# ML based Analysis of the Impact of Agri-Food Industry on Temperature Change

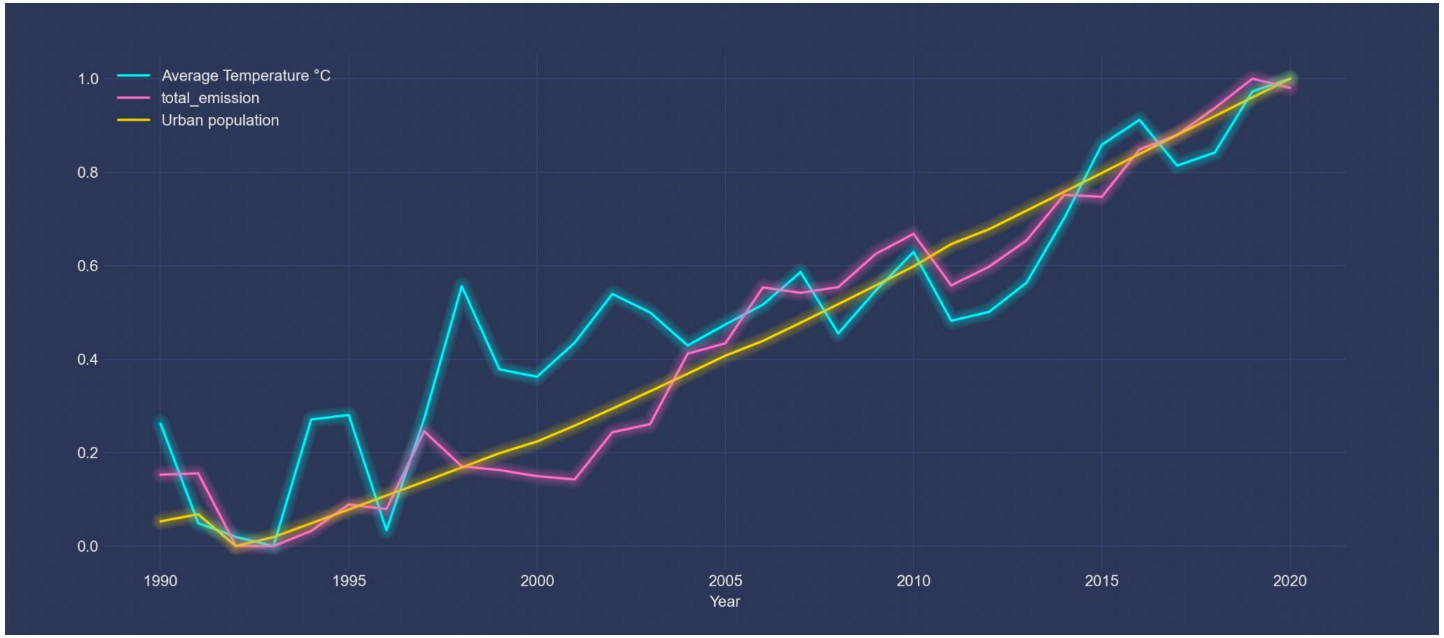
Ritwik Harit, 2021557
Vasan Vohra, 2021572
Tanmay Singh, 2021569
Shreyas Kabra, 2021563

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
**DELHI**

# Motivation

The increasing global population demands growth in the agri-food industry. Understanding and addressing the environmental impact of the agri-food industry is crucial for mitigating climate change and developing sustainable practices within this sector.

The motivation for this project is the need to develop tools for predicting and managing the temperature change caused by CO2 emissions. By harnessing the power of ML, the project aims to create a predictive model that will enable stakeholders and anyone connected to the agri-food industry to make informed decisions and reduce carbon emissions.



# Literature Review

Machine learning and deep learning have gained immense attention for their ability to unveil insights from data in diverse fields like image recognition, natural language processing, healthcare, agriculture, and more. Notably, they've played a pivotal role in climate science, especially in predicting global temperature changes.

In the paper titled "A Machine Learning-Based Model for Predicting Temperature Under the Effects of Climate Change" by Mahmoud Y Shams et al. (2023), the authors explored global climatic patterns. Using the "Climate Change" dataset, they employed various models such as Linear Regression, Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine, and Cat Boost Regressor. Their results indicated that the Cat Boost Regressor (CBR) Model outperformed others, achieving a 0.003 Mean Squared Error, 0.054 Root Mean Squared Error, 0.0036 Mean Absolute Error, and an R2 score of 0.92.

Other significant studies include "Monthly prediction of air temperature in Australia and New Zealand with machine learning models" by S. Salcedo-Sanz et al. (2016) and "Climate Change Analysis Using Machine Learning" by Himanshu Vishwakarma (2018) . These studies explored temperature prediction using Support Vector Regressor (SVR), Multi-layer Perceptron (MLP), Linear Regression, SVR, Lasso Regression, and Elastic Net based on historical climate and greenhouse gases data.

# Dataset Details

The Agri-food CO2 emission dataset  available on Kaggle has been curated by combining and meticulously processing multiple distinct datasets sourced from the Food and Agriculture Organization (FAO) and data provided by the Inter governmental Panel on Climate Change (IPCC).

As shown by the dataset, these emissions make a significant and noteworthy contribution to the annual global emissions.

