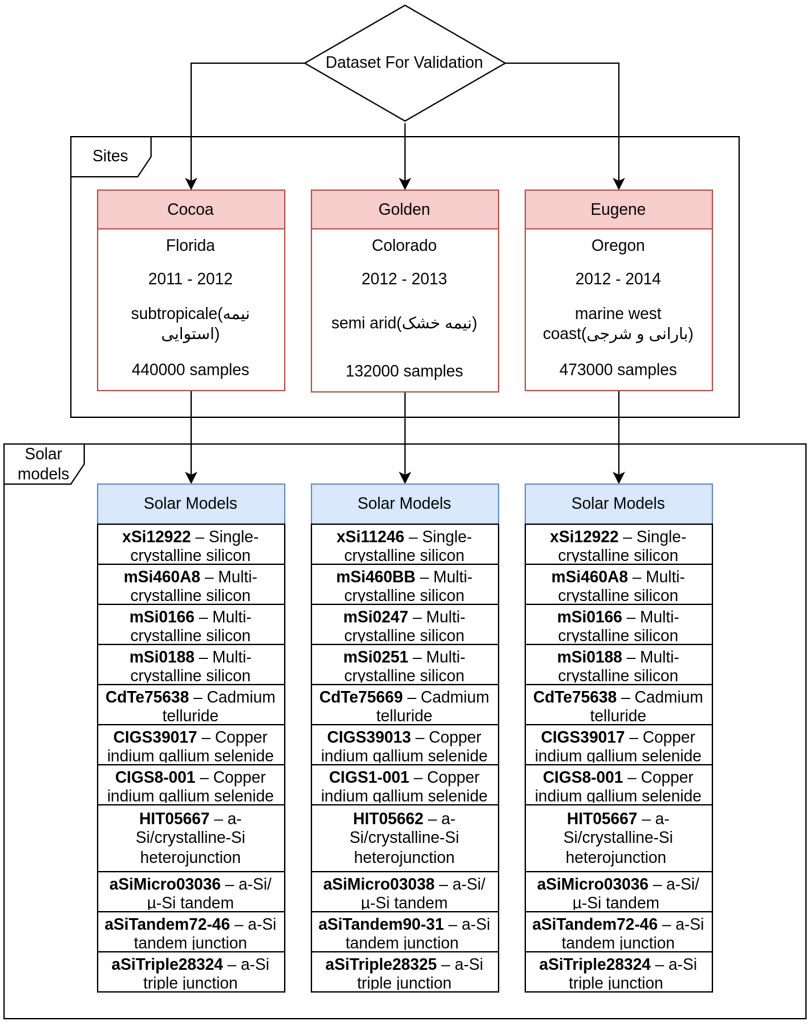
# General View of Dataset



# Columns

- **Date and Time :** Timestamp of the start of the I-V sweep
- **POA Irradiance (W/m²:)** Plane-of-array irradiance measured by a Kipp & Zonen CMP-22 pyranometer
- **POA Irradiance Uncertainty (%)** Expanded uncertainty of the POA irradiance, calculated from seven elemental components (MT5 DMM accuracy, calibration, incidence-angle error, temperature coefficient, linearity, stability, and zero-offset)

$$U_{Irr} = \sqrt{u^2_{Irr\text{-}DMM} + u^2_{Irr\text{-}Cal} + u^2_{Irr\text{-}Angle} + u^2_{Irr\text{-}Tcoef} + u^2_{Irr\text{-}Lin} + u^2_{Irr\text{-}Stab} + u^2_{Irr\text{-}TypeA}}$$

- **Back-Surface Temperature (C):** Measured with a Type T thermocouple bonded to the module back; read by the MT5 DMM channe
- **Back-Surface Temperature Uncertainty (C):** Absolute uncertainty from three sources—mounting method, MT5 measurement, and thermocouple accuracy

$$U_{T,\mathrm{mod}} = \sqrt{u^2_{T\text{-}Mount} + u^2_{T\text{-}DMM} + u^2_{TC}}$$

- **Isc (A):** Short-circuit current, determined by extrapolating the linear fit of the I-V curve near zero voltage
- **Isc Uncertainty (%)** Combined uncertainty from MT5 current accuracy, linear-fit error, resolution (bit-digit) error, and calibration source.

$$I_{sc,R} = \frac{1000}{E_{poa}} \times \frac{I_{sc}}{S_d} \times \frac{1}{1 + \alpha\,(T_c - 25)}$$

- **Voc Uncertainty (%)** From MT5 voltage accuracy and linear-fit error near Voc

$$U_{Voc} = \sqrt{u^2_{V\text{-}DMM} + \left(\frac{s_{\mathrm{VocFit}}}{N_{Voc}}\right)^2 + u^2_{V\text{-}Cal}}$$

- **Pmpp (W):** Maximum power

- **Pm Uncertainty (%)** Uncertainty combining voltage and current MT5 errors, polynomial-fit error, and calibration uncertainties

$$U_{P_{max}} = \sqrt{u^2_{V\text{-}DMM} + u^2_{I\text{-}DMM} + u^2_{P,\text{fit}} + u^2_{V\text{-}Cal} + u^2_{I\text{-}Cal}}$$

