



INTRODUCTION TO MACHINE LEARNING APPLICATIONS FOR
SUSTAINABILITY RESEARCH

Lecture 1: Introduction to ML & Sustainability

Foundations of supervised & unsupervised learning, framed through real sustainability and operations challenges.

Prof. Lucas Zanon

Prof. Tiago Sigahi

University of São Paulo (USP) — Department of Production Engineering

July 13–22, 2026 · Remote Lectures

Today's Roadmap



1. What Is Machine Learning?

Definitions, intuition, and a classic statistical-learning example.



2. A Map of ML Methods

How supervised and unsupervised techniques fit together.



3. Core Concepts You'll Need

Regression vs. classification, parametric vs. non-parametric, overfitting.



4. Sustainability Case: CO₂ Emissions

Framing a real ML problem on country-level emissions data.



5. Getting Started with Python

Why Python, basic syntax, and your first lines of pandas & scikit-learn.



PART 1

What Is Machine Learning?

Where the term comes from, what problem it solves, and why it overlaps so heavily with classical statistics.

Machine Learning: What Is It? What Is It For?



“A set of methods that can reveal patterns of variation in data and use that pattern to predict future outcomes.”

— adapted from Murphy, K. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.



“A set of methods and algorithms capable of extracting knowledge from data, learning from experience accumulated over time.”

— adapted from Bertolini et al. (2021). Machine learning for industrial applications: A comprehensive literature review.

Machine Learning: Examples

Example 1 · Manufacturing

Analyzing the relationship between noise, cutting force, and geometric error in a machining process can reveal a pattern that relates process variables (input) to the outcome.

Example 2 · Customer Risk

Analyzing the relationship between customer income, invoice value, and payment delay can reveal a pattern that relates customer profile (input) to the outcome.

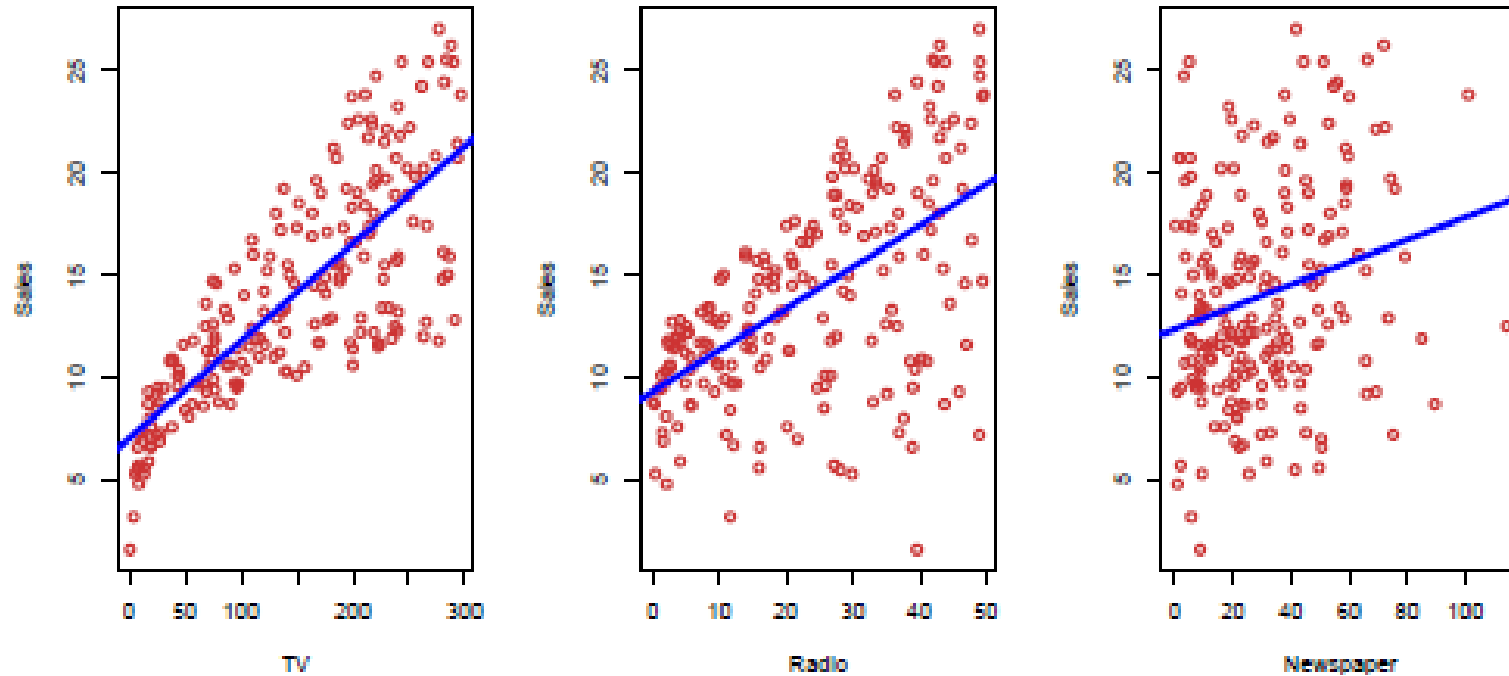
In both cases, the input variables can be controlled to avoid undesirable outcomes.

Machine Learning or Statistical Learning

- Machine learning methods are also known as statistical learning methods.
- Statistical learning refers to a broad set of multivariate statistical techniques applied to model and understand the relationships within data.

Statistical Learning: Example

- Sales data as a result of advertising spend on TV, radio, and newspaper.



James, G. et al., An introduction to statistical learning, Springer, 2021.

Statistical Learning: Example

In this example, the input variables — also called *predictors* or *independent variables* — are:

X_1 : budget spent on TV advertising;

X_2 : budget spent on radio advertising;

X_3 : budget spent on newspaper advertising;

The output variable — also called the *dependent variable* or *response variable* — is:

Y : sales.

Statistical Learning: Example

We assume there is a relationship between the predictor variables and the response variable. Mathematically:

$$Y = f(X) + \varepsilon$$

Where $X = (X_1, X_2, \dots, X_p)$

Why Estimate $f(X)$?

Prediction: in many situations, we have the input variables but not the response, and we want to predict it.

$$\hat{Y} = \hat{f}(X)$$

$\hat{f}$ represents the estimate of f , and $\hat{Y}$ the prediction of Y .

Example: suppose $X = (X_1, X_2, \dots, X_p)$ are a patient's blood characteristics, and Y is the risk of an allergic reaction to a given medication, as a function of X . We want to predict Y based on X .

Why Estimate $f(X)$?