The dataset contains 6965 rows and 31 columns, including 30 distinct features and 1 target column.

The target column, labelled “Average Temperature °C“, indicates the yearly average temperature rise.

# Dataset Details – Key Features

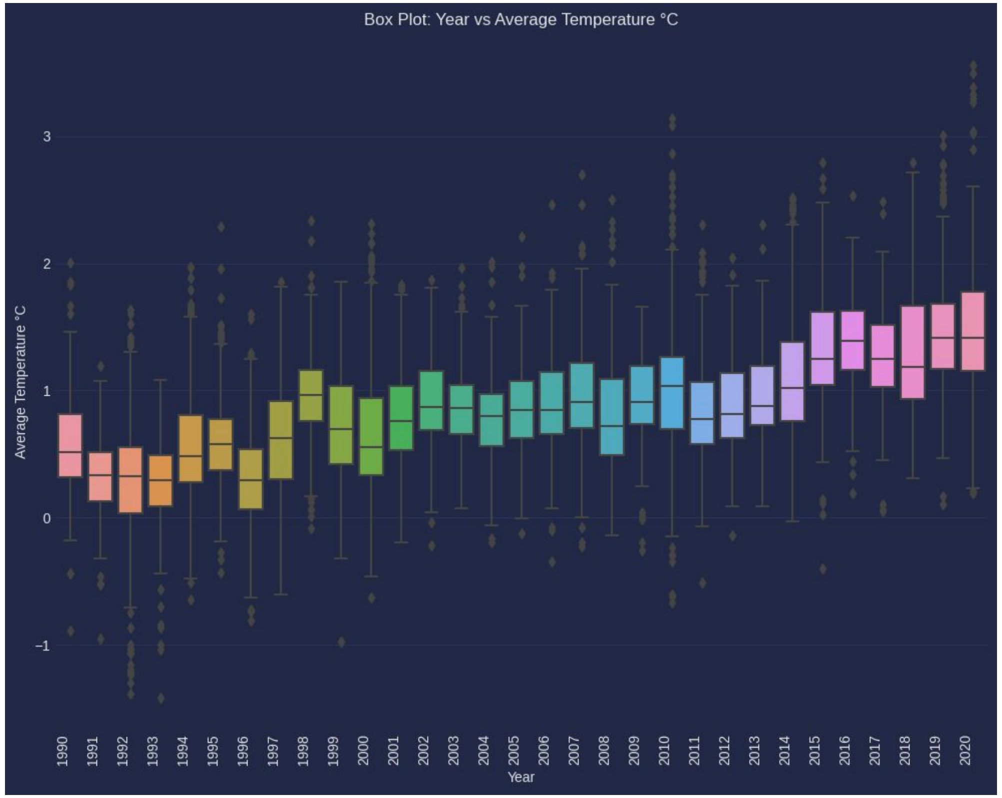
1. **Area:** Denotes the respective country.
2. **Year:** Represents the specific year of data.
3. **Savanna Fires:** Reflects emissions resulting from fires occurring in savanna ecosystems.
4. **Forest Fires:** Represents emissions originating from fires within forested regions.
5. **Crop Residues:** Signifies emissions from the combustion or decomposition of residual plant material post harvesting.
6. **Rice Cultivation:** Indicates emissions stemming from methane release during rice cultivation.
7. **Drained Organic Soils (CO2):** Quantifies emissions linked to carbon dioxide release during organic soil drainage.
8. **Pesticides Manufacturing:** Quantifies emissions associated with pesticide production.
9. **Food Transport:** Measures emissions resulting from the transportation of food products.
10. **Forestland:** Specifies the land area covered by forests.
11. **Net Forest Conversion:** Depicts changes in forest areas resulting from deforestation and afforestation.
12. **Food Household Consumption:** Records emissions attributable to food consumption at the household level.
13. **Food Retail:** Encompasses emissions generated by the operation of food retail establishments.
14. **On-farm Electricity Use:** Represents electricity consumption on agricultural farms
15. **Food Packaging:** Quantifies emissions arising from producing and disposing of food packaging materials.
16. **Agri-food Systems Waste Disposal:** Represents emissions emanating from waste disposal within the agri-food system.

