Vietnam National University - Ho Chi Minh City

University of Information Technology

REPORT FINAL PROJECT

CS313 - DATA MINING

TOPIC: PREDICT CO2 EMISSION

Teacher: PhD. Vo Nguyen Le Duy

Student:

22521498 - Nguyễn Thị Ngọc Trâm

22521516 - Dương Thành Trí

22521610 - Phạm Nguyễn Anh Tuấn

22521626 - Nguyễn Mạnh Tường

22521671 - Lưu Khánh Vinh

A. PROJECT OVERVIEW

1. Introduction

Climate change and air pollution, largely driven by **CO_2 emissions**, pose significant global environmental challenges. **Monitoring and predicting** these emissions in specific areas is vital for effective environmental protection policies and sustainable development. This project focuses on building a **machine learning model to predict weekly CO_2 emissions** at defined geographical locations. The research aims to provide an effective predictive tool and **enhance our understanding** of the factors influencing CO_2 emissions, thereby supporting environmental management and climate change mitigation efforts. The report will detail the implementation process, results, challenges, and future development directions.

2. Data Description

This dataset is part of the **Kaggle Playground Series 2023 challenge,** designed to predict CO_2 emissions using satellite observations from Sentinel-5P. It includes weekly data collected **between 2019 and November 2022** from around **497 locations in Rwanda**. The dataset contains seven key atmospheric indicators such as **SO_2, CO, NO_2, O_3, HCHO, aerosol, cloud**. The aim is to train machine learning models to predict CO_2 emissions based on historical and atmospheric data.

Feature	**Name of column**
Location	latitude, longitude
Time	year, week_no
Measured Gases & Features	
● Sulfur Dioxide (SO_2)	9 columns related this feature
● Carbon Monoxide (CO)	7 columns related this feature
● Nitrogen Dioxide (NO_2)	11 columns related this feature
● Formaldehyde (HCHO)	8 columns related this feature
● Ozone (O_3)	9 columns related this feature
● Aerosol (UV absorbing index & aerosol layer height)	10 columns related this feature
● Cloud	11 columns related this feature

3. Input-Output

Input:

The training dataset is $\boldsymbol{D = \{(x_i, y_i)\}}$, where:

- $x_i = [x_{i1}, x_{i2}, ..., x_{ip}]$: x_{ij} is the value of the j_{th} feature for observation i, and p is the total number of features.
- y_i: The CO_2 emission value corresponding to observation i ($y_i \in \mathbb{R}$).
- $x_{predict} = [x_{predict1}, x_{predict2}, ..., x_{predictp}]$

Output:

- $\widehat{y}_{predict}$: The predicted CO_2 emission value for the input observation $x_{predict}$, produced by the model f ($\widehat{y}_{predict} \in \mathbb{R}$).

B. METHODOLOGY & EXPERIMENTS

1. Exploratory Data Analysis (EDA)

Perform EDA to gain a better understanding of the characteristics, distributions, and relationships among data features.

Initial Data Inspection

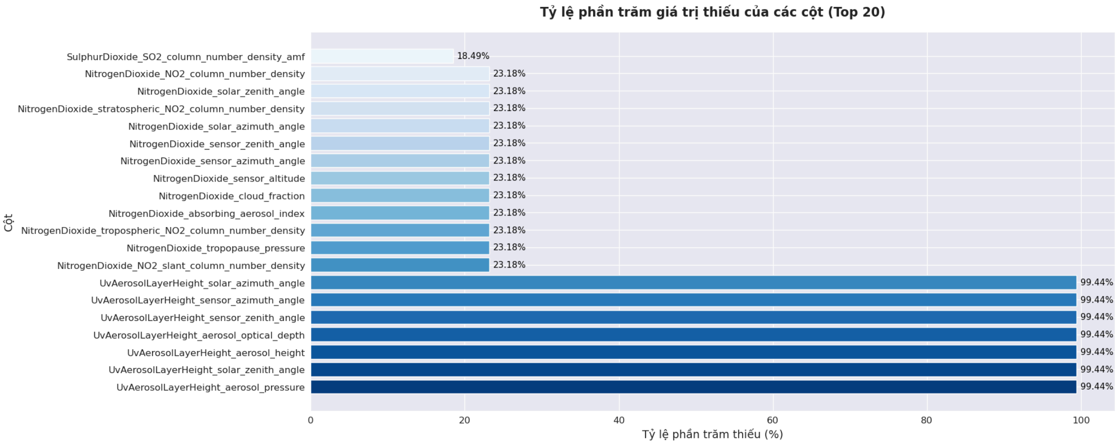
The dataset includes 497 geographical points defined by coordinate pairs (latitude-longitude). These points are present in both the training and test sets. Specifically, for each geographical point:

- **159 rows** of data in the **training set**, corresponding to **3 years (2019, 2020, 2021)**, with each year consisting of **53 weeks numbered from 0 to 52**.
- **49 rows** of data in the **test set**, corresponding to weeks from **0 to 48 of the year 2022**.
- The total number of training data rows: **79,023**.
- The total number of test data rows: **24,353**.