Inference: in other situations, we want to understand the relationships between Y and $X = (X_1, X_2, \dots, X_p)$:

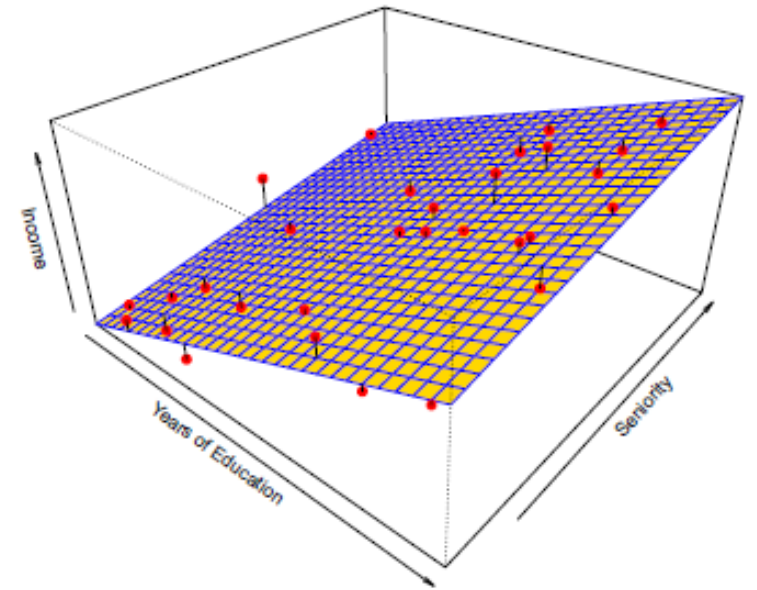
- Which of the predictor variables $X = (X_1, X_2, \dots, X_p)$ significantly affect the response Y ;
- What is the relationship between each predictor variable and the response? Positive, negative, linear, non-linear?

How Do We Estimate $f(X)$?

We start from a dataset, called the *training data*.

Example: dataset with $n = 30$:

- x_1 : years of education;
- x_2 : age;
- $x_{ip} = (x_{i1}, x_{i2})^T, i = 1, 2, \dots, n.$
- Y_i : income;
- $\{(x_{11}, x_{12}, Y_1), (x_{21}, x_{22}, Y_2), \dots, (x_{301}, x_{302}, Y_{30})\}$



How Do We Estimate $f(X)$?

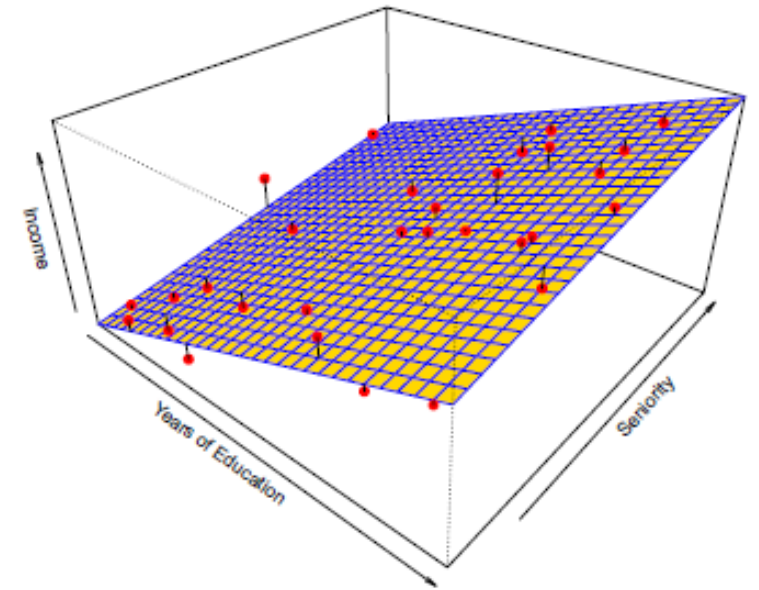
Based on the dataset

$$\{(x_{11}, x_{12}, Y_1), (x_{21}, x_{22}, Y_2), \dots, (x_{301}, x_{302}, Y_{30})\}$$

we apply a statistical learning method to estimate the function $f(X)$. That is:

$$Y \approx \hat{f}(X), \text{ for any observation } (X, Y)$$

Based on $\hat{f}$, we can predict the value of Y from new values of the input variables.





PART 2

A Map of Machine Learning Methods

From the supervised/unsupervised split down to the individual algorithms you'll use this week.

Supervised vs. Unsupervised: The Big Picture



Supervised Learning

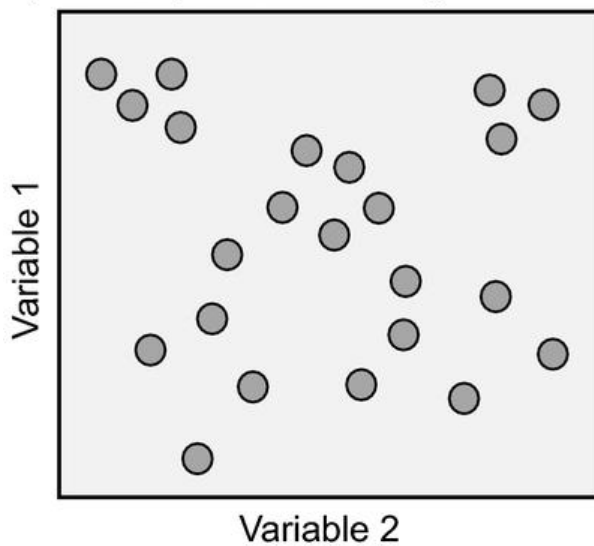
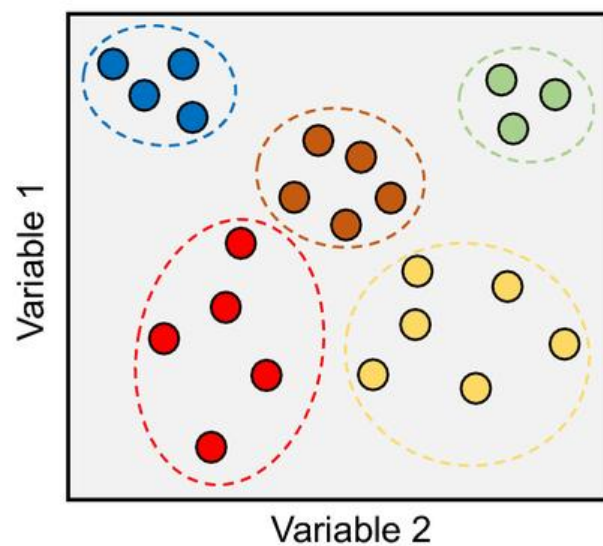
- We have labeled data: every observation has a known input X and a known output Y .
- Goal: learn f so that, given new X , we can predict Y .
- Quantitative $Y \rightarrow$ regression problem.
- Qualitative $Y \rightarrow$ classification problem.