# Dataset Details – Key Features

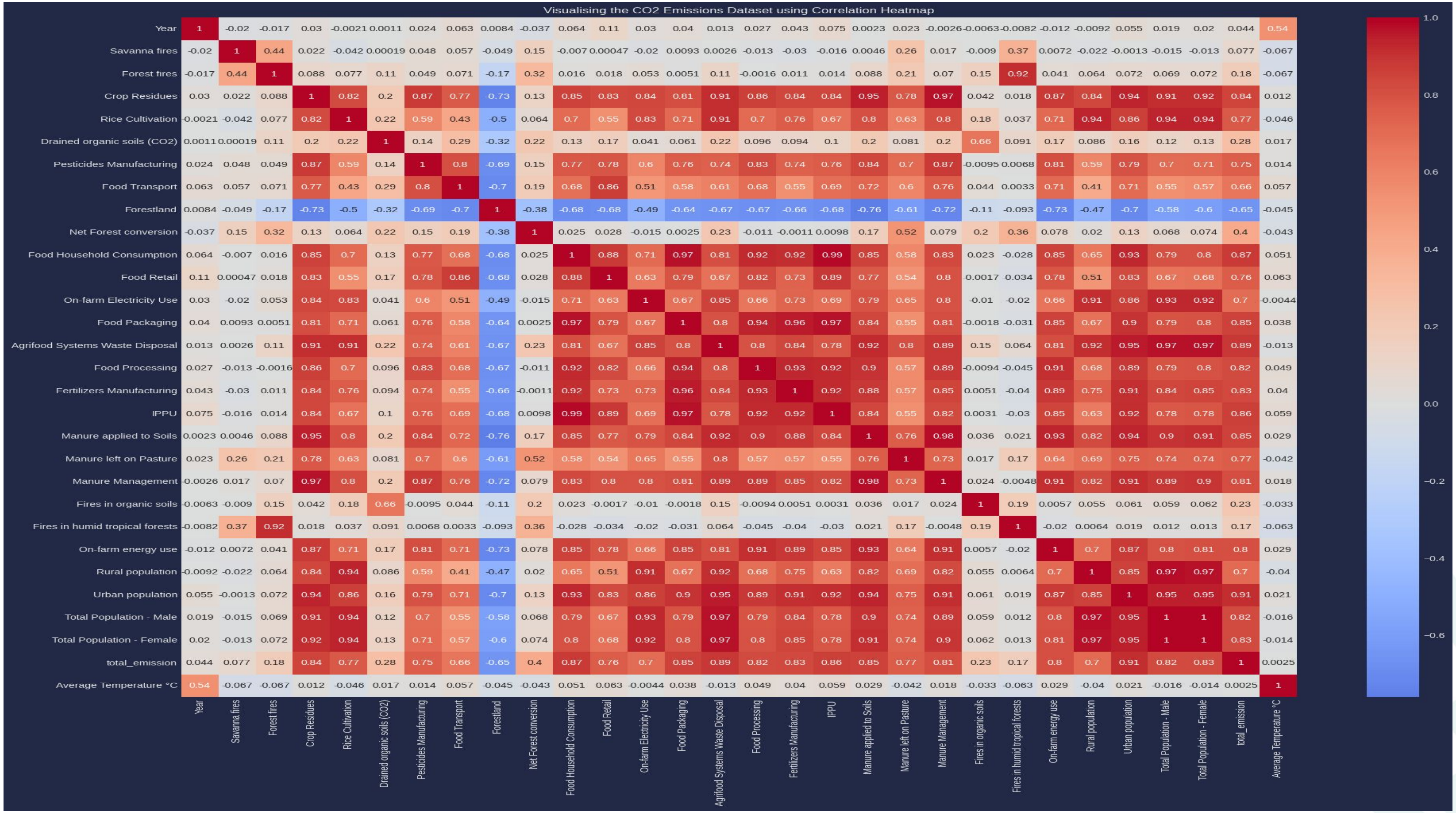
17. **Food Processing:** Represents emissions associated with the processing of food products.
18. **Fertilizers Manufacturing:** Quantifies emissions linked to the production of fertilisers.
19. **IPPU (Industrial Processes & Product Use):** Encompasses emissions originating from industrial processes & product use.
20. **Manure Applied to Soils:** Reflects emissions from animal manure's application to agricultural soils.
21. **Manure Left on Pasture:** Quantifies emissions arising from the presence of animal manure on pasture or grazing land.
22. **Manure Management:** Indicates emissions related to the management and treatment of animal manure.
23. **Fires in Organic Soils:** Captures emissions generated by fires occurring in organic soils.
24. **Fires in Humid Tropical Forests:** Measures emissions resulting from fires in humid tropical forests.
25. **On-farm Energy Use:** Represents energy consumption on agricultural farms.
26. **Rural Population:** Signifies the number of individuals residing in rural areas.
27. **Urban Population:** Represents the number of individuals residing in urban areas.
28. **Total Population – Male:** Quantifies the total male population.
29. **Total Population – Female:** Quantifies the total female population.
30. **Total Emission:** Aggregates total greenhouse gas emissions from diverse sources.

# Dataset Details – Target Column

**Average Temperature °C:** This column records the average annual temperature increase in degrees Celsius, serving as the dataset’s primary target variable for analysis and prediction.



# Correlation Heatmap



# Dataset Details- Preprocessing Techniques

- **Handling Null Values:** Rows containing Null values were removed from the dataset to ensure data integrity and prevent potential bias in the analysis.
- **Eliminating Duplicate Rows:** Duplicate rows were identified and removed from the dataset to avoid redundancy and ensure the accuracy of the analysis.
- **Feature Removal:** Features exhibiting a high correlation, represented by a correlation coefficient greater than or equal to 0.99, were identified and removed. This step helps in improving the model's interpretability and generalization.
- **Encoding for Categorical Features:** We used 2 techniques for encoding the categorical features:
    - **Label Encoding:** Used Label Encoding to transform the categorical feature "Area" into a numerical format for modelling and analysis.
    - **One-Hot Encoding:** The categorical feature "Area" was transformed using One-Hot encoding. This technique creates binary columns for each category, allowing machine learning algorithms to work effectively with categorical data.
- **Standard Scaling:** Standard scaling was used to standardize the data. Standardization helps ensure that features with different units and scales do not disproportionately influence the modeling process and helps machine learning algorithms converge faster.