- **Imp (A):** Current at maximum power
- **Imp Uncertainty (%):** Estimated as halfway between the Isc and Pm uncertainties (≈0.31%)
- **Vmp (V):** Voltage at maximum power.
- **Vmp Uncertainty (%):** Estimated as halfway between the Voc and Pm uncertainties (≈0.30%)
- **FF (%):** Fill factor

$$FF = 100\frac{P_{mpp}}{V_{oc}.I_{sc}}$$

- **FF Uncertainty (%):** Propagated from Voc, Isc, and Pm uncertainties

$$U_{FF} = \sqrt{\left(\frac{U_{P_{max}}}{P_{max}}\right)^2 + \left(\frac{U_{V_{oc}}}{V_{oc}}\right)^2 + \left(\frac{U_{I_{sc}}}{I_{sc}}\right)^2}$$

- **delta CMP-22 POA (W/m²):** Difference between the CMP-22 reading just before and just after the 1˜s I-V sweep.
- **delta LI-COR POA (W/m²):** Same as above but using a fast-response LI-COR photodiode to flag variable-irradiance conditions
- **MT5 Cabinet Temperature (C):** Internal air temperature of the MT5 electronics cabinet, logged by the MT5 itself
- **Dry Bulb Temperature (C):** Ambient air temperature from a Vaisala WXT520 sensor
- **Dry Bulb Temp Uncertainty (C)**
- **Relative Humidity (%):** From the WXT520
- **Relative Humidity Uncertainty (%RH):** 3.46 %RH

- **Atmospheric Pressure (mbar):** Barometric pressure from the WXT520
- **Pressure Uncertainty (mbar):** ≈0.11% of reading at 1,013 mbar
- **Precipitation (mm):** Accumulated rainfall from the WXT520 tipping-bucket gauge
- **Direct Normal Irradiance (DNI, W/m²):** Measured by a CMP-22 on a two-axis tracker pointed at the sun; logged by CR1000. Acceptance angle 0.75°
- **DNI Uncertainty (%):** 1.36%
- **DNI Std Dev (W/m²):** 1-sec standard deviation over the 5-sec averaging window
- **Global Horizontal Irradiance (GHI, W/m²):** CMP-22 mounted level; tilt-angle error 0.1°, other methods as POA
- **GHI Uncertainty (%):** 2.14%
- **GHI Std Dev (W/m²):** 1-sec standard deviation (5-sec avg).
- **Diffuse Horizontal Irradiance (DHI, W/m²):** CMP-22 with a shading ball to block the sun disc, mounted on the same tracker as DNI
- **DHI Uncertainty (%):** 2.56%
- **DHI Std Dev (W/m²):** 1-sec standard deviation (5-sec avg)
- **Solar QA Residual (W/m²):** Difference between $I_h = I_{dh} + I_{dn} \cos\theta_z$ and $Ih$ flags inconsistent irradiance measurements.
- **Soiling Derate:** Daily ratio of clean vs. uncleaned module performance (1.000 = no loss)

- **Daily Maintenance Start/End (HH:MM):** Times when routine cleaning/inspection began and ended
- **Precipitation Prior to Maintenance (mm):** Rainfall accumulated before that day's maintenance.

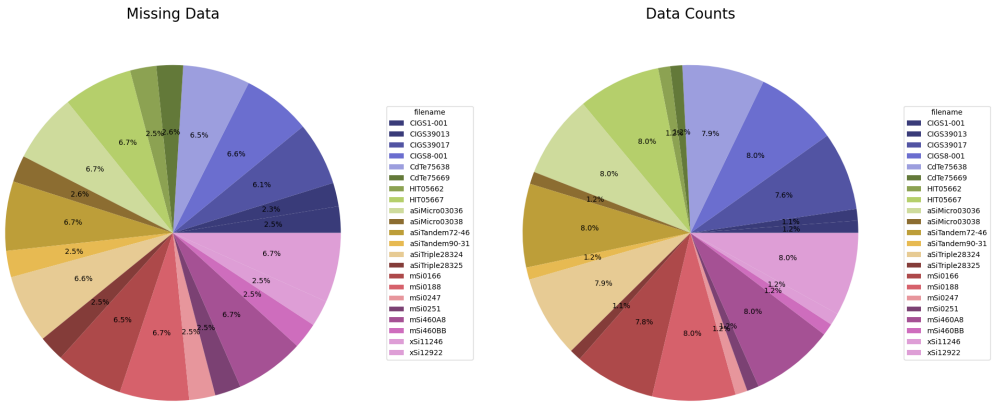
- **I-V Curve Data Pairs (N):** Number of (current, voltage) samples in the sweep, set by the MT5 tracer.
- **I-Values (A):** The N individual current readings, sampled at ≈200 Hz by the MT5.
- **V-Values (V):** The N corresponding voltage readings from the MT5.

## Task

هدف اصلی این گزارش تخمین زدن $V_{mpp}$ با استفاده از مدل های ماشین لرنینگ می باشد.