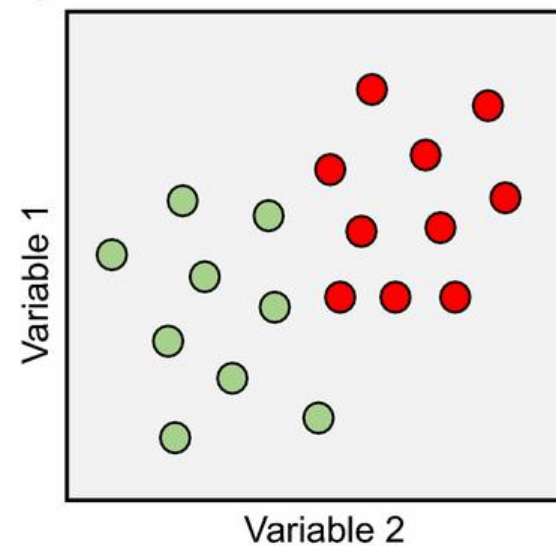
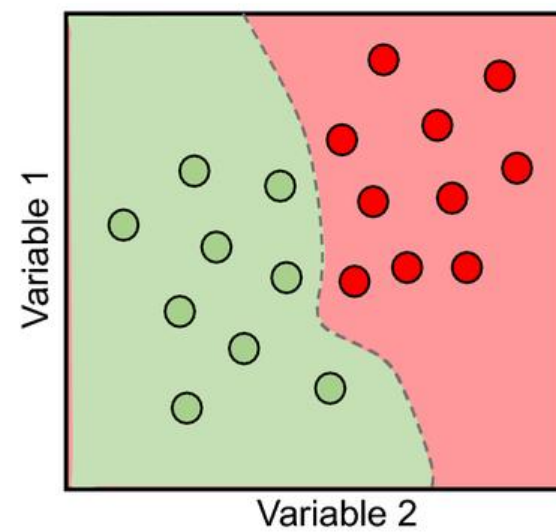
Unsupervised Learning

- We only observe inputs X — there is no labeled response Y .
- Goal: discover structure, groupings, or a simpler representation of the data.
- Examples: clustering (grouping similar observations).
- Examples: dimensionality reduction (fewer variables, same information).

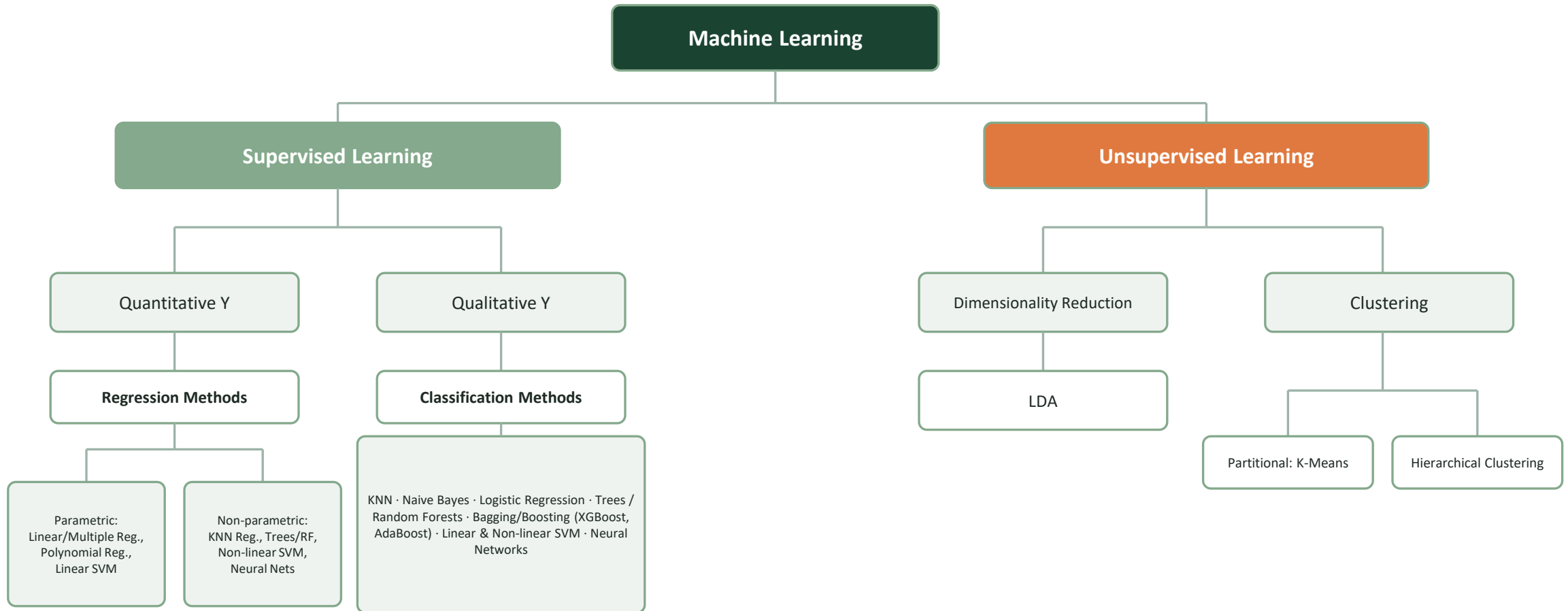
a) Unsupervised learning



b) Supervised learning



The Full Map: ML Methods at a Glance



We'll zoom into each branch over the next few slides.

Zooming In: Regression Methods

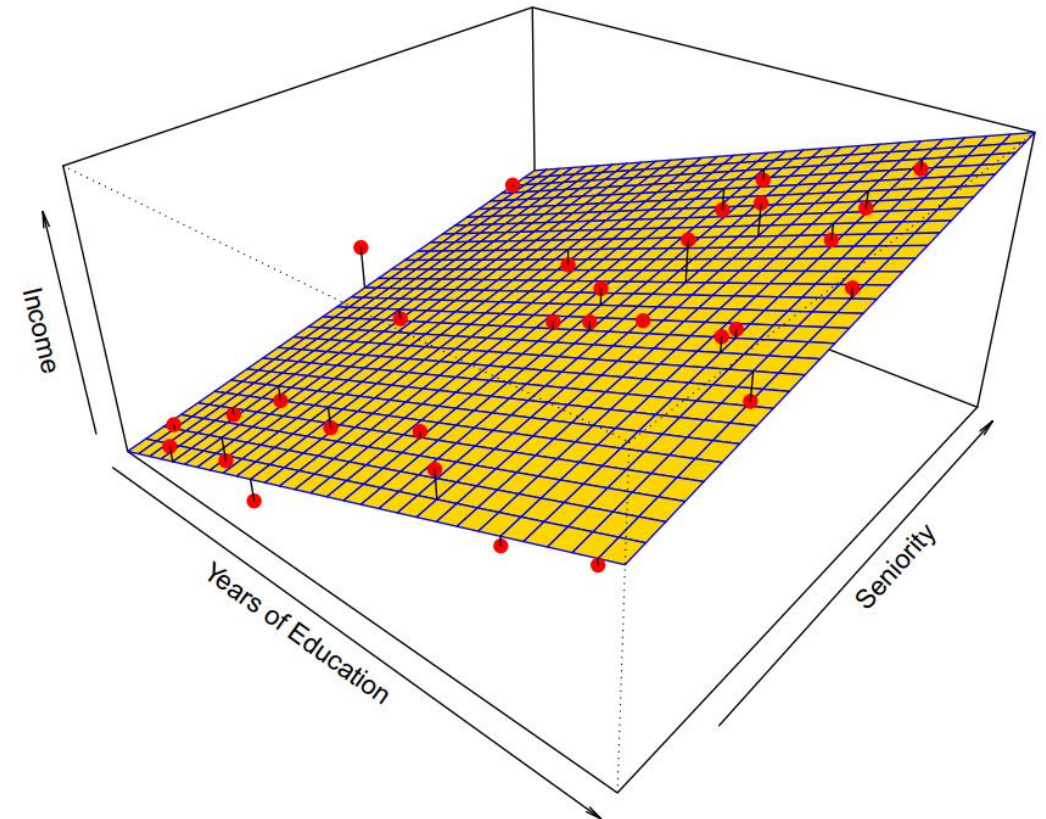
Used when the response Y is quantitative (continuous).