# Dataset Details- Preprocessing Techniques

- **Dimensionality Reduction for One-Hot Encoded dataset :** We utilized Principal Component Analysis (PCA) to reduce feature dimensionality from 183 to 160. We reduced the features to 160 because the graph drawn for *Explained Variance* vs *Number of Components* (see Fig. 1) shows that 160 features are able to explain more than 95% of the variance in the data. This process improved efficiency, increased interpretability, and lowered computational complexity in our analysis.

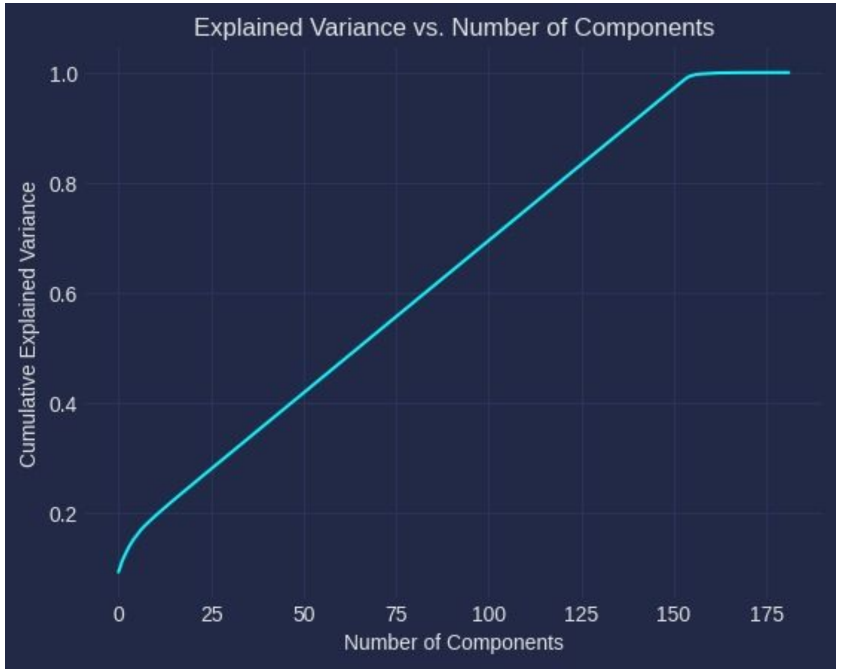
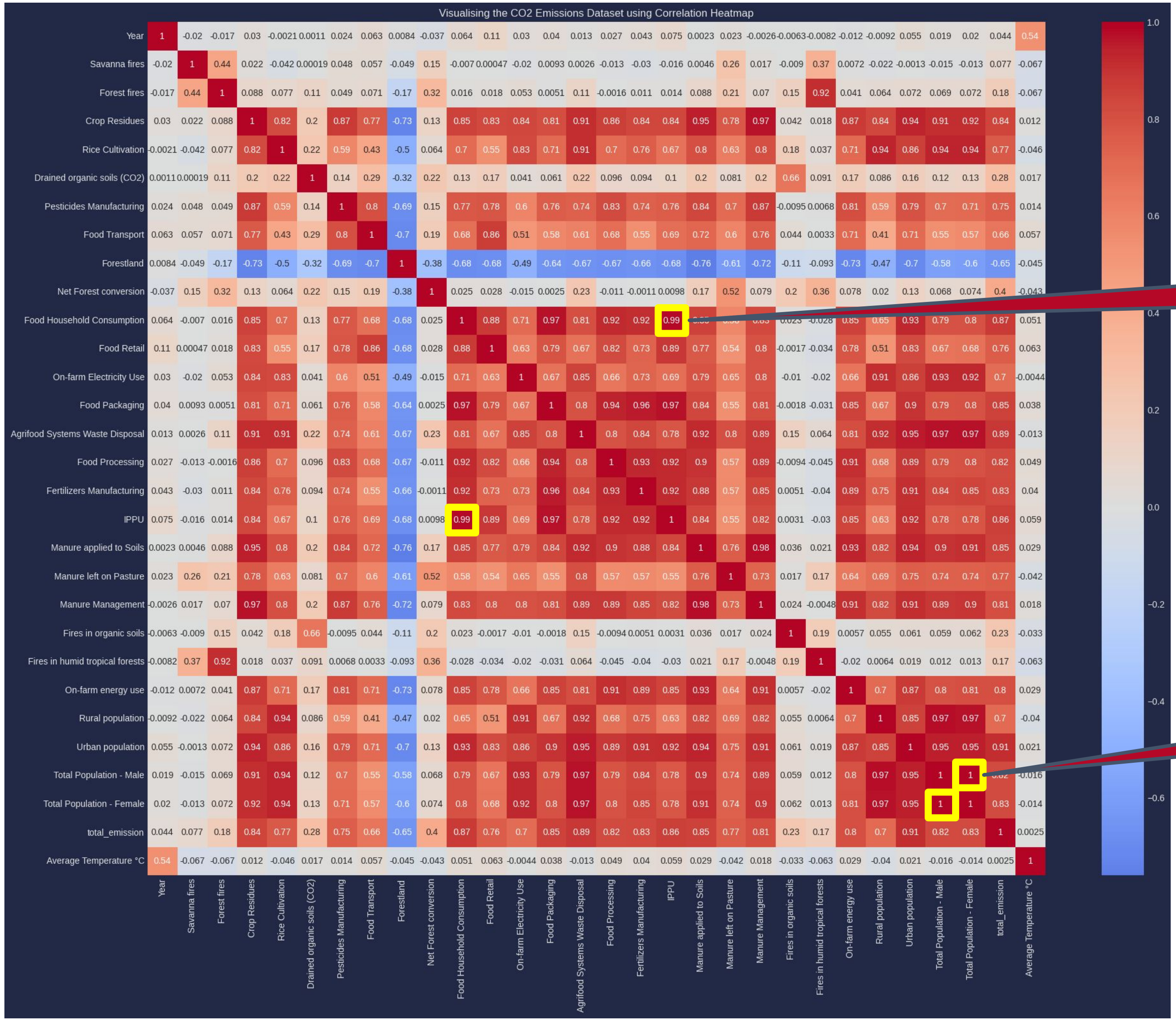



