	Temperature change
ET	Temperature change	2	Afghanistan	7271	Temperature change
ET	Temperature change	2	Afghanistan	7271	Temperature change
~~ET~~	~~Temperature change~~	~~2~~	Afghanistan	7271	Temperature change



Months Code	Months	Year Code	Year	Unit	Value	Flag
Months Code	Months	Year Code	Year	centigrade	Temperature Change	centigrade
7001 7020	17 unique values	1961 2020	1961 2020	1 unique value	-9.3 11.8	Fc NV
7001	January	1961	1961	?C	0.746	Fc
7001	January	1962	1962	?C	0.009	Fc
7001	January	1963	1963	?C	2.695	Fc
7001	January	1964	1964	?C	-5.277	Fc
7001	January	1965	1965	?C	1.827	Fc
7001	January	1966	1966	?C	3.629	Fc
7001	January	1967	1967	?C	-1.436	Fc
7001	January	1968	1968	?C	0.388	Fc
7001	January	1969	1969	?C	-2.26	Fc
7001	January	1970	~~1970~~	~~?C~~	~~0.013~~	Fc

The following link provides further information on the dataset: https://www.kaggle.com/datasets/sevgisarac/temperature-change?select=FAOSTAT_data_1-10-2022.csv

RAISE CENTRE OF EXCELLENCE

DATA CLEANSING AND TRANSFORMATION:

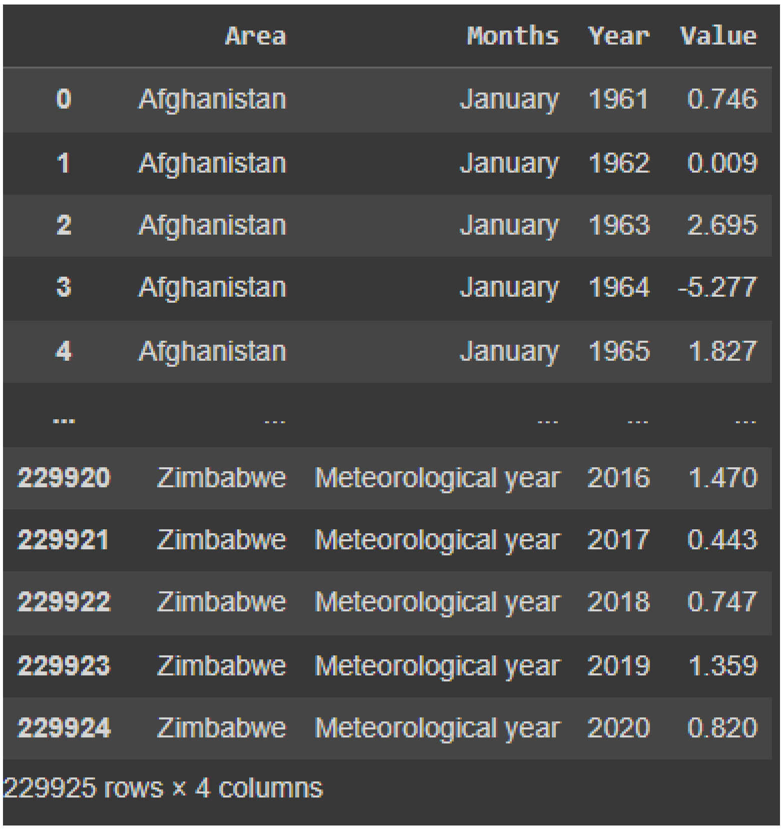
Before the patterns in the data were visualized, the dataset was evaluated to check and remove any irrelevant data

After careful evaluation, it was inferred that for this regression model, the only data that are required are from these columns:

- Area
- Year
- Months
- Value

According to the publisher of the dataset, the other columns are provided to be used for merging with other tables in the future which may contain the leading factor that affects the mean surface temperature of the earth.