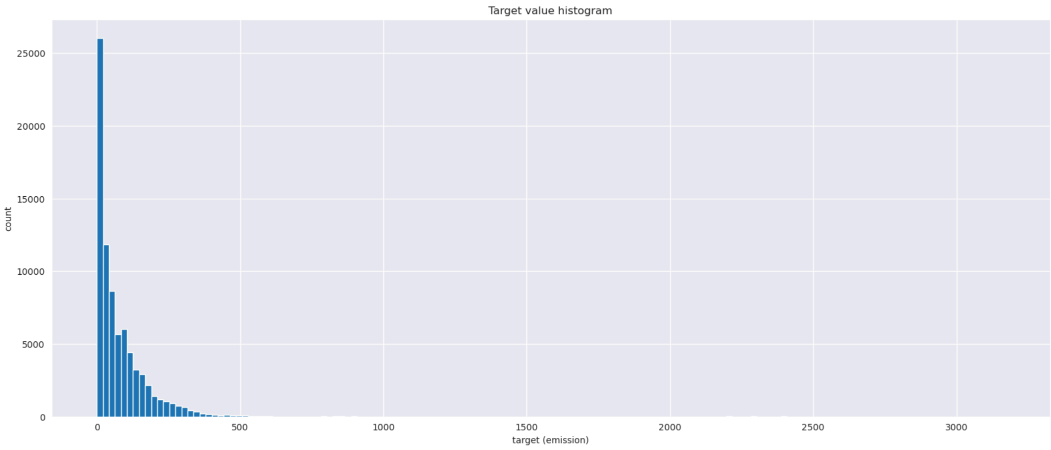
Data Quality Assessment

Initial inspection reveals a significant number of missing values (NaNs) in several important columns related to emissions and environmental indices, such as **Nitrogen Dioxide (NO_2)**, **Formaldehyde (HCHO)**, and **Ozone (O_3)**. Specifically:

- Column **SulphurDioxide_SO2_column_number_density** has approximately **18.5%** missing values.
- Column **NitrogenDioxide_NO2_column_number_density** contains a substantial amount of missing values (~**23%**).
- **Formaldehyde_tropospheric_HCHO_column_number_density** has about **9.2%** missing values.
- **Ozone_O3_column_number_density** has fewer missing values (~**0.7%**).
- Features within the **UVAeroSollayerHeight** group have more than **99%** missing values.