FIG1 : Explained Variance V/S Number of Components

# Correlation Heatmap




IPPU vs Food HouseHold Consumption

0.99

Total Population - Female vs Total Population - Male

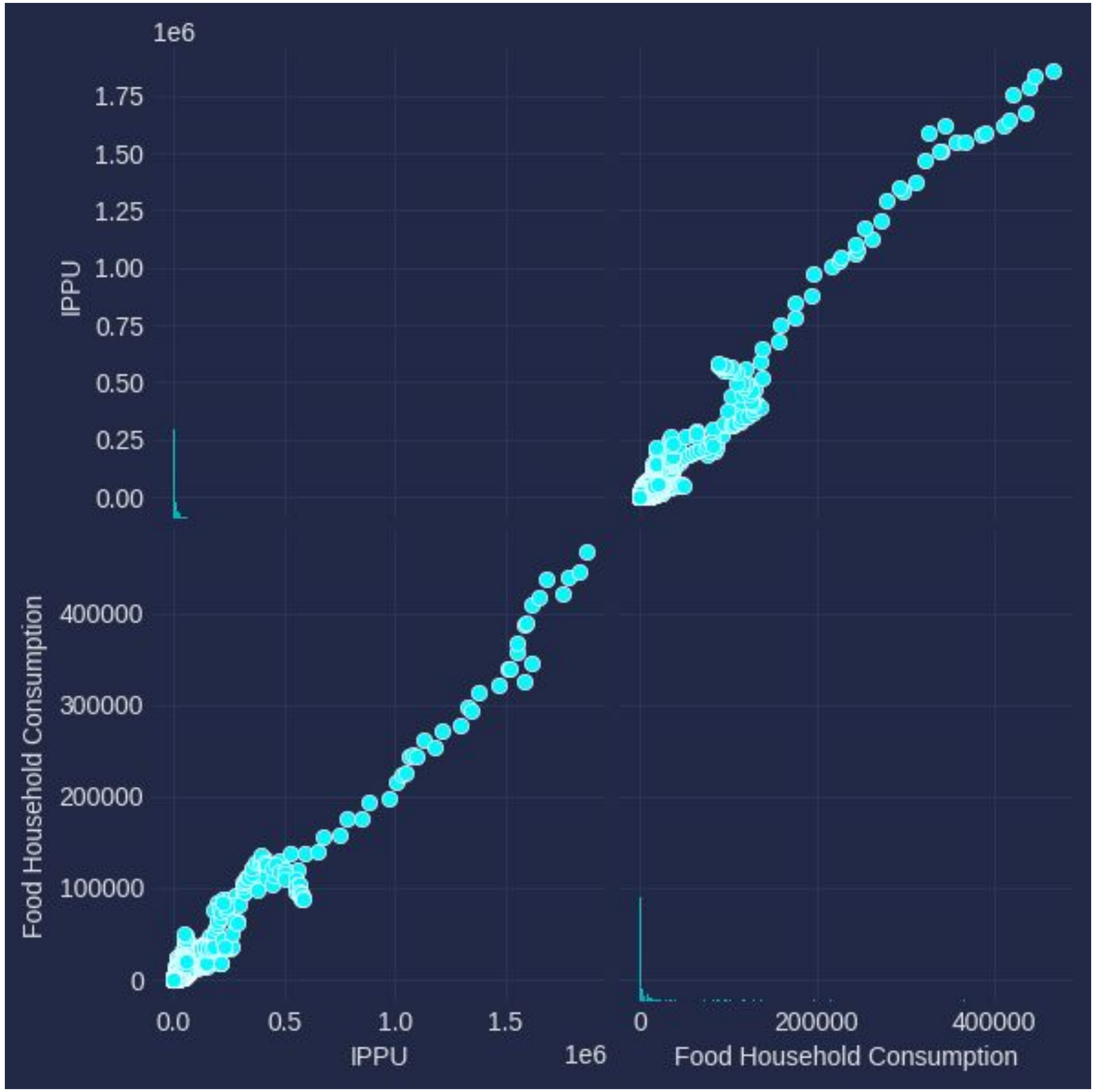
1

# Methodology

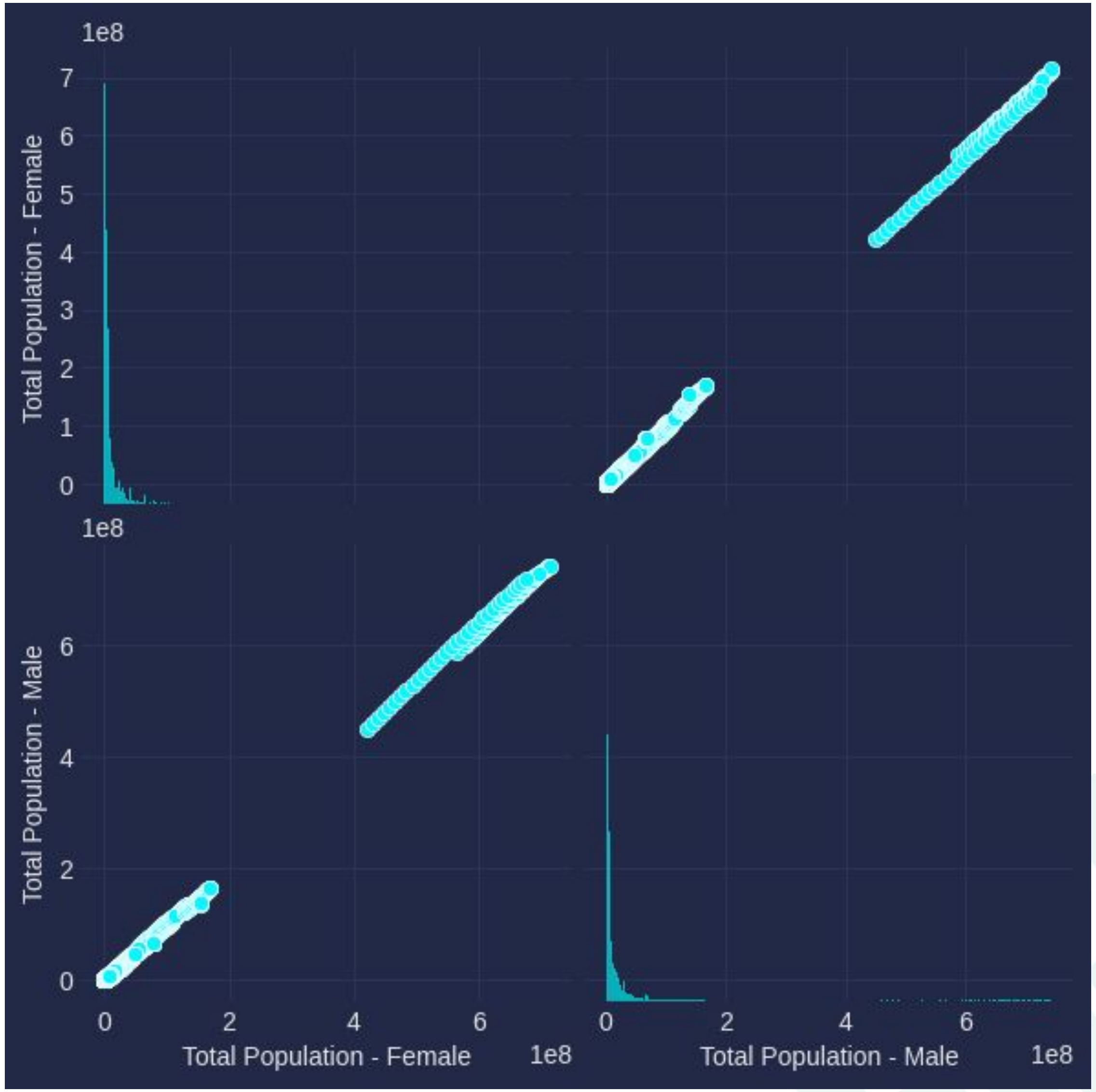
## Data Preprocessing

- Removed rows containing Null values.
- Eliminated duplicate rows.
- Created various graphs, including scatter plots, histograms, boxplots, pie charts, and t-SNE graphs, for data visualization and analysis.
- We removed features *IPPU* and *Total Population-Female* due to their high correlation (greater than or equal to 0.99) with *Food Household Consumption* and *Total Population-Male*, respectively, as identified through correlation heatmaps(prev slide) and pair plots (next slide).
- We created 3 scenarios by applying the following techniques techniques for encoding the categorical features:
  - **Scenario 1:** Label Encoding to transform the categorical feature into the numerical format.
  - **Scenario 2:** Applied one-hot encoding to handle categorical features, increasing the number of features to 183.
  - **Scenario 3:** Used one-hot encoding and subsequently, we attempted to reduce the feature set from 183 to 160 using PCA.

# Pairplots



1e6
IPPU
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Food Household Consumption
400000
300000
200000
100000
0
0.0
0.5
1.0
1.5
1e6
IPPU
0
200000
400000
Food Household Consumption

1e8
Total Population - Female
7
6
5
4
3
2
1
0
1e8
Total Population - Male
6
4
2
0
0
2
4
6
Total Population - Female
1e8
0
2
4
6
Total Population - Male
1e8

# Methodology

## Model Validation

Utilized K-Fold cross-validation with K=5 to validate machine learning models. This approach helps assess the model's performance across different subsets of the data.