Parametric

- Linear / Multiple Regression
- Polynomial Regression
- Linear SVM

Parametric statistical learning assumes that data follows a specific distribution or model and the structure of the model is defined by a fixed set of parameters.

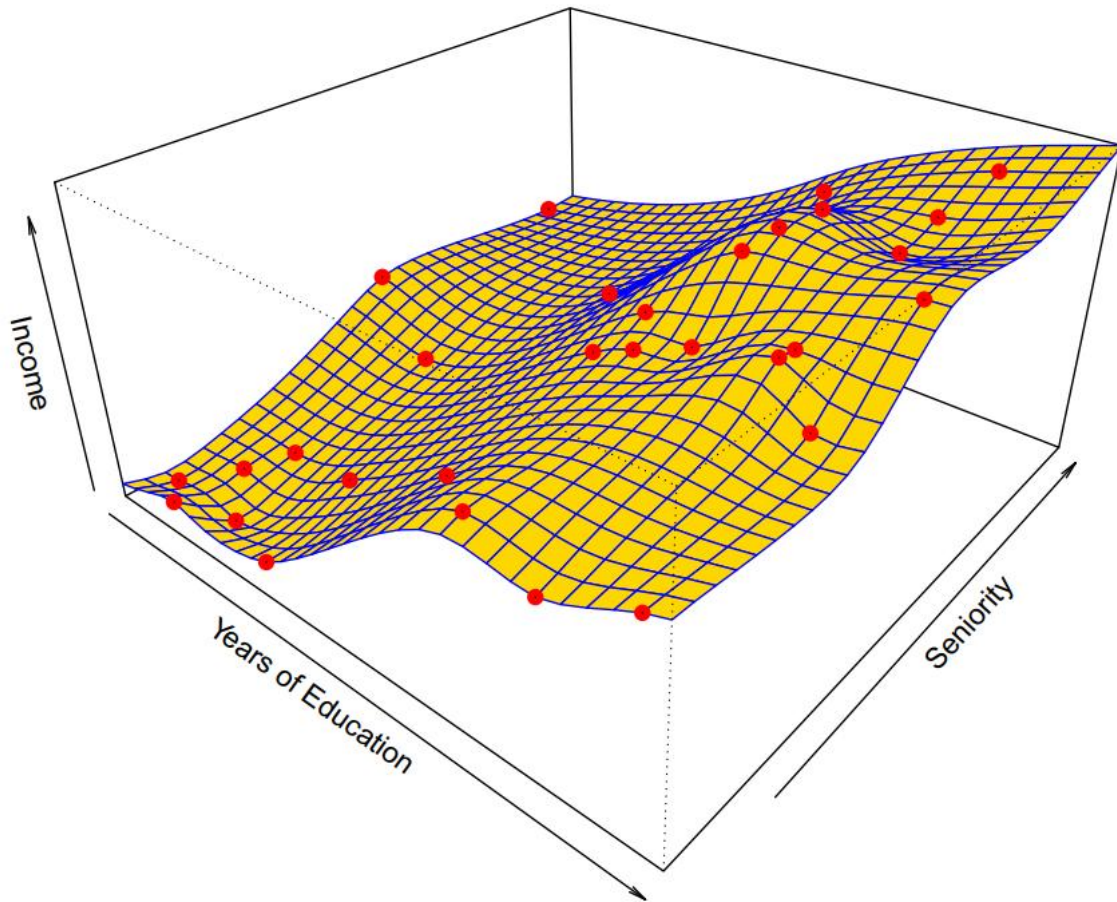
Model parameters are estimated from the data.



Zooming In: Regression Methods

Used when the response Y is quantitative (continuous).

Non-parametric statistical learning does not take a specific functional form for the model and the number of parameters can grow with the size of the data.



Non-parametric

- KNN Regression
- Classification Trees / Random Forests
- Non-linear SVM (Polynomial, RBF kernels)
- Neural Networks

Zooming In: Classification Methods

Used when the response Y is qualitative (categorical).

KNN Classification

Naive Bayes

Logistic Regression

Classification Trees / Random Forests

Bagging / Boosting (XGBoost, AdaBoost)

Linear SVM

Non-linear SVM (Polynomial, RBF)

Neural Networks

Zooming In: Unsupervised Methods



Dimensionality Reduction

Compresses many correlated variables into a smaller set that preserves most of the information.

LDA (Linear Discriminant Analysis)



Clustering

Groups observations so that items in the same group are more similar to each other than to items in other groups.

Partitional:
K-Means

Hierarchical Clustering



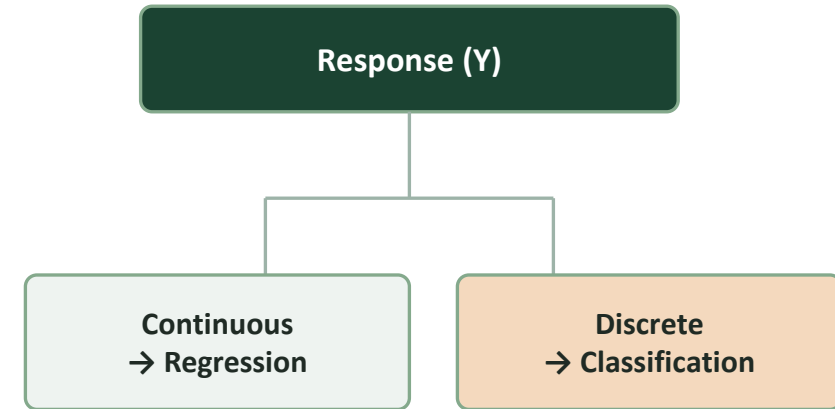
PART 3

Core Concepts You'll Need

The vocabulary and intuitions that make every later lecture easier to follow.

Regression vs. Classification

- Variables can be characterized as quantitative or qualitative.
- Quantitative: age, weight, height, income, CO₂ emissions, etc.
- Qualitative (categorical): gender, marital status, product brand, etc.
- When the response (Y) is quantitative → regression problem: the model estimates a numeric value.
- When the response (Y) is qualitative → classification problem: the model estimates which class the response belongs to.



Parametric vs. Non-Parametric Methods

Parametric

Assumes the data follow a specific distribution or model form. The model's structure is defined by a fixed set of parameters, estimated from the data.

Examples: Linear Regression, Logistic Regression

Non-Parametric

Does not assume a specific functional form. The number of effective parameters can grow as more data becomes available.