After removing the other irrelevant columns using the **drop()** function in pandas, the updated table is as follows:

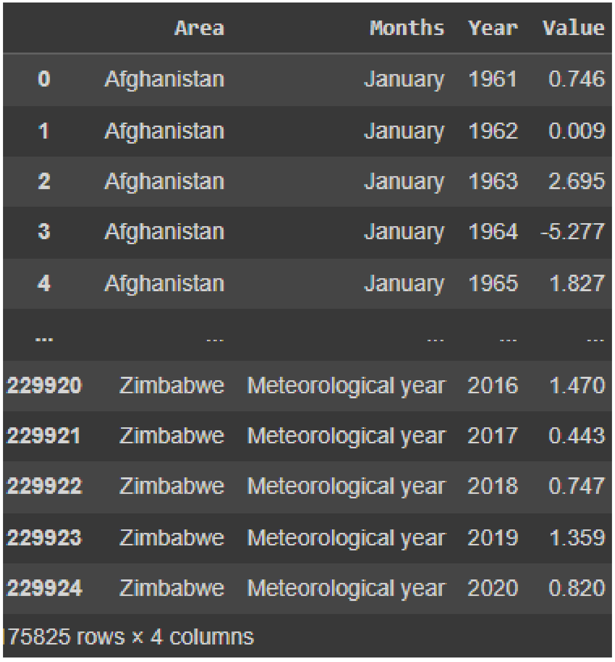


	Area	Months	Year	Value
0	Afghanistan	January	1961	0.746
1	Afghanistan	January	1962	0.009
2	Afghanistan	January	1963	2.695
3	Afghanistan	January	1964	-5.277
4	Afghanistan	January	1965	1.827
...	...	...	...	...
229920	Zimbabwe	Meteorological year	2016	1.470
229921	Zimbabwe	Meteorological year	2017	0.443
229922	Zimbabwe	Meteorological year	2018	0.747
229923	Zimbabwe	Meteorological year	2019	1.359
229924	Zimbabwe	Meteorological year	2020	0.820

229925 rows × 4 columns

After the removal of the columns, the data was re-examined and it was noted that some of the data were of the quarterly changes which might not be required for training the data

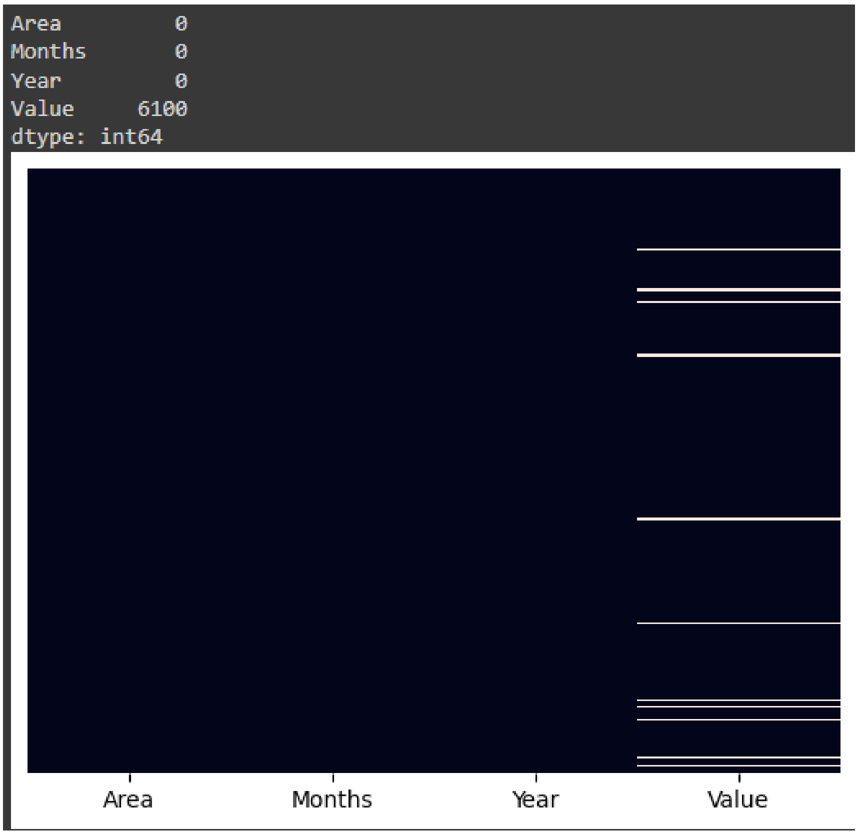
The resulting table:



	Area	Months	Year	Value
0	Afghanistan	January	1961	0.746
1	Afghanistan	January	1962	0.009
2	Afghanistan	January	1963	2.695
3	Afghanistan	January	1964	-5.277
4	Afghanistan	January	1965	1.827
...	...	...	...	...
229920	Zimbabwe	Meteorological year	2016	1.470
229921	Zimbabwe	Meteorological year	2017	0.443
229922	Zimbabwe	Meteorological year	2018	0.747
229923	Zimbabwe	Meteorological year	2019	1.359
229924	Zimbabwe	Meteorological year	2020	0.820