## Evaluation Metrics

Used Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 score as evaluation metrics for both training and validation. These metrics provide insights into the accuracy and performance of the models.

# Methodology

## Machine Learning Models Used

1. Random Forest Regressor
2. Linear Regressor
3. Gradient Boosting Regressor
4. Adaboost Regressor
5. XGB Regressor
6. Lasso Regressor
7. Ridge Regressor
8. Support Vector Regression (SVR) with various kernels:
    a. RBF kernel
    b. Linear kernel
    c. Polynomial kernel
    d. Sigmoid kernel

# Result & Analysis

The results of various evaluation metrics for different models and preprocessing techniques (Label Encoding and One-Hot Encoding) are provided in Tables 1, 2, and 3. The evaluation metrics used include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 Score for both training and validation sets.

**Type of Data used to evaluate Model Performance:**

- **Scenario 1:** Label Encoded Data
- **Scenario 2:** One-Hot Encoded Data
- **Scenario 3:** One-Hot Encoded Data with PCA (Principal Component Analysis, a dimensionality reduction algorithm)

# Result & Analysis – Scenario 1

**Random Forest Regressor** and **XGBRegressor** performed the best, showing low errors on both training and validation sets.

Their high R2 score indicates they effectively captured the data patterns, highlighting their strong predictive ability.

Support Vector Regression with Sigmoid Kernel is the worst performer, displaying extremely high training and validation errors and negative R2 scores on the training set (-17463.75), indicating an inappropriate fit.

Moreover, the negative R2 score on the validation set (-18086.22) emphasizes the model's inability to learn from the data.

| **Table** | **Training** | | | **Validation** | | |
|---|---|---|---|---|---|---|
| **Models** | ***MAE*** | ***RMSE*** | ***R2 Score*** | ***MAE*** | ***RMSE*** | ***R2 Score*** |
| RFR | 0.103 | 0.141 | 0.940 | 0.276 | 0.375 | 0.581 |
| LR | 0.356 | 0.471 | 0.339 | 0.359 | 0.474 | 0.330 |
| GBR | 0.265 | 0.356 | 0.622 | 0.293 | 0.394 | 0.536 |
| ABR | 0.342 | 0.436 | 0.433 | 0.350 | 0.454 | 0.386 |
| XGBR | 0.088 | 0.116 | 0.959 | 0.277 | 0.373 | 0.585 |
| LassoR | 0.356 | 0.472 | 0.338 | 0.358 | 0.474 | 0.329 |
| RidgeR | 0.356 | 0.471 | 0.339 | 0.358 | 0.474 | 0.330 |
| SVR (RBF) | 0.330 | 0.446 | 0.408 | 0.342 | 0.456 | 0.379 |
| SVR (Linear) | 0.354 | 0.474 | 0.331 | 0.358 | 0.478 | 0.319 |
| SVR (Poly) | 0.357 | 0.482 | 0.309 | 0.403 | 1.139 | -7.049 |
| SVR (Sigmoid) | 22.134 | 76.705 | -17463.75 | 22.380 | 76.825 | -18086.22 |

TABLE I: Evaluation Metrics for Label-Encoded Data

# Result & Analysis – Scenario 2

**Random Forest Regressor** and **XGBRegressor** continue to perform as the best models in our analysis.

There isn't a significant difference observed between the results of models on label encoding and one-hot encoded data. One-hot encoding typically leads to a larger feature space due to the creation of binary columns for categorical data.

However, this increase in feature complexity has not yielded better results for our models; instead, it has added complexity without a corresponding improvement in performance.

With the increased number of features, Linear regression becomes the worst performer.
The huge difference between the training and validation scores shows that it overfits the data.

| **Table** | **Training** | | | **Validation** | | |
|---|---|---|---|---|---|---|
| **Models** | ***MAE*** | ***RMSE*** | ***R2 Score*** | ***MAE*** | ***RMSE*** | ***R2 Score*** |
| RFR | 0.102 | 0.141 | 0.940 | 0.277 | 0.375 | 0.580 |
| LR | 0.295 | 0.399 | 0.525 | 2.322e+8 | 4.952e+9 | -1.77e+20 |
| GBR | 0.272 | 0.364 | 0.604 | 0.297 | 0.398 | 0.527 |
| ABR | 0.345 | 0.438 | 0.428 | 0.352 | 0.457 | 0.377 |
| XGBR | 0.109 | 0.144 | 0.937 | 0.277 | 0.374 | 0.583 |
| LassoR | 0.294 | 0.400 | 0.524 | 0.308 | 0.418 | 0.478 |
| RidgeR | 0.294 | 0.399 | 0.525 | 0.308 | 0.418 | 0.478 |
| SVR (RBF) | 0.284 | 0.397 | 0.531 | 0.307 | 0.421 | 0.472 |
| SVR (Linear) | 0.292 | 0.404 | 0.515 | 0.310 | 0.422 | 0.469 |
| SVR (Poly) | 0.316 | 0.441 | 0.422 | 0.344 | 0.469 | 0.343 |
| SVR (Sigmoid) | 1.217 | 5.682 | -95.385 | 1.296 | 5.771 | -107.049 |

TABLE II: Evaluation Metrics for One-Hot Encoded Data

# Result & Analysis – Scenario 3

On performing PCA on the one-hot encoded data, the performance of most of the models is reduced.