Examples: KNN, Decision Trees, SVM with kernels, Neural Networks

The Flexibility – Interpretability Trade-off



Methods toward the top-left are easier to interpret; methods toward the bottom-right are more flexible but harder to explain. There is rarely a free lunch.

What Is Overfitting?

Overfitting occurs when a model fits the training data excessively — capturing not just the underlying pattern, but also the noise, errors, and peculiarities specific to that dataset.

As a result, the model performs very well on training data, but fails to generalize to new, unseen data.

We'll guard against this with train/test splits, cross-validation, and hyperparameter tuning, covered in upcoming lectures.



What Is Overfitting?

Overfitting occurs when a model fits the training data excessively — capturing not just the underlying pattern, but also the noise, errors, and peculiarities specific to that dataset.

As a result, the model performs very well on training data, but fails to generalize to new, unseen data.

We'll guard against this with train/test splits, cross-validation, and hyperparameter tuning — covered in upcoming lectures.



Supervised Techniques You'll Learn



Regression

- Simple linear regression
- Multiple linear regression
- Non-linear regression
- Neural networks (deep learning)



Classification

- Naive Bayes
- K-nearest neighbors (KNN)
- Logistic regression
- Linear discriminant analysis (LDA)
- Decision trees
- Support vector machines (SVM)

Unsupervised Techniques You'll Learn



K-Means

Partitions data into k groups by minimizing within-cluster variance.



Principal Component Analysis (PCA)

Reduces many correlated variables into a smaller set of uncorrelated components.



Hierarchical Clustering

Builds a tree of nested clusters, useful when the number of groups isn't known in advance.



PART 4

Today's Sustainability Case: CO₂ Emissions

Putting the taxonomy to work on a real, messy, country-level dataset.

The Dataset: CO₂ Emissions Across Countries

A country-year panel combining emissions with the economic and energy variables that plausibly explain them.

Variable	Description	Type
Country / Year	Geographic and time identifiers	Categorical / Discrete
CO ₂ emissions (total & per capita)	Tonnes of CO ₂ emitted, overall and per person	Quantitative
GDP (total & per capita)	Economic output, a common driver of emissions	Quantitative
Population	Number of inhabitants	Quantitative
Energy use per capita	Primary energy consumption per person	Quantitative
Share of renewable energy	Percentage of energy from renewable sources	Quantitative

Framing the ML Problem: Discussion

What is Y?

Could be total CO₂ emissions (a number) — or a category, like “high” vs. “low” emitter.

Regression or classification?

Predicting next year's emissions → regression. Flagging countries above a threshold → classification.

Supervised or unsupervised?

Predicting emissions from GDP/energy use → supervised. Grouping countries by emission profile, with no target → unsupervised (clustering).

What could X be?

GDP per capita, population, energy mix, industrial output, renewable energy share — and how they interact.



PART 5

Getting Started with Python

The language we'll use to turn every concept in this lecture into working code.

Why Python for Data Science & ML?

- Free, open-source, and the most widely used language in applied machine learning research.
- Readable syntax — you can focus on the modeling idea, not the language.
- A mature ecosystem of libraries covering every step: data wrangling, visualization, modeling, and text analysis.



Data loading, cleaning & manipulation



Predictive models & evaluation



Foundational plotting



Statistical visualization



Interactive charts



NLP & sentiment analysis

Python Basics: Variables, Types & Operators

- Variables are created on assignment — no need to declare a type.
- Common types: str (text), int (whole numbers), float (decimals), bool (True/False).
- Arithmetic (+ - * / **) and comparison (== != > <) operators work as expected.

```
country = "Brazil"           # str
year = 2026                   # int
co2_emissions = 462.4         # float (Mt CO2)
is_high_emitter = co2_emissions > 400    # bool -> True

print(f"{country} emitted {co2_emissions} Mt CO2 in {year}.")
```

Python Basics: Lists, Loops & Functions

- Lists hold ordered collections of items, indexed from 0.
- for loops iterate over a list (or any iterable) one item at a time.
- def defines a reusable function — the backbone of every analysis script.

```
countries = ["Brazil", "Germany", "India"]

for c in countries:
    print(f"Processing {c}...")

def classify_emitter(co2):
    return "high" if co2 > 400 else "low"
```

First Look at Data: pandas in Action

- pandas loads tabular data into a DataFrame — think spreadsheet, but programmable.
- `head()` previews rows; `describe()` summarizes each numeric column.
- This is exactly how we'll start exploring the CO₂ emissions dataset next class.

```
import pandas as pd

df = pd.read_csv("co2_emissions.csv")
df.head()          # preview the first 5 rows
df.describe()      # mean, std, min, max for numeric columns
df["co2_per_capita"].mean()
```

Your First Model: scikit-learn in 5 Lines

- Every scikit-learn model follows the same pattern: split data, fit, predict.
- `train_test_split` holds out data to check generalization — our defense against overfitting.
- Swapping `LinearRegression()` for another model is almost always this easy.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42)