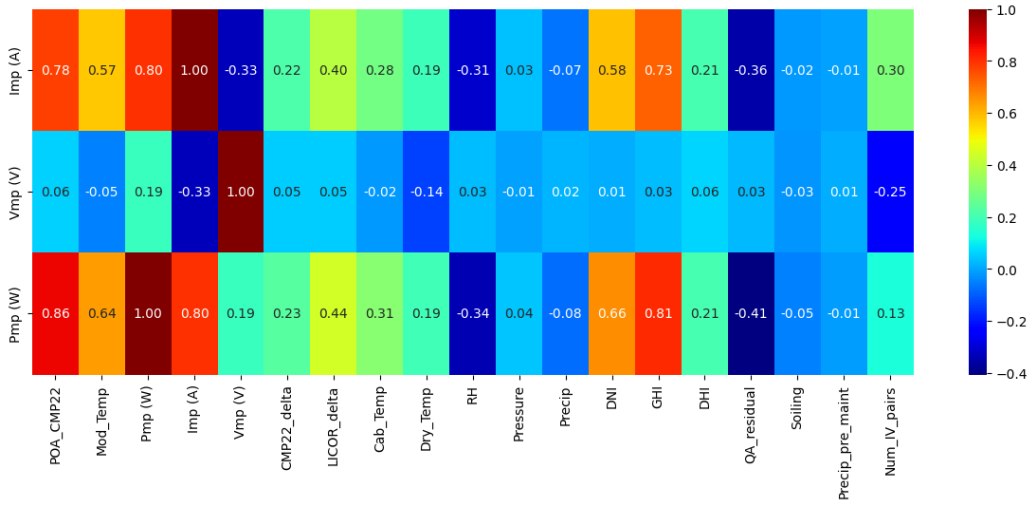
## Dataset Analysis

ابتدا بررسی های لازم برای پیدا کردن داده های گم شده را انجام دادم .وضعیت داده های گم شده به تفکیک مدل پنل ها به اینصورت است :



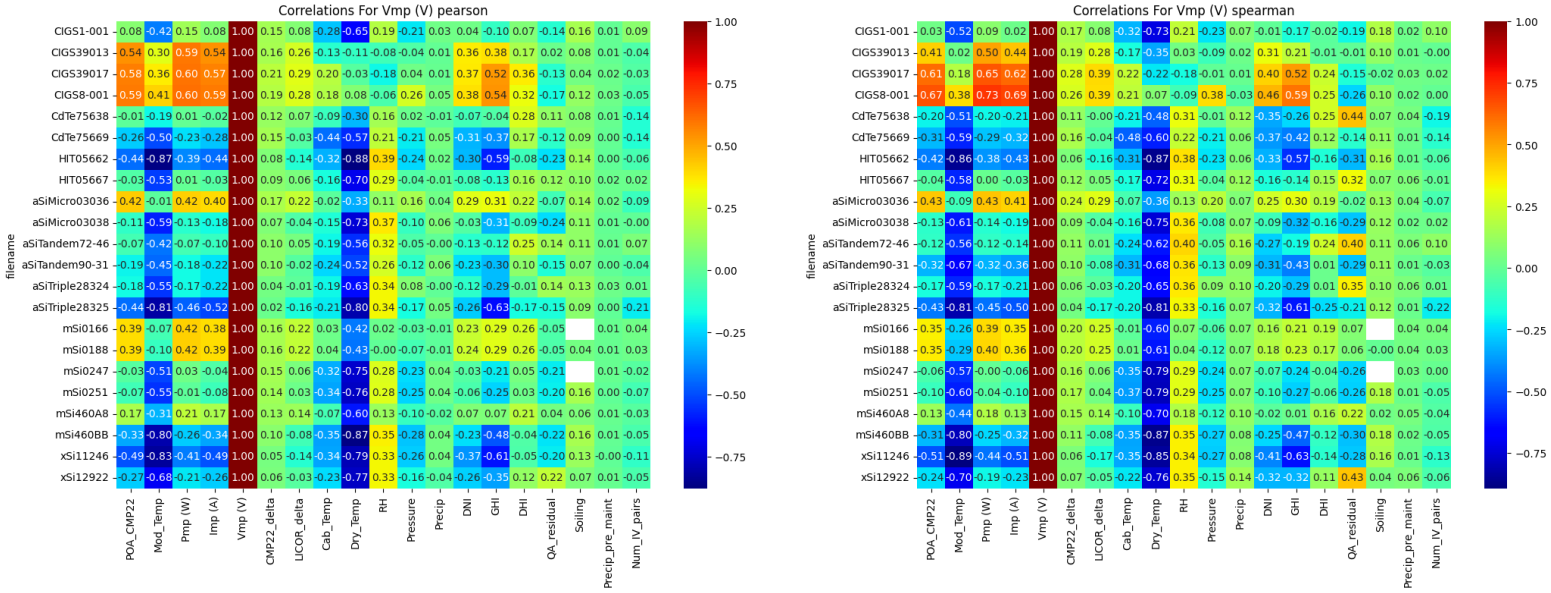

به دلیل اینکه بسیاری از داده های گم شده مربوط به ویژگی های mppt بودند مجبور شدم این داده هارا از دیتاست حذف کنم که در نهایت دیتاست از $1\,025\,599$ تا داده به $769\,595$ تا تغییر یافت .
ویژگی هایی که در مدل باعث leakage می شدند و به نوعی مستقیما از ویژگی های

mppt به دست می آمدند را حذف کردم.
در مرحله بعد یک heatmap از همبستگی بین داده های mppt و ویژگی های باقی مانده رسم کردم :