75825 rows × 4 columns

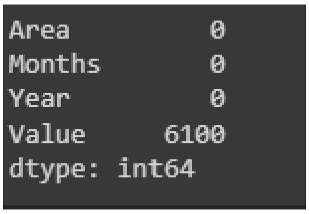
Then, using the heatmap from the seaborn library and the isnull() function from the numpy library, missing values were discovered in the dataset.



```
Area         0
Months       0
Year         0
Value     6100
dtype: int64
```

To handle these missing values, pandas provide us with several techniques such as b-fill, f-fill or just dropping the entire row.
But the **Value** column is a sensitive value to the model as it requires the values for all the months in the year to properly compute the pattern in the data.

Firstly, locate where in the dataset there exists no 'value' for 'meteorological year' in months and delete the entire year from the data for that particular area since, in this dataset, every year in every area has a value for the meteorological year when there is sufficient data to calculate the value for the meteorological year otherwise, they cannot produce any meaningful information or be of any help in filling up the missing values in the dataset



```
Area          0
Months        0
Year          0
Value      6100
dtype: int64
```

After this operation, there remains 6100 missing data and about 54,100 records have been eliminated from the dataset which could have produced improper results.

The rest of the values can be filled using the kth nearest neighbors imputer to get a value that closely matches the values of the same month from different years for the missing values.
But the categorical columns need to be encoded into numerical values for the KNN imputer to interpret the values,

The Non-numerical values are:

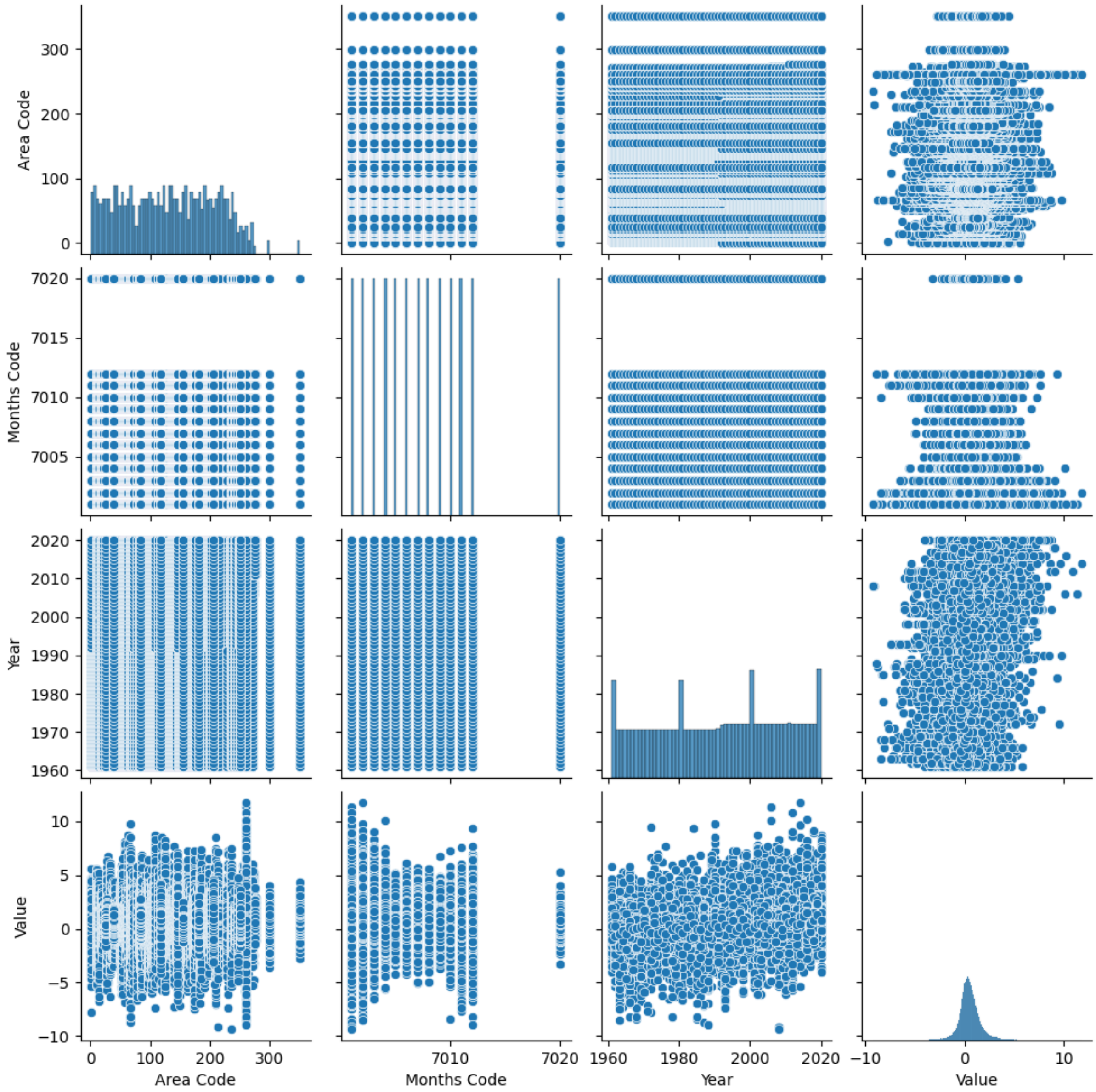
- Area
- Months

Since the dataset already contains encoded formats of these columns. The old columns are replaced with the coded columns and then the data is fed into the KNN imputer.

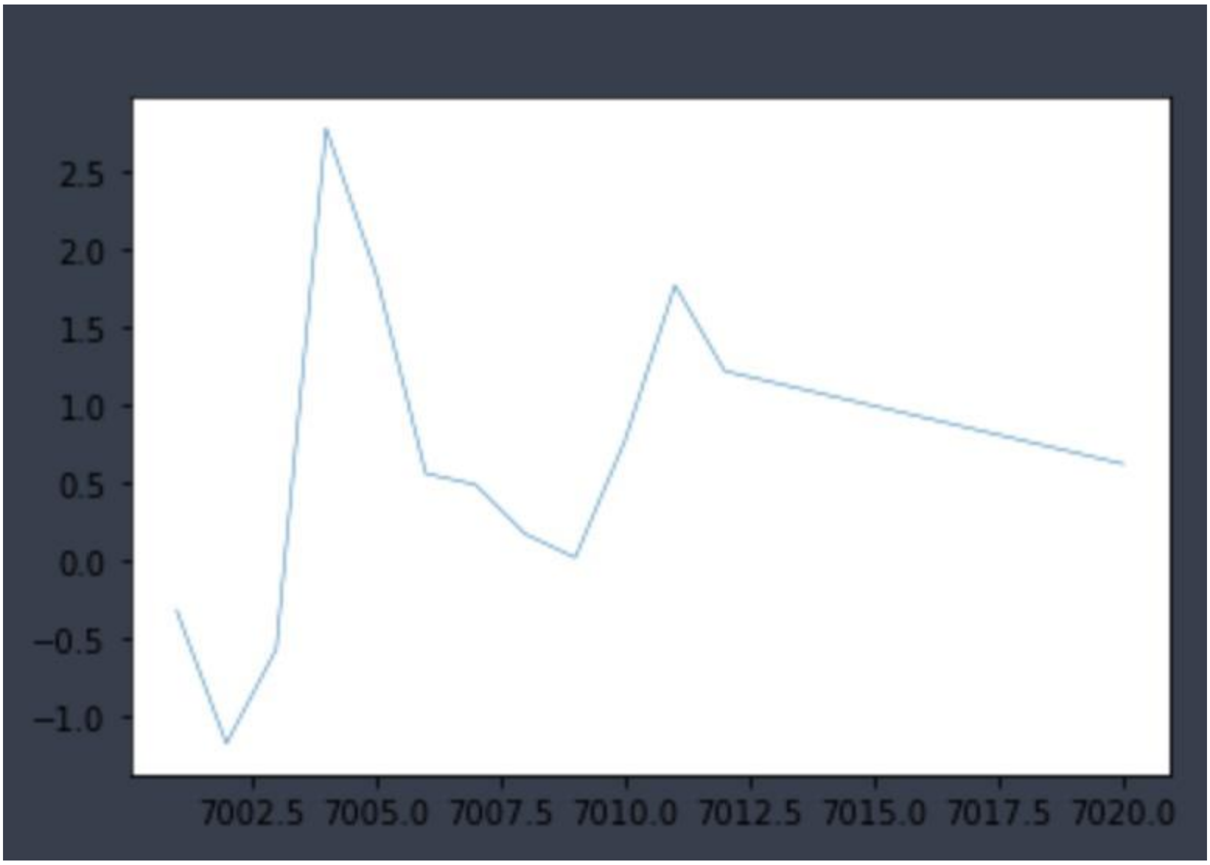
Now that all missing values have been handled and the variability between each pair of columns in the dataset is tested using the seaborn pairplot() function:

Exploratory Data Analysis

Using the pairplot() function of the seaborn library, the following inferences were drawn:



Graph showing the variance of the changes in the mean surface temperature with the changing months in a year:



- Some areas experienced higher changes in the mean surface temperature over the years than others.
- Some months showed a higher rise in the mean surface temperature than the rest.