model = LinearRegression().fit(X_train, y_train)
predictions = model.predict(X_test)
```



WRAP-UP

Key Takeaways

- Machine learning = statistical learning: revealing patterns in data to predict or explain outcomes.
- The first fork in the road is supervised (labeled Y) vs. unsupervised (no Y) — everything else follows from that choice.
- Regression vs. classification depends only on whether Y is quantitative or qualitative.
- Flexibility buys predictive power but costs interpretability — and overfitting is the risk you pay for unchecked flexibility.
- Python (pandas + scikit-learn) is how every one of these ideas becomes a working model.



INTRODUCTION TO MACHINE LEARNING APPLICATIONS FOR
SUSTAINABILITY RESEARCH

Lecture 1: Introduction to the sustainability context

Carbon emissions: Global context, socioeconomic challenges, and paths toward decarbonization.

Prof. Tiago Sigahi

Prof. Lucas Zanon

University of São Paulo (USP) — Department of Production Engineering

July 13–22, 2026 · Remote Lectures

Today's Sustainability Roadmap: Carbon emissions

1

What is CO₂?

Carbon dioxide, greenhouse effect and human-induced imbalance.

2

CO₂ emissions as a global sustainability problem

Economic, social and environmental aspects.

3

Where emissions come from

Fossil fuels, industry and land-use change.

4

Who emits, and how we count

Total, per-capita, historical, territorial and consumption perspectives.

5

Development, inequality and transition

Energy access, decoupling, climate justice and policy dilemmas.

Is CO₂ the same as greenhouse gas emissions?

Term	Meaning	Examples
CO ₂ emissions	Emissions of carbon dioxide	Fossil fuels, cement, deforestation
Greenhouse gas emissions	CO ₂ plus other gases that contribute to warming	CO ₂ , methane, nitrous oxide
CO ₂ -equivalent — CO ₂ e	A common unit expressing the warming effect of different gases	Methane and other gases converted into comparable units

Why the distinction matters

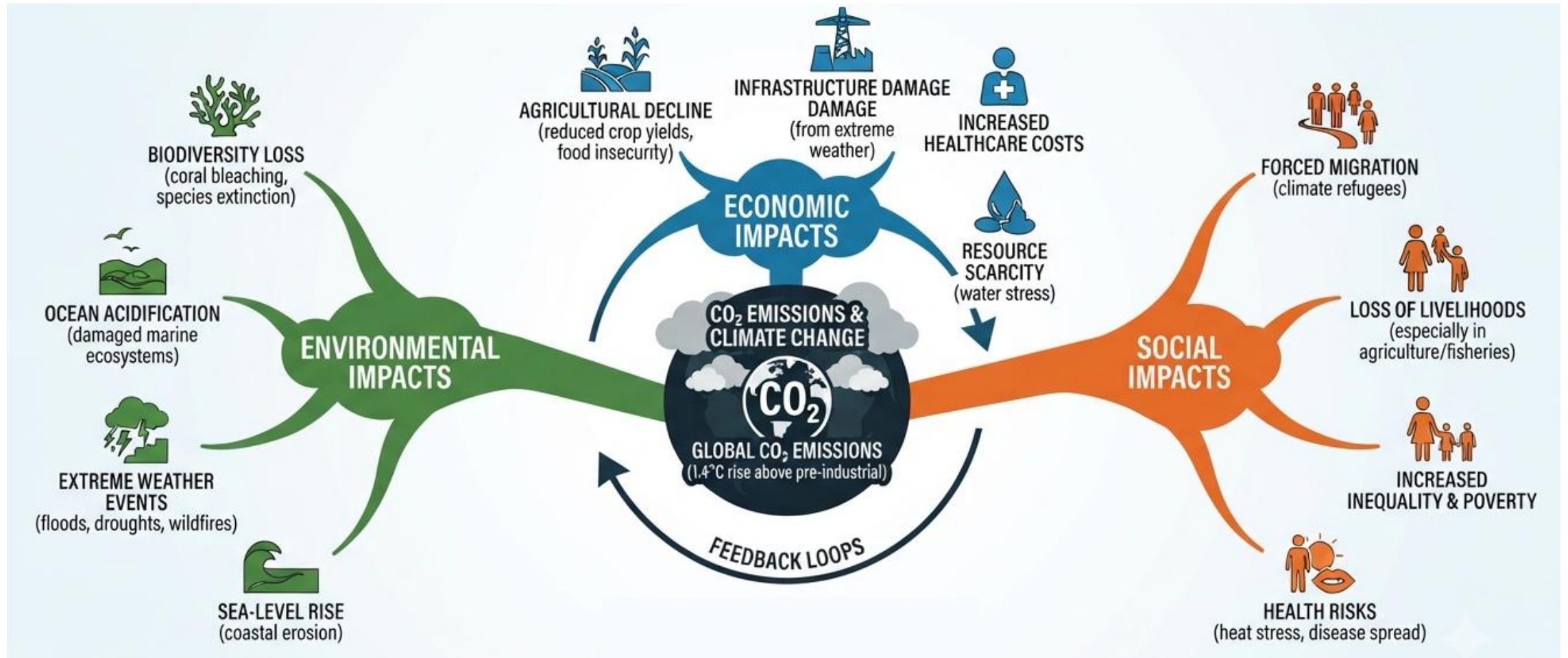
A sector may have limited direct CO₂ emissions while producing substantial methane or nitrous oxide. Climate assessments must state which gases are included.

Example

Agriculture can have major climate impacts through methane and nitrous oxide even when direct fossil CO₂ emissions are limited.



CO₂ emissions: a global sustainability challenge



CO₂ emissions: a global sustainability challenge

38.1 Gt

projected fossil CO₂ emissions in 2025

≈ 4.8 tonnes

per person worldwide per year

1 day ≈ 13 kg CO₂
per person per day

4.1 Gt

projected CO₂ from land-use change in 2025

deforestation, forest degradation and land conversion

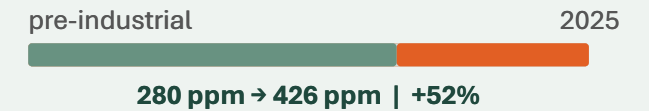
+11% added on top of CO₂ emissions



425.7 ppm

projected atmospheric CO₂ concentration in 2025

≈ 4 CO₂ molecules in every 10,000 air molecules



+1.4 °C, instability and extreme climate, ecosystems imbalance

The difficult question we must ask:

Why do global CO₂ emissions continue to increase despite decades of climate agreements, technological progress and public debate?

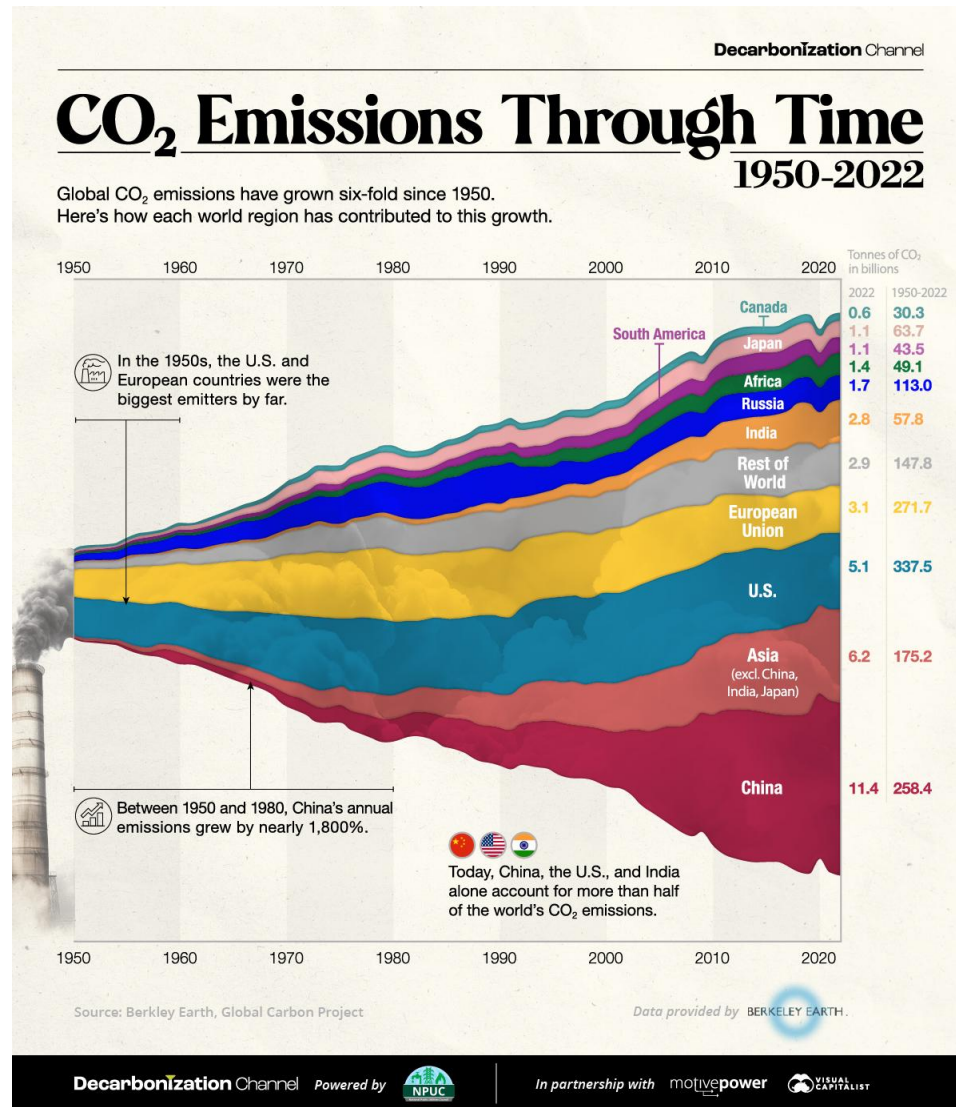
What the numbers show

emissions are still rising
land use remains a major source
concentrations keep accumulating

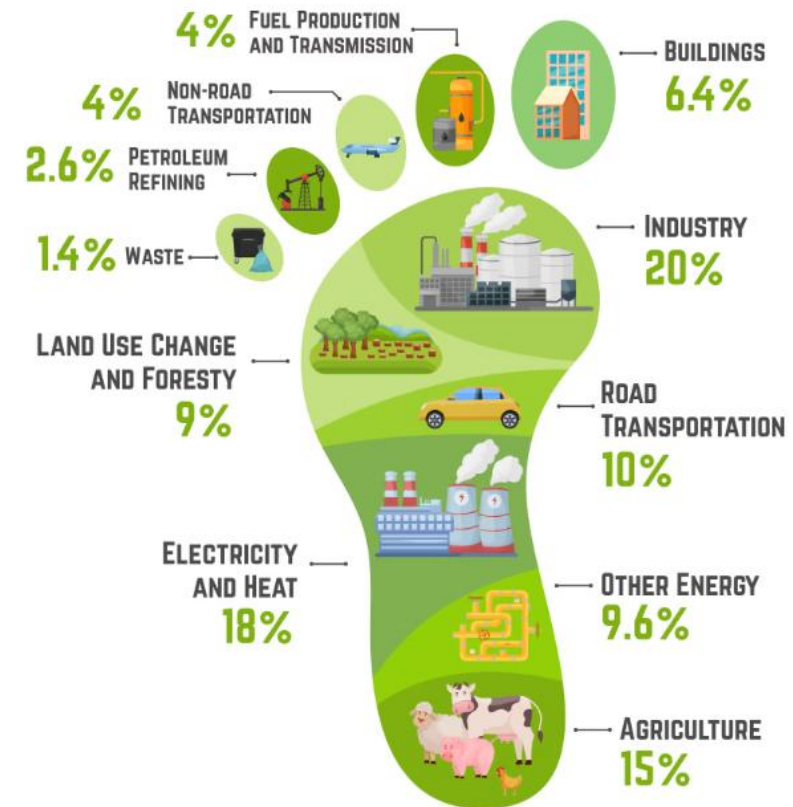
What they do not show

who is responsible
who has capacity to act
who is most vulnerable

CO₂ emissions: a global sustainability challenge



CARBON FOOTPRINT



A problem of accumulation



Think of CO₂ as a filling bathtub