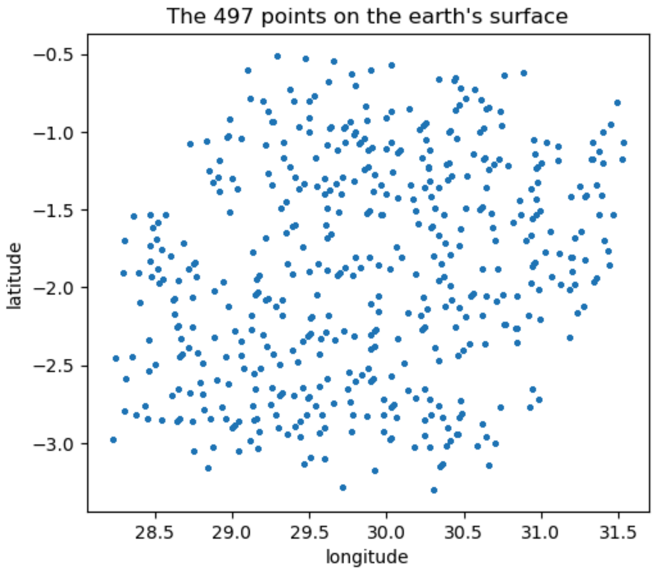
Descriptive Statistical Analysis



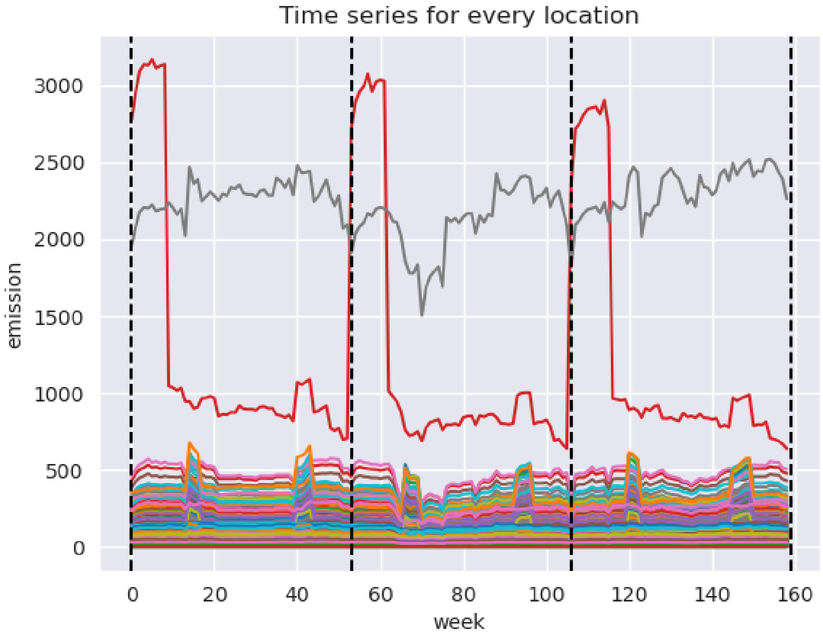
- Most emission indices do not exceed a value of 200.

Data Visualization

Geographical Distribution of Measurement Points:



Time Series of Emissions:



- At most locations, CO_2 emissions decreased from 2019 to 2020, likely due to the impact of the COVID-19 pandemic. However, emissions rose again from 2020 to 2021.

Conclusions

- All features are numerical, with no categorical data.
- Features (excluding location columns) and the target variable can be grouped into **eight main categories**, facilitating organization and analysis.
- The dataset comprises **497 distinct geographical points**.
- CO_2 emission values span an extensive range, **from 0 to 3200**.
- The **UV Aerosol** feature group has a high rate of missing data and exhibits significant distribution discrepancies between the training and test sets, and should thus be removed.
- CO_2 emissions are strongly influenced by seasonal factors and holidays, highlighting the critical role of time.
- Two locations with exceptionally high emission values require special attention.
- The test set consists of data collected from the same locations as the training set, covering the first 49 weeks of 2022, which are the prediction target.

2. Data Preprocessing

2.1 Handling missing values

- **Dropping columns with high missing values**: to ensure data quality and reduce noise, all columns related to UvAerosolLayerHeight with missing value percentages over 99.44% were dropped from the dataset. Removing these features helps avoid introducing bias or error from excessive imputation.
- **Imputing missing values with mean:** for the remaining columns with 18–23% missing values, mean imputation was applied. This mainly includes sulfur dioxide and nitrogen dioxide features, helping preserve data distribution and ensure model compatibility.

2.2 Normalization

Standard Scaler is a data normalization technique that transforms each feature to have a mean of 0 and a standard deviation of 1. It is especially useful for models sensitive to feature scales, such as linear regression.

$$z = \frac{x - \mu}{\sigma}$$

Where: x - original value, $\boldsymbol{\mu}$ - mean of the feature, **σ** - standard deviation of the feature

2.3 Time Transformation

- Created a **date** column from **year** and **week_no** for time-based processing.
- Extracted **season** based on the **month**.
- Marked holiday weeks often change routines, reducing travel and production, which can affect emissions.
  - **Week 0:** The first week of the year, around New Year's Day
  - **Week 51:** The last week of the year, during Christmas and New Year's holidays.
  - **Week 12:** Late March, often related to Easter or school holidays.
  - **Week 30:** Late July, usually the mid-year vacation.

2.4 Cyclical Week Transformation

Used sine and cosine transformations (**sin_week**, **cos_week**) to encode the cyclical nature of weeks.

2.5 Geometric Features

Generated rotated coordinates (**rot_15_x**, **rot_15_y**, **rot_30_x**, **rot_30_y**) by rotating latitude and longitude by 15° and 30° to help the model capture non-linear spatial relationships.