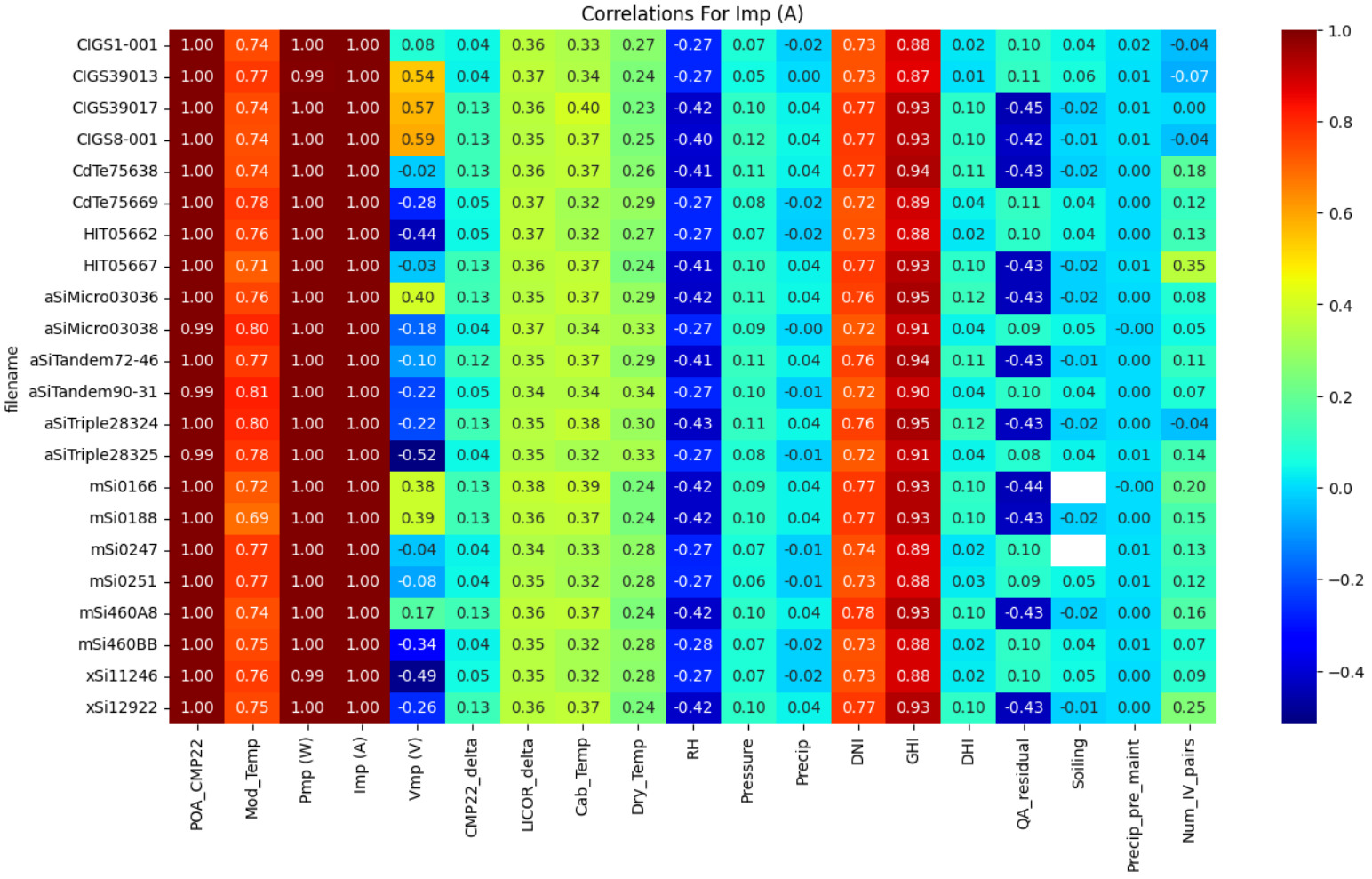



شکل ۱ : همبستگی کل داده ها نسبت به سه ویژگی $V_{mpp}, P_{mpp}, I_{mpp}$

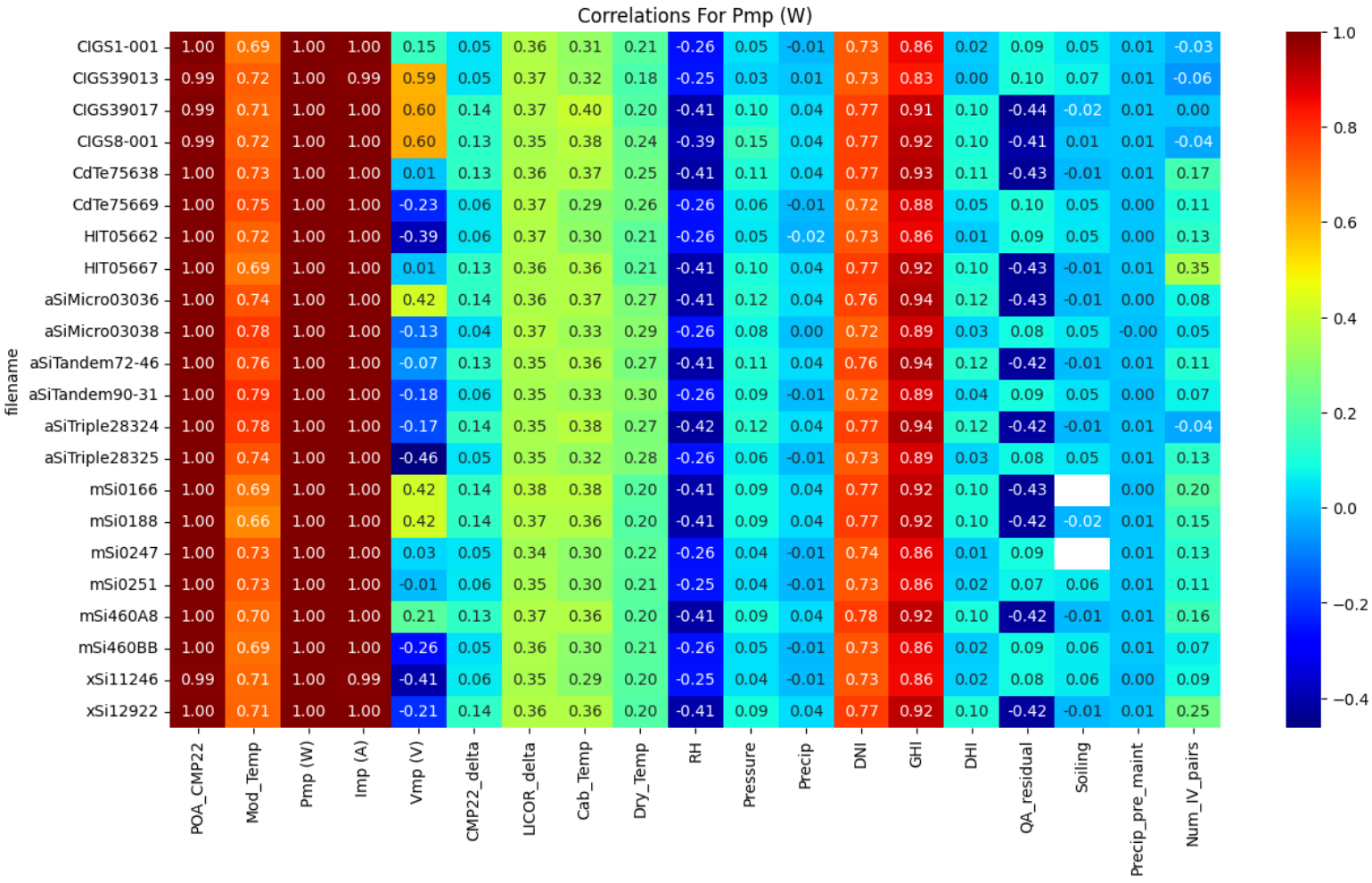
به دلیل اینکه همبستگی نسبت به کل دیتاست پایین به نظر میرسد سعی کردم تا ابتدا داده ها را بر اساس مدل پنل ها تفکیک کنم و سپس همبستگی را برای آن ها پیدا کنم :