The climate problem is driven by the stock that accumulates, not by a single year in isolation.

- 1 Annual emissions**
the inflow entering every year
- 2 Cumulative CO₂**
the stock already in the system
- 3 Net-zero CO₂**
the point where the stock stops growing

Why accumulation makes forecasting necessary

Past emissions define the stock already accumulated.
Future emissions determine how fast the “bathtub” keeps filling.
A snapshot is not enough: sustainability decisions depend on trajectories.

From measuring the past to anticipating the future

Forecasts help ask: if current patterns continue, where are emissions heading?
They also show when intervention is still possible — before the stock becomes larger and harder to reverse.
This motivates the next step: estimating future CO₂ trajectories from observed country-level patterns.

Who emits? The answer depends on how we count

Annual total emissions

Current scale of a country's contribution.

Emissions per person

Average emission level associated with each resident.

Historical cumulative emissions

Contribution made over decades of industrialization.

Territorial emissions

Emissions attributed to the country where production occurs.

Consumption emissions

Emissions attributed to final consumption of goods and services.

Carbon intensity

Emissions relative to economic output, energy or activity.

Implication There is no single ranking of the world's "largest emitters." Different indicators answer different questions and imply different interpretations of responsibility.

Progress is visible, but insufficient

3 Gt

annual CO₂ emissions avoided in 2025 by low-emission technologies deployed since 2019

IEA, 2026

2.8°C

projected warming this century under current policies

UNEP, 2025

55%

global reduction needed by 2035 for a 1.5°C pathway, relative to 2019

UNEP, 2025

What is improving?

solar and wind deployment
electric vehicles
some national emission reductions

Why is it insufficient?

global energy demand continues to rise
fossil fuels remain central
Linear economy model dominates

Implication

Technological progress matters, but its scale and speed must exceed the growth of global energy and material demand.

Opening debate

Who should reduce CO₂ emissions, how rapidly, and who should bear the costs?

Historical responsibility

Countries that **emitted more** during industrialization should **contribute more**.

Current contribution

Countries that currently produce the **largest emissions** should act most rapidly.

Capacity to act

Wealthier countries with **greater technological and financial resources** should carry a **larger burden**.

Development needs

Lower-income countries should retain greater **space to expand essential services and infrastructure**.

If you want to learn more



Twenty-five years of adaptation finance through a climate justice lens

Mizan Khan¹ · Stacy-ann Robinson²  · Romain Weikmans³  · J. Timmons Roberts⁵



Theorising environmental justice: the expanding sphere of a discourse

David Schlosberg 

Example

scientific reports

[nature](#) > [scientific reports](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 03 June 2025

A machine learning approach to carbon emissions prediction of the top eleven emitters by 2030 and their prospects for meeting Paris agreement targets

[Arju Manara Begum](#) & [Mahadee Al Mobin](#) 

[Scientific Reports](#) **15**, Article number: 19469 (2025) | [Cite this article](#)