2.6 Adjustment for COVID Impact in 2020

- Calculated average weekly emissions from non-COVID years (2019, 2021) and compared to 2020 to compute adjustment ratios.
- Applied these ratios to correct 2020 data.
- Applied power transformation (pow(emission, 1/1.5)) to smooth the spike in week 52.

2.7 Location Identification

- Combined **latitude** and **longitude** to create a **location** identifier, supporting grouping and sorting by space-time.

2.8 Rolling Mean Calculation

- Computed 7-week rolling means for the top 10 features most correlated with emissions.
- Replaced resulting NaN values with corresponding medians.

2.9 Feature Selection

- Selected rolling means of the top 10 correlated features.
- Included key features: **sin_week**, **cos_week**, **latitude**, **season**, **holidays**, **rotated coordinates**, **year**.
- Removed redundant or duplicate rolling mean features.

3. Modeling

XGBoost is a high-performance boosting algorithm using decision trees. It builds sequential models with regularization to prevent overfitting. Its accuracy and ability to handle imbalanced data make it widely used in competitions.

Random Forest is a robust machine learning algorithm that creates multiple decision trees using random data and feature subsets to enhance prediction accuracy. Its ability to prevent overfitting and handle various data makes it popular for prediction tasks.

LGBMRegressor is a gradient boosting model built on decision trees. It is designed for high performance and fast training, even on large datasets. The model is commonly used for regression tasks due to its accuracy and scalability.

CatBoostRegressor is a gradient boosting model designed to handle categorical features efficiently. It provides high accuracy with minimal data preprocessing. The model is robust, fast, and well-suited for a wide range of regression tasks.

4. Evaluation

- **Metric**

To evaluate the performance of the models, we use the **Root Mean Squared Error (RMSE)** as the main metric:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE measures the average magnitude of the prediction errors, providing an indication of how close the predicted values are to the actual CO_2 emissions.

- **Cross-Validation**

K-Fold Cross-Validation is a model evaluation method that splits the data into **k equal parts**. The model is trained and tested **k times**, with each part used once as the test set. The final result is the **average of the k evaluations**, providing more reliable and unbiased performance estimates.

C. EXPERIMENTAL RESULTS

Model	Pre-processing results	Post-processing results
CatBoost Regressor	27.37104 ± 1.41471	23.89653 ± 3.31377
XGB Regressor	22.01177 ± 1.98852	20.40372 ± 2.94385
LGBM Regressor	19.43554 ± 3.49667	18.68310 ± 4.06312
Random Forest Regressor	22.47780 ± 3.21664	16.11353 ± 5.16918

Observation:

Data processing significantly improves the performance of all models (as indicated by a reduction in error values).

- **Before data processing:**
  - LGBM achieved the best result (19.44 ± 3.50).
  - CatBoost performed the worst (27.37 ± 1.41).
- **After data processing:**
  - Random Forest Regressor achieved the best result (16.11 ± 5.17).
  - CatBoost Regression showed significant improvement but remained the lowest-performing model (23.90 ± 3.31).

Conclusion:

- Data processing is an essential and indispensable step in building machine learning models.
- Proper processing helps reduce prediction errors and improves the accuracy of all models.
- **Random Forest Regressor** demonstrates a strong ability to leverage processed data, while **LGBM Regressor** maintains stable performance both before and after data processing.

D. CONCLUSION AND FUTURE WORK

By employing the Random Forest Regressor, the project successfully predicted CO_2 emissions for 2022 with an average error (Mean RMSE) of **16.11 ± 5.17**, demonstrating stable predictive performance and an acceptable error margin. However, due to limitations in data availability beyond 2021, the current model relies solely on historical trends to forecast emissions for 2022, potentially reducing accuracy when extending predictions to subsequent years.

In the future, it is essential to incorporate new observational data, integrate supplementary data sources, and explore advanced algorithms (e.g., deep learning and ensemble methods) to enhance prediction coverage and accuracy. Additionally, implementing an automated monitoring and alert system will significantly improve the project's practical applicability.

TEAM TASK ASSIGNMENT

ID	Name	Task	Contribution
22521498	Nguyễn Thị Ngọc Trâm	Report writing, Experiment, Slide	22%
22521516	Dương Thành Trí	Report writing	15%
22521610	Phạm Nguyễn Anh Tuấn	Experiment, Demo	24%
22521626	Nguyễn Mạnh Tường	Experiment, Demo	22%
22521671	Lưu Khánh Vinh	Slide, Presentation	17%