شکل ۲: همبستگی نسبت به $V_{mpp}$



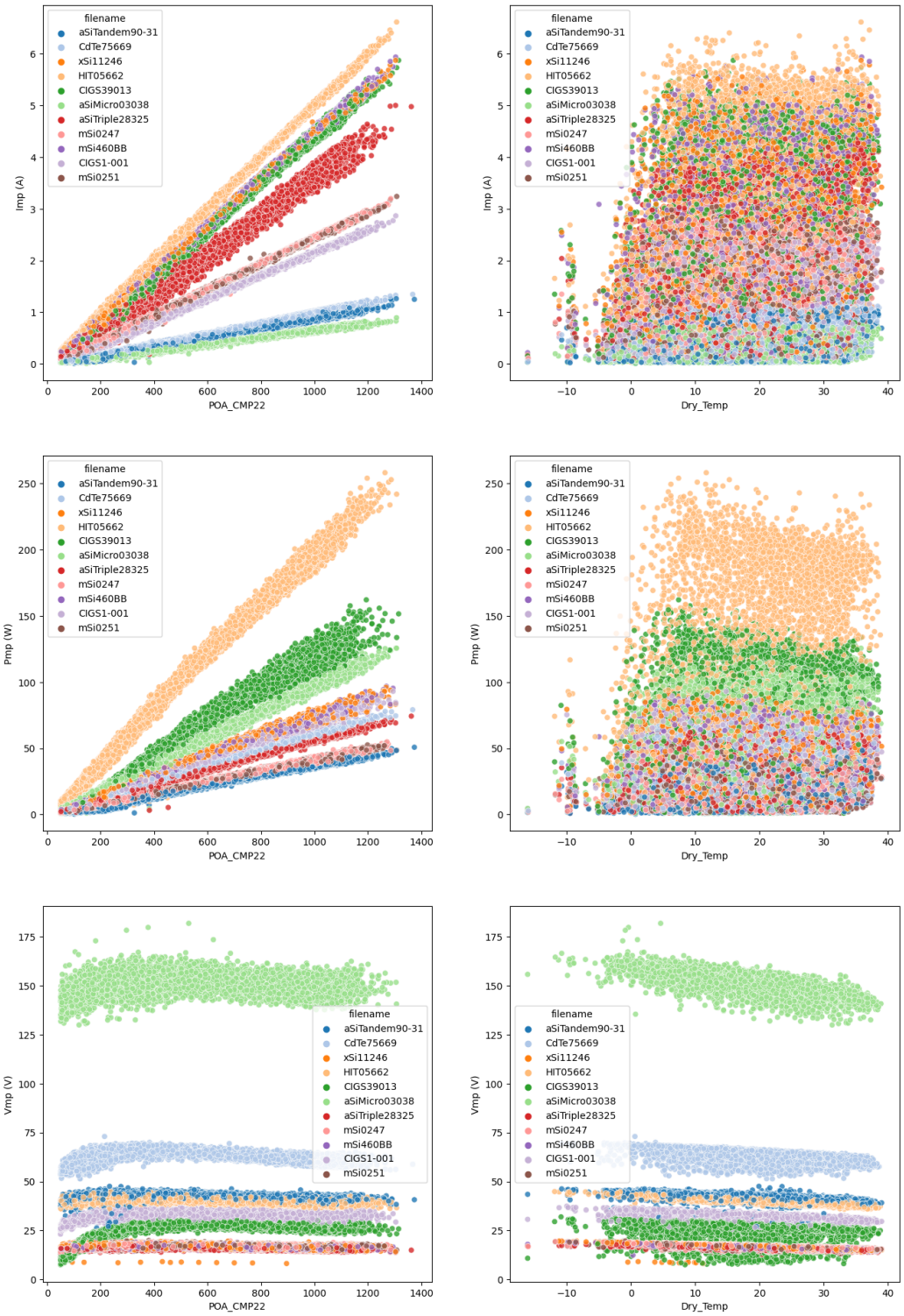

شکل ۳: همبستگی نسبت به $I_{mpp}$




شکل ۴: همبستگی نسبت به $P_{mpp}$

به وضوح میزان همبستگی بالاتر رفته است پس به نظر میرسد که هرکدام از مدل های پنل های خورشیدی دارای توزیع های مختلفی از ویژگی های mppt میباشد پس به همین دلیل بهتر است که برای هر نوع پنل خورشیدی یک مدل به صورت جداگانه آموزش داده شود.

یک scatter plot نیز نسبت به دو ویژگی POA_CMP22 و Dry_Temp برای هر سه تا متغیر mppt رسم شده است:



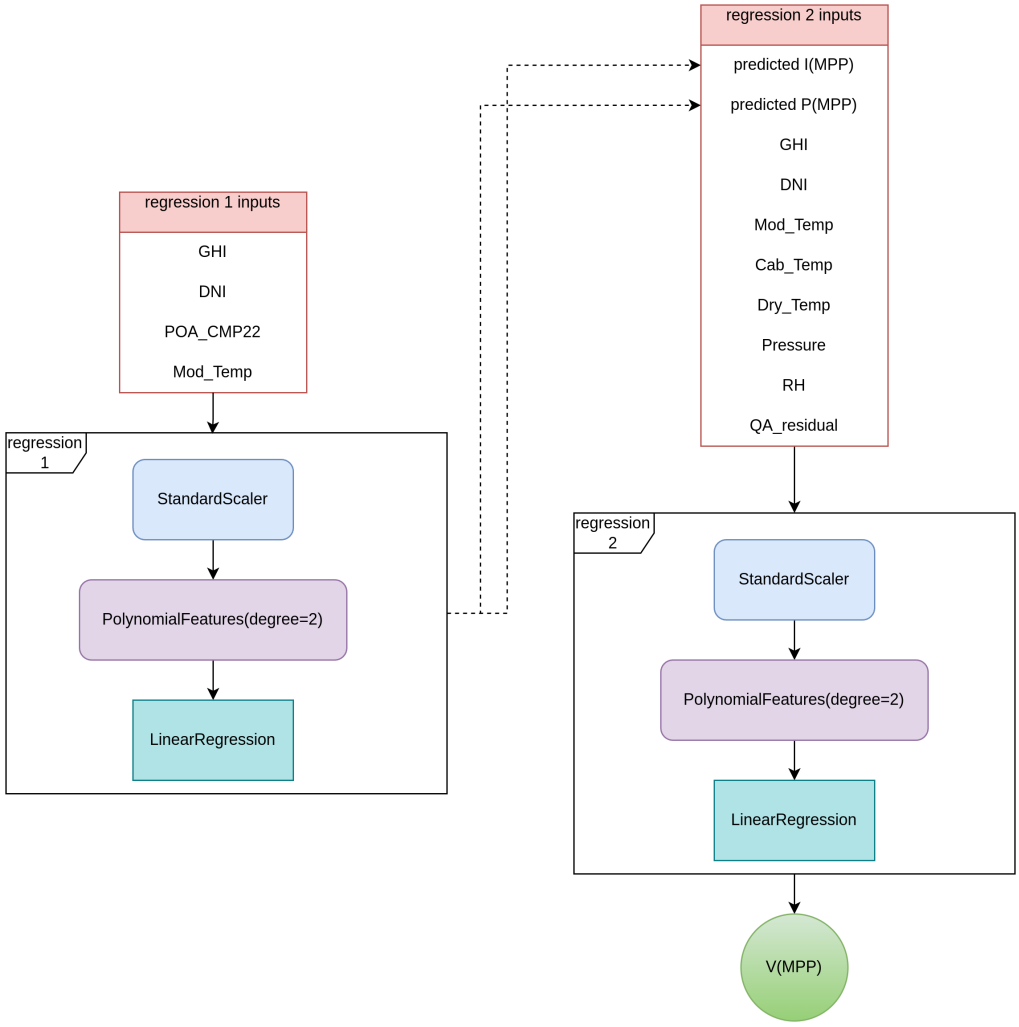

شکل ۵: scatter plot بین دو ویژگی خوب و متغیر های mppt

این نمودار نیز نشان میدهد که رفتار mppt بر روی هر مدل جداگانه سولار پنل متفاوت است .
همچنینی به نظر میرسد که ویژگی $V_{mpp}$ با ویژگی های مربوط به دما ارتباطات بهتری دارد .

# Models

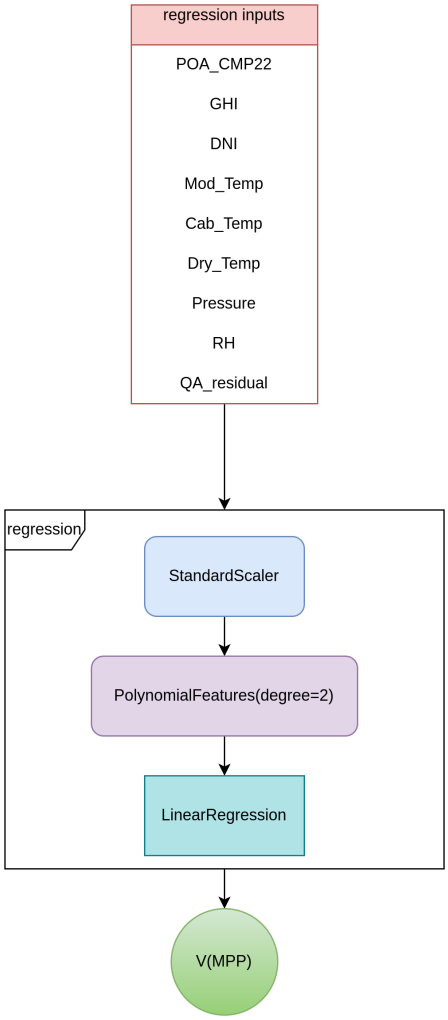
## Regression

با استفاده از regression دو مدل درست کردم؛
مدل اول ترکیبی از دو مدل رگرسیونی است که بخش اول سعی می کند دو ویژگی $I_{mpp}, P_{mpp}$ را تخمین بزند و سپس با استفاده از این تخمین ها و برخی ویژگی های مربوط به دما و ویژگی هایی که همبستگی بالایی با $V_{mpp}$ دارند $V_{mpp}$ را تخمین بزند :




شکل ۶: مدل اول متشکل از دو مدل رگرسیونی

در مدل بعدی سعی کردم به صورت مستقیم $V_{mpp}$ را تخمین بزنم یعنی در اینجا فقط از یک مدل استفاده شده است :