There is a **significant improvement in the results of Linear Regression** by reducing some features because the model becomes less prone to overfitting, allowing it to generalize better to unseen data.

However, with the reduced performance of our best models, it is not preferred to implement PCA in this particular case, as it negatively impacts the overall predictive power of our models.

| **Table** | **Training** | | | **Validation** | | |
|---|---|---|---|---|---|---|
| **Models** | ***MAE*** | ***RMSE*** | ***R2 Score*** | ***MAE*** | ***RMSE*** | ***R2 Score*** |
| RFR | 0.116 | 0.159 | 0.924 | 0.312 | 0.425 | 0.461 |
| LR | 0.296 | 0.403 | 0.517 | 0.309 | 0.420 | 0.472 |
| GBR | 0.275 | 0.364 | 0.605 | 0.311 | 0.419 | 0.476 |
| ABR | 0.345 | 0.441 | 0.421 | 0.357 | 0.468 | 0.347 |
| XGBR | 0.118 | 0.156 | 0.927 | 0.326 | 0.444 | 0.413 |
| LassoR | 0.296 | 0.403 | 0.517 | 0.309 | 0.420 | 0.473 |
| RidgeR | 0.296 | 0.403 | 0.517 | 0.309 | 0.420 | 0.472 |
| SVR (RBF) | 0.285 | 0.398 | 0.529 | 0.307 | 0.422 | 0.471 |
| SVR (Linear) | 0.294 | 0.407 | 0.507 | 0.311 | 0.424 | 0.464 |
| SVR (Poly) | 0.318 | 0.442 | 0.417 | 0.345 | 0.473 | 0.334 |
| SVR (Sigmoid) | 1.239 | 5.752 | -97.886 | 1.318 | 5.844 | -110.239 |

TABLE III: Evaluation Metrics for One-Hot Encoded Data with PCA

# Conclusion

To sum it up, our analysis showed that using either Label Encoding or One-Hot Encoding doesn't make a big difference in our results.

We found that models like **Random Forest Regressor and XGB Regressor consistently perform well**, making them our top choices.

On the other hand, the simpler Linear Regression didn't work as effectively, especially with more complicated data. This highlights the need for more advanced models that can handle complex patterns in the data.

When we tried simplifying our data using PCA, it helped Linear Regression but hurt other models. So, while these techniques can be helpful, they need to be applied carefully.

In essence, choosing the right model is crucial. Our focus will be on exploring more sophisticated models and fine-tuning them. This approach will lead us to a strong and reliable predictive model tailored to our specific requirements.

# Timeline

1. **Data Exploration and Analysis (1 week)**
    - Preprocessing the data to ensure consistency and quality and conducting initial exploratory data analysis to understand the data's structure and characteristics.
2. **Data Preprocessing (1 week)**
    - Encoding categorical values using appropriate techniques.
    - Handling duplicate rows and missing values.
    - Standardising numerical features.
3. **Feature Engineering (1 week)**
    - Performing Dimensionality reduction (using PCA).
4. **Model Selection, Training and Evaluation (2 weeks)**
    - Using different types of machine learning models suitable for regression tasks.
    - Training the selected models using the training dataset.
    - Evaluating the models using appropriate metrics (RMSE, MAE, etc).
5. **Model Optimisation (1 week)**
    - Optimising hyperparameters using the validation dataset.
6. **Reporting and Presentation (1 week)**
    - Creating a comprehensive report and presentation summarising the project's objectives and methods.

# Future Timeline

1. **Data Preprocessing (1 week)**
    - Handling Outliers.
2. **Feature Engineering (1 week)**
    - Creating new features (through techniques such as clustering) that may improve the model's performance.
    - Performing Dimensionality reduction using different techniques such as LDA, t-SNE, etc.
3. **Model Selection, Training and Evaluation (2 weeks)**
    - Using more complex machine learning models suitable for regression tasks such as neural networks, polynomial regression, etc.
    - Training the selected models using the training dataset.
    - Evaluating the models using appropriate metrics (RMSE, MAE, etc).
4. **Model Optimisation (1 week)**
    - Optimising hyperparameters using techniques such as Grid Search, Bayesian Search, etc.
    - Refining the model through ensemble techniques.
5. **Reporting and Presentation (1 week)**
    - Creating a comprehensive report and presentation summarising the project's objectives and methods.

# Contributions

**Tanmay:** Data Exploration and Analysis, Benchmark model creation and Model Development.
**Shreyas:** Data Preprocessing and Analysis, Feature Engineering and Model Development.
**Ritwik:** Data Preprocessing, Model Development, Training and Evaluation.
**Vasan:** Data Preprocessing, Model Development with Model Optimisation and Refinement.

***<u>Although the tasks have been divided, all four team members have contributed equally towards each task.</u>***

# References


- Agri-food CO2 emission dataset - Forecasting ML
- Mahmoud Y. Shams et al. "A Machine Learning-Based Model for Predicting Temperature Under the Effects of Climate Change". DOI: 10.1007/978-3-031-22456-0 4
- S. Salcedo-Sanz et al. "Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms". DOI: 10.1007/s00704-015-1480-4
- Himanshu Vishwakarma. "Climate Change Analysis Using Machine Learning". DOI: 10.21275/SR20722101621
- Code