

Emissions and Predictions: Evaluating the Role of CO₂ Data in ML Projects

Ivan Maslov

University of Silesia in Katowice, Poland
Faculty of Science and Technology
i.maslov@gsuite.us.edu.pl

Paulina Trybek

University of Silesia in Katowice, Poland
Institute of Physics, Faculty of Science and
Technology
paulina.trybek@us.edu.pl

Abstract. *Mitigating greenhouse gas (GHG) emissions, particularly carbon dioxide (CO₂), has become a central priority in environmental policy frameworks adopted by many nations. Governments and international bodies implement a wide range of strategies aimed at reducing emissions, with forecasting models playing an increasingly vital role in guiding adaptive and forward-looking decision making. In recent years, the emergence of advanced machine learning (ML) and artificial intelligence (AI) techniques has enabled more sophisticated approaches to emission prediction. However, despite this technological progress, comparatively little attention has been paid to the nature and origin of the underlying datasets used in these models. This paper presents a review of the most widely used sources of CO₂ emission data, the key concepts behind them, and the methodological frameworks of which emission values are composed, often through estimation formulas involving activity data, such as fuel consumption, and standardized emission factors. We approach this topic from the perspective of ML-based analysis, emphasizing the critical importance of dataset transparency, provenance, and credibility. Our goal is to initiate a constructive dialogue around the limitations of predicting values that are themselves derived estimates rather than direct measurements. In doing so, we also call for a more nuanced and critical approach to dataset exploration and construction, one that can serve as a more reliable foundation for future prediction efforts in both scientific and policy contexts.*

Keywords. Machine Learning, CO₂ Emissions, Dataset Transparency

References

- [1] Crippa, Monica, et al. "GHG emissions of all world countries." *Publications Office of the European Union, Luxembourg* 10 (2023): 953322.
- [2] Directorate-General for "Joint Research Centre". (n.d.). *EDGAR — emissions database for global atmospheric research*. Retrieved May 20, 2025, from <https://edgar.jrc.ec.europa.eu/methodology>
- [3] Eggleston, H. S., et al. "2006 IPCC guidelines for national greenhouse gas inventories." (2006).
- [4] Ritchie, Hannah, Pablo Rosado, and Max Roser. "CO₂ and greenhouse gas emissions." *Our world in data* (2023).