Thus, it was concluded that the changes in the surface temperature varied and depended on the time of the year and the area.

Model Selection

Model selection is a critical step in the project, as it determines the algorithm or combination of algorithms that will be used to build the predictive model. While the specific models chosen for this project are not mentioned, I can provide you with an overview of common model selection techniques that are often employed in data analysis and predictive modelling tasks.

1. Linear Regression

2. Sine Regression

3. Decision Trees

4. Random Forests

5. Support Vector Machines (SVM)

6. K-Nearest Neighbors (KNN)

7. XGBoost

8. Deep Neural Networks

The choice of models depends on several factors, including the problem requirements, the available data, computational resources, and the trade-off between model performance and interpretability. It is common to experiment with multiple models, compare their performance using evaluation metrics, and select the one that best suits the project's objectives.

Without further information on the specific models selected in the project, it's challenging to provide more detailed insights. However, the models mentioned above are commonly used and can serve as a starting point for model selection in climate change prediction tasks.

MODEL TRAINING AND EVALUATION:

During the model training and evaluation phase, the selected models were trained using the prepared dataset. The dataset was likely divided into two subsets: a training set and a testing set. The training set was used to train the models, while the testing set was held back to evaluate the models' performance on unseen data.

To train the models, the input features from the dataset were used to predict the target variable, which in this case would be the future changes in mean surface temperature. The models were trained to learn the underlying patterns and relationships between the input features and the target variable.

After the models were trained, they were evaluated using performance evaluation metrics. Commonly used metrics for regression tasks include:

1. Mean Absolute Error (MAE)

2. Mean Squared Error (MSE)

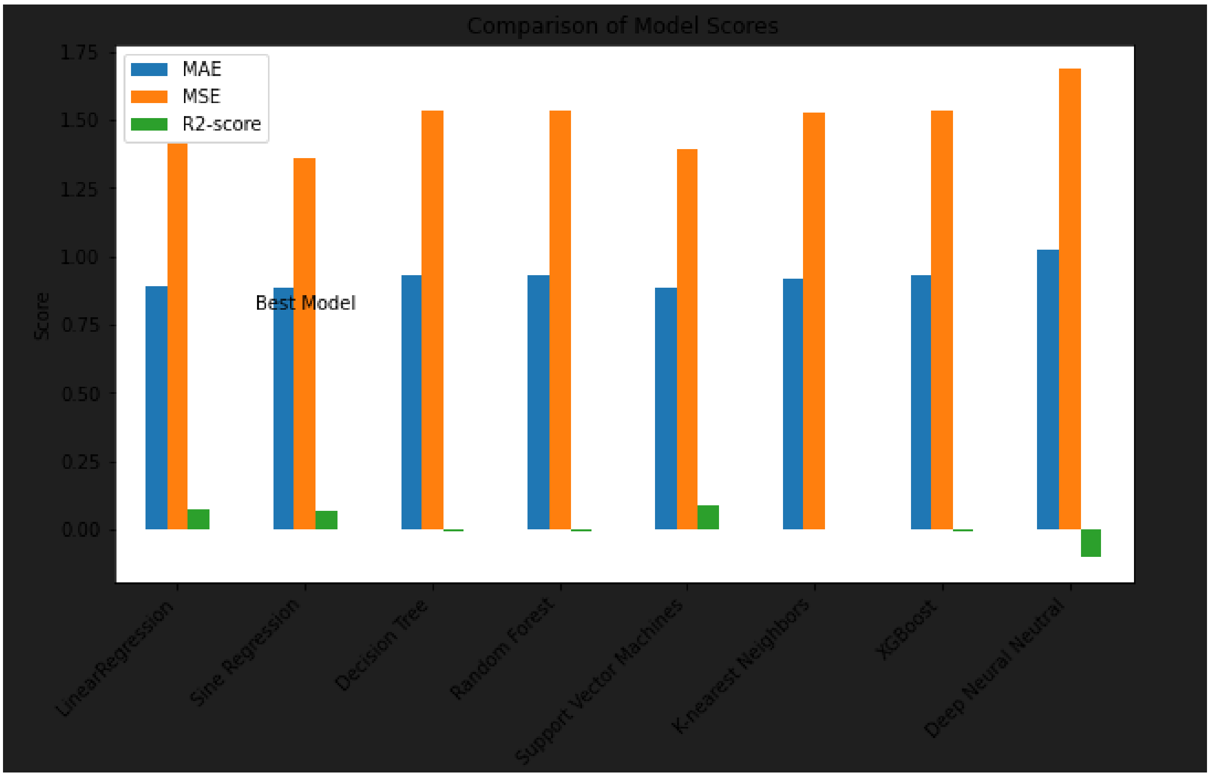
3. R2 Score

These metrics allow for a quantitative assessment of the models' accuracy in making predictions. By comparing the models' performance on the testing set using these evaluation metrics, it becomes possible to identify which model performs better and is more suitable for the given problem.

It is important to note that model training and evaluation is an iterative process. Different models, hyperparameters, and feature engineering techniques may be explored to improve the models' performance. Cross-validation techniques, such as k-fold cross-validation, can also be employed to obtain a more robust estimate of the models' performance.

By evaluating the models using appropriate metrics, the project team can assess the models' effectiveness in predicting future changes in mean surface temperature and make informed decisions about which model to select for deployment.

Here is a comparison chart depicting the various metrics for each model that are used for evaluating and choosing the one with the best results:



The sine regression model showed the best results relative to the other models.
Hence, the sine-regression model has been selected for the project.

COMPARISON AND ANALYSIS OF METRICS USING DIFFERENT PARAMETERS (PERFORMANCE OF THE MODEL)

In the comparison and analysis of metrics using different parameters, the performance of the trained models was assessed using various evaluation metrics. These metrics, such as MAE, MSE, or R2 score, provide quantitative measures of how well the models are performing in understanding the data's trends and making accurate predictions.

By calculating these metrics for each model, the performance of different models can be compared and analysed. The purpose is to identify which model performs better in terms of accuracy and ability to capture the underlying patterns in the data.

By comparing these metrics across different models, the project team can identify which model is performing better in terms of accuracy, precision, and capturing the patterns in the data. This comparison helps in selecting the most suitable model for the project's objectives.

In addition to the metrics mentioned above, other evaluation metrics and techniques can also be employed depending on the specific requirements and characteristics of the project. It's important to consider the trade-offs between different metrics and choose the ones that align with the project goals and the nature of the data.

Overall, the comparison and analysis of metrics using different parameters provide valuable insights into the models' performance, enabling the project team to make informed decisions about the most suitable model to use for predicting future changes in mean surface temperature.

Conclusion:

In conclusion, the project aimed to automate the analysis of climate change data and predict future changes in mean surface temperature. The following key findings and outcomes were obtained:

1. Data Acquisition: The project utilized a dataset obtained from Kaggle, consisting of mean surface temperature change statistics by country. The dataset covered the period 1961-2019 and provided monthly, seasonal, and annual temperature anomalies.

2. Data Cleansing and Transformation: The dataset underwent a cleansing process to remove irrelevant data and handle missing values. Columns that were not required for the regression model were dropped, and missing values were filled using techniques like kth nearest neighbors imputation.

3. Exploratory Data Analysis: Through visualizations such as heatmaps and seaborn's pairplot, exploratory data analysis revealed variations in mean surface temperature changes across different areas and months. This provided valuable insights into the patterns and relationships within the data.

4. Model Selection: While specific models were not mentioned, the project involved selecting appropriate models for predicting mean surface temperature changes. Common model selection techniques include linear regression, decision trees, random forests, support vector machines, k-nearest neighbors, XGBoost, and deep learning models.

5. Model Training and Evaluation: The selected models were trained using the prepared dataset, and their performance was evaluated using metrics such as mean absolute error (MAE), mean squared error (MSE), or R2 score. These metrics assessed the models' accuracy in predicting future changes in mean surface temperature.

6. Comparison and Analysis of Metrics: The performance of the trained models was compared using different evaluation metrics to identify the model that best understood the data's trends and provided accurate predictions.

While specific details on the comparative analysis and the chosen models were not provided, the project's findings would contribute to a better understanding of climate change patterns and help raise awareness about its impacts. The deployment details and specific recommendations for action or future research were not included in the given project description.

Overall, the project aimed to utilize machine learning techniques to analyze climate change data and make predictions regarding mean surface temperature changes. By automating the analysis process, the project sought to contribute to the understanding of climate change and support decision-making processes in addressing its effects.