**GHG Emission Prediction Project**

# **Day 2 -** Understanding Dataset and Step 1 Code

# **Date -** 18 June 2025 (Week 1)

# **Internship Name** - Edunet-Shell Skills4Future AICTE Internship

# **Intern Name:** Abhinay Singh

**✅ Tasks Completed**

**1. Explored and Understood Dataset Structure:**

* Viewed both 2016\_Summary\_Commodity and 2016\_Summary\_Industry Excel sheets.
* Learned the difference between commodity-based and industry-based GHG emission data.

Understood the purpose and scope of data columns, especially:

* Commodity / Industry Name
* Substance (e.g., CO₂, CH₄)
* Emission Factor with Margin → model output
* **Data Quality (DQ) metrics** (reliability, temporal, geographical, etc.)

**2. Understood the Role of Data Quality Metrics:**

* Identified them as numerical input features reflecting the **credibility of data**.
* Recognized their value in improving **model prediction accuracy**.

**3. Compared Industry vs Commodity Datasets:**

* **Industry =** Economic sectors (e.g., agriculture, healthcare)
* **Commodity =** Materials used across sectors (e.g., fertilizer, plastics)
* Understood the need for both perspectives in real-world policymaking.

**4. Completed Step 1 of Code: Imported Required Libraries**

**Successfully ran and understood the role of:**

* **pandas, numpy →** data handling
* **matplotlib, seaborn →** visualization
* **sklearn modules →** training, evaluation, tuning
* **joblib →** saving the model for future use

❓**Doubts Resolved Today**

| **Doubt** | **Clarification** |
| --- | --- |
| What is GHG and why is it harmful? | GHGs trap heat → global warming. Essential in moderation, dangerous in excess. |
| How can a student project help reduce GHGs? | By identifying high-emission industries/commodities, this project informs better environmental decisions. |
| Difference between industry and commodity datasets? | Industry = source of emission (e.g., farming), Commodity = what causes it (e.g., diesel) |
| What’s the output and input of the ML model? | Inputs = DQ scores, Substance, Name; Output = Emission Factor with Margin |
| Can beginners handle this? | Yes — with structured support, learning by doing, and consistent curiosity 💪 |

**✍️ Self Reflection**

* *"Today was incredibly insightful. I started with limited knowledge, but by the end, I understood the structure of real-world emission datasets and the foundational tools of machine learning. Though I am a beginner, I now appreciate how this project can genuinely contribute to a better world. I'm excited to move into preprocessing and modeling next. Looking forward to making it count — one step at a time."* 🌍💻.

**📚 References**

* **Dataset Source:** [data.gov - GHG Emissions Supply Chain Factors]
* National Grid Article on Greenhouse Effect (for theoretical understanding)
* Mentor-provided sessions materials and code
* ChatGPT for step-by-step breakdown and doubt clarification