<https://doi.org/10.1038/s41598-025-04236-5>

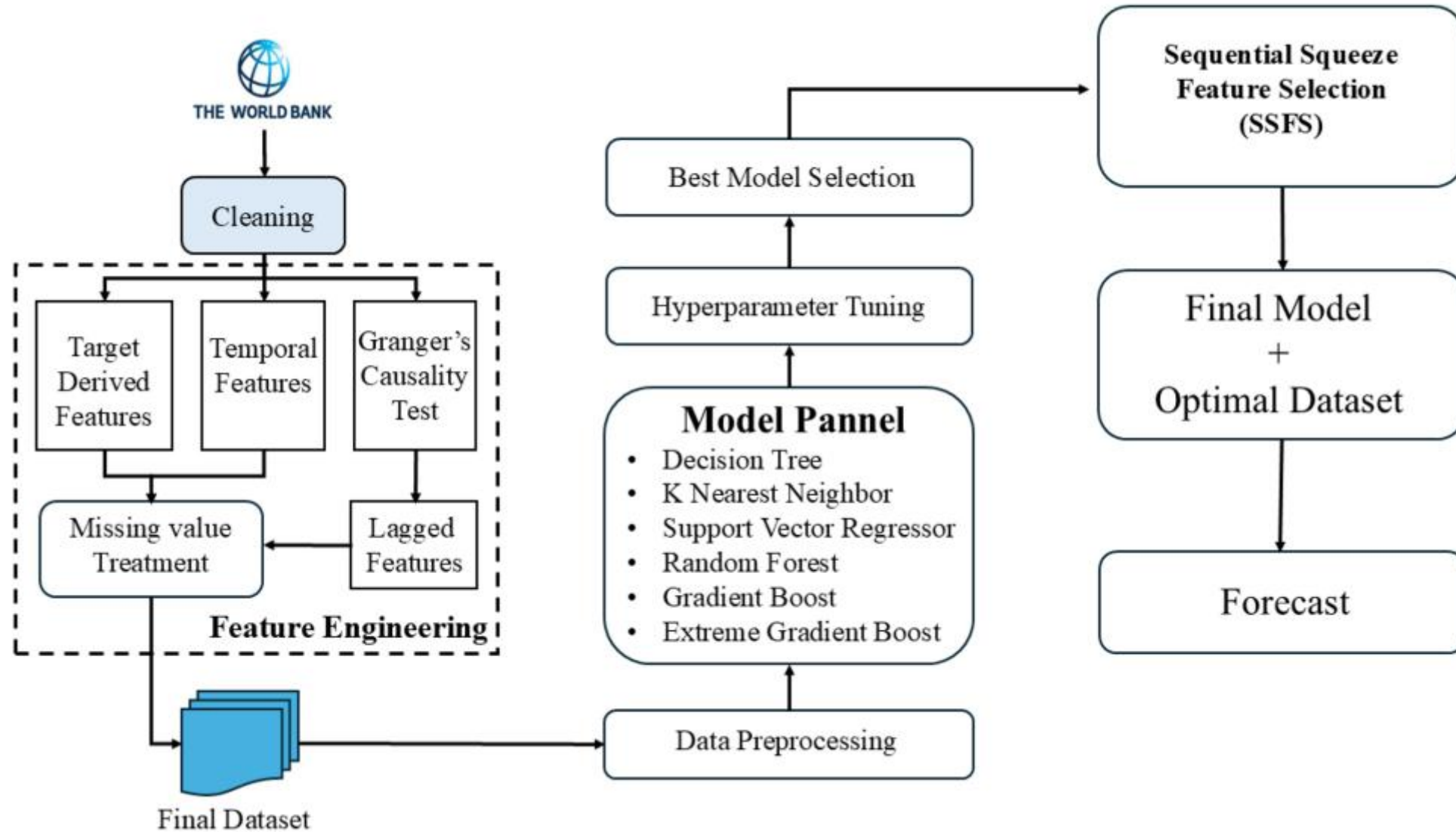
- Data from 1990 to 2023
- Forecast CO2 emissions through 2030 according to targets (NDCs)



- 6 machine learning models
- 11 variables (economic, industrial, and energy consumption)
- Average prediction accuracy of 96.21%.

Table 1. Summary of key contributions in CO_2 emissions forecasting literature.

Author(s)	Region/Country	Methodology	Variables Used
C. Dong et al. (2018) ⁴	Top 10 carbon emitters	Trend extrapolation + BP Neural Network	Economic, energy
Huang et al. (2019) ²⁰	China	PCA + LSTM	16 economic/industrial variables
A. Mardani et al. (2020) ²⁶	G20 countries	ANFIS, ANN with SVD, SOM	Energy consumption, GDP
Li (2020) ¹³	China	Hybrid LSTM + SVM	GDP, energy, manufacturing
Ofosu-Adarkwa et al. (2020) ²¹	China (Cement)	Grey prediction model	Cement sector emissions
X. Li et al. (2020) ¹³	China	ARIMA, SARIMAX, ANN, RF, LSTM	Univariate (emissions only)
Jena et al. (2021) ¹²	Global (17 nations)	MLANN	GDP, population, emissions
Ahmed et al. (2022) ³	China & India	LSTM	Energy, emissions (1990-2014)
Chang et al. (2023) ²³	China	PPR + ANN	Electricity, Energy consumption
M. Ahmed et al. (2023) ²⁵	China, India, Russia and the USA	SVM, ANN and LSTM	Energy consumption factors
S. Kumari et al. (2023) ²²	India	ARIMA, SARIMAX, Holt-Winters, RF, LR, LSTM	Univariate
Wang et al. (2023) ²⁸	China	Two-stage ML (SVR-ANN)	Quarterly economic data
Chukwunonso et al. (2024) ²⁴	USA	L-RNN, ANN	Energy use (1973-2022)

Fig. 1. Workflow of the study.

Literature validation of feature selection (Table 12)

Variable	Conceptual Role	Supported By
Fossil Fuel Energy Consumption	Carbon intensity of energy	4,26
GDP	Economic scale of emissions	3,13
Electric Power Consumption	Energy intensity	3,27
Population	Demographic pressure	3,12
Renewable Energy Output	Clean energy transition	4,6
Industry (% of GDP)	Industrial structure	13
Manufacturing	Production structure	14
Trade	Economic globalization and emissions outsourcing	17
Renewable Energy Consumption	Clean energy adoption and mitigation	4,6
Domestic Credit to Private Sector	Financial development and investment-driven emissions	4
Fertilizer Consumption	Agricultural modernization	16

Sl.	Country	Carbon Emission in 2030	Base Year Carbon Emission	Change of Carbon Emission from base year (%)	NDCs
1	Canada	0.312	0.415	−24.82%	At least 40-45% below 2005 levels by 2030.
2	China	0.616	1.318	−53.11%	“Over 65%” in 2030 from the 2005 level.
3	Germany	0.156	0.408	−61.76%	65% reduction below 1990 levels.
4	India	0.833	1.039	−19.92%	45% below 2005.
5	Iran	1.419	1.335	+6.07%	12% below BAU by 2030 from 2010 level.
6	Indonesia	0.537	0.632	−15.03%	32% <i>against the 2030 BAU from 2010 level.</i>
7	Japan	0.241	0.290	−16.90%	46% <i>reduction in 2030 from 2013.</i>
8	Russia	0.989	1.863	−46.70%	30% below 1990 levels by 2030.
9	Saudi Arabia	0.856	0.734	+16.62%	Reduce emissions by 278 $MtCO_2e$ annually by 2030, with 2019 as the base year.
10	South Korea	0.325	0.393	−17.30%	40% below 2018 by 2030.
11	USA	0.206	0.370	−44.33%	50 – 52% below 2005 levels by 2030.

Findings

- **Russia** is **on track** to exceed its reduction targets, while **Germany** and **USA** will **fall slightly short**
- **China, India, Japan, Canada, South Korea, and Indonesia** are projected to miss their commitments by **significant margins**
- At the same time, **Iran** and **Saudi Arabia** are expected to **increase emissions** rather than reduce them.