**Dataset overview**

Each row = one **(observation × day)** record.

* **Observation**: an iNaturalist sighting of *Pyracantha angustifolia*.
* **Day**: a NASA POWER daily weather record for the observation’s coordinates on a specific date (spanning N years back to the sighting date).
* Engineered **feature aggregates** are repeated per row (keyed by inat\_id) so you can train from the flat file or pivot as needed.

**Keys & joins**

* Primary keys in practice:
  + inat\_id (iNat observation id)
  + date (YYYY-MM-DD)
* Location:
  + latitude\_x, longitude\_x — from iNaturalist (original sighting)
  + latitude\_y, longitude\_y, elevation — from NASA POWER response (gridded location rounded by the API)

**Observation fields (from iNaturalist)**

* inat\_id — unique observation id (same as id in iNat).
* uuid — iNat UUID for the observation.
* user\_id — iNat user identifier.
* time\_observed\_at — ISO timestamp when the sighting happened (local TZ per iNat).
* created\_at — ISO timestamp when the record was created on iNat.
* scientific\_name — taxon scientific name (e.g., *Pyracantha angustifolia*).
* common\_name — taxon common name.
* quality\_grade — research / needs\_id (filter cue for quality).
* place\_guess — human-readable location.
* latitude\_x, longitude\_x — reported coordinates.
* positional\_accuracy — accuracy in meters (lower = better).
* image\_url — first image URL (if any).

**NASA POWER daily weather (per date)**

Units follow NASA POWER:

* ALLSKY\_SFC\_SW\_DWN — All-sky surface shortwave irradiance (kW·h/m²/day).
* CLRSKY\_SFC\_SW\_DWN — Clear-sky shortwave irradiance (kW·h/m²/day).
* PRECTOTCORR — Precipitation corrected (mm/day).
* RH2M — Relative humidity at 2 m (%).
* T2M — Mean air temperature at 2 m (°C).
* T2M\_MAX — Max air temperature (°C).
* T2M\_MIN — Min air temperature (°C).
* TQV — Total column precipitable water (kg/m²).
* TS — Earth skin (surface) temperature (°C).
* WS2M — Wind speed at 2 m (m/s).
* date — daily record date (UTC-aligned daily aggregation from POWER).
* latitude\_y, longitude\_y — POWER grid point used.
* elevation — POWER elevation (m).

**Engineered feature aggregates (per inat\_id)**

Computed over a **lookback window** (default 3 years) ending on the observation date:

* **Temperature**
  + feat\_t2m\_mean — mean daily T2M.
  + feat\_t2mmax\_mean — mean of T2M\_MAX.
  + feat\_t2mmin\_mean — mean of T2M\_MIN.
  + feat\_gdd\_base10 — Growing Degree Days (base 10°C): Σ max( (T2M\_MAX+T2M\_MIN)/2 − 10, 0 ).
  + feat\_heat\_days\_30c — count of days with T2M\_MAX ≥ 30°C.
  + feat\_frost\_days\_0c — count of days with T2M\_MIN ≤ 0°C.
* **Moisture & rainfall**
  + feat\_prectot\_sum — total precipitation (mm).
  + feat\_rh2m\_mean — mean relative humidity (%).
  + feat\_tqv\_mean — mean precipitable water (kg/m²).
* **Radiation & cloudiness**
  + feat\_allsky\_mean — mean ALLSKY\_SFC\_SW\_DWN.
  + feat\_clrsky\_mean — mean CLRSKY\_SFC\_SW\_DWN.
  + feat\_cloud\_index\_mean — mean cloudiness proxy = 1 − (ALLSKY / CLRSKY), clipped to [0,1].
* **Wind**
  + feat\_ws2m\_mean — mean wind speed at 2 m (m/s).
* **Seasonality context**
  + feat\_month\_obs — month of observation (1–12).
  + feat\_season\_obs — derived season label at the location (e.g., DJF, MAM, JJA, SON or custom SA seasons).

Note: All feat\_\* columns repeat per daily row for the same inat\_id to keep the dataset flat. You can deduplicate to one row per observation if you prefer modeling purely on the feature aggregates.

**Example interpretation**

A single row means:

“For observation inat\_id=210128381 at (latitude\_x=-34.006, longitude\_x=18.438), on **2024-04-27**, the NASA grid at (latitude\_y=-34.006, longitude\_y=18.438, elevation=209.37 m) recorded **T2M=19.66°C**, **PRECTOTCORR=0.05 mm**, etc. Across the previous 3 years up to 2024-04-27, the site accumulated **feat\_gdd\_base10=XXXX**, had **feat\_heat\_days\_30c=YY**, **feat\_frost\_days\_0c=ZZ**, etc.”