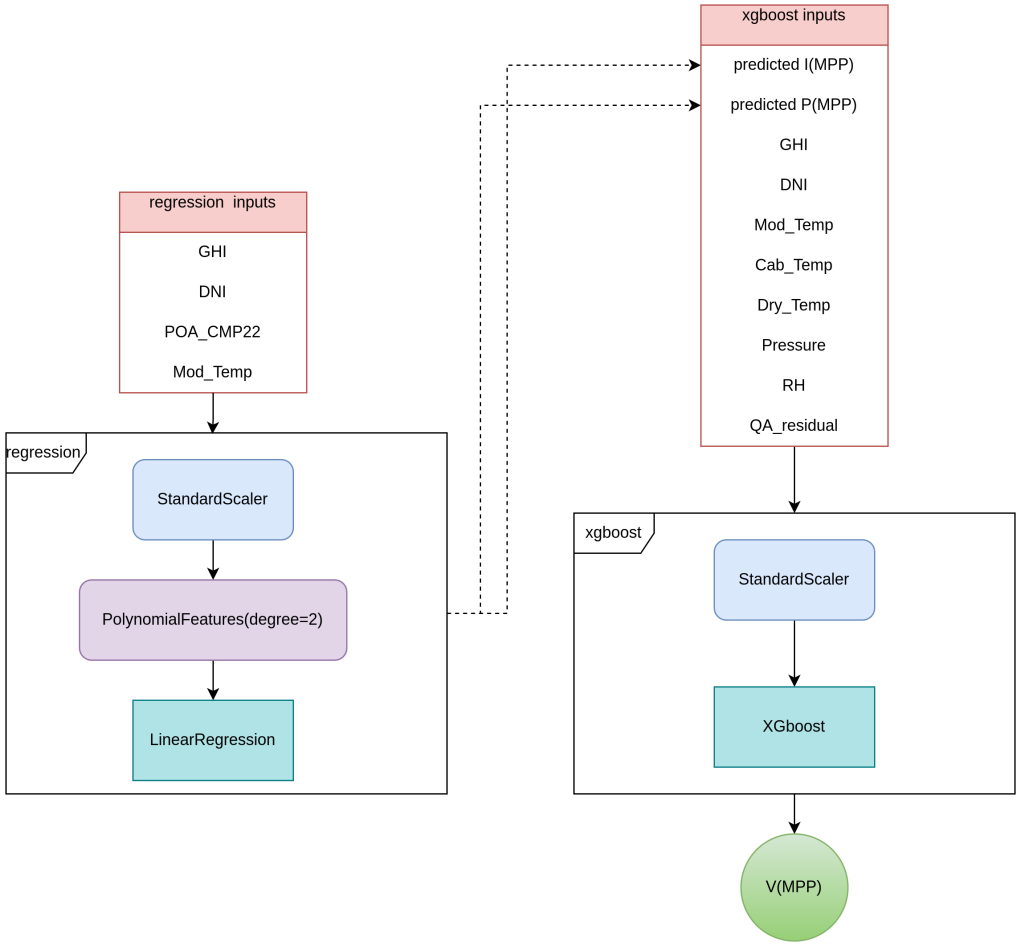



شکل ۷: مدل دوم متشکل از یک مدل رگرسیونی

## HistGradBoost

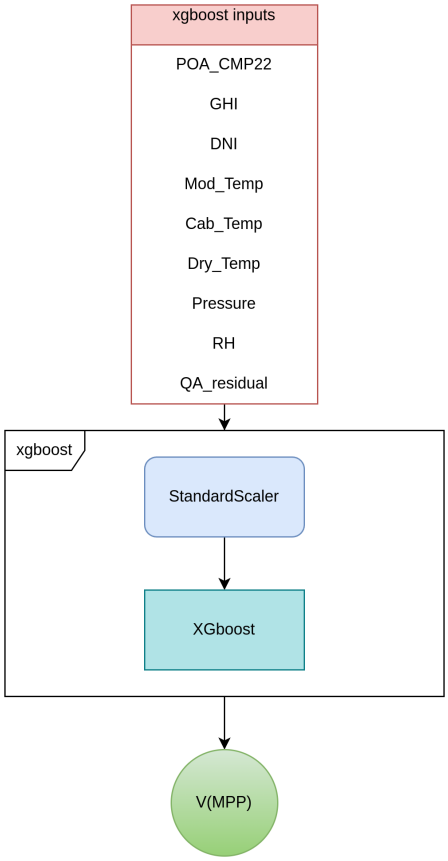
دو مدل دیگر به وسیله HistGradBoost طراحی کردم:

مدل اول نیز مثل مدل اول رگرسیونی دارای دو بخش است : بخش اول سعی می کند دو ویژگی $I_{mpp}, P_{mpp}$ را تخمین بزند و بخش دوم سعی می کند به وسیله HistGradBoost $V_{mpp}$ را تخمین بزند




شکل ۸: مدل اول متشکل از دو مدل رگرسیونی + HistGradBoost

مدل دوم سعی میکند که $V_{mpp}$ به صورت مستقیم با استفاده از HistGradBoost تخمین بزند:




شکل ۹: مدل دوم متشکل از یک مدل HistGradBoost

پارامتر های تنظیم شده برای HistGradBoost به این صورت است:

- max_iter : 200
- max_depth: None
- learning_rate: 0.1
- random_state :42

# Results

| Model | R2 Train | R2 Test | ADJ-R2 Train | ADJ-R2 Test |
|---|---|---|---|---|
| double regression | 0.8907 | 0.8835 | 0.8906 | 0.8830 |
| regression | 0.8324 | 0.8264 | 0.8323 | 0.8259 |
| **regression+HistGradBoost** | 0.9801 | 0.9541 | 0.9801 | 0.9536 |
| HistGradBoost | 0.9278 | 0.9123 | 0.9277 | 0.9119 |

بهترین نتیجه ها مربوط به مدل **regression+HistGradBoost** میباشد:

Models Summary

| **Model Name** | **R۲ Train** | **R۲ Test** | **Total Time** | **Pred Time** |
|---|---|---|---|---|
| AdaBoostRegressor | 0.89604 | 0.87883 | $4m24.16s$ | $0.14767s$ |
| **HistGradientBoostingRegressor** | 0.98926 | 0.95877 | $49.2s$ | $0.14421s$ |
| RandomForestRegressor | 0.93233 | 0.90886 | $3m0.07s$ | $0.07549s$ |
| LinearRegression | 0.89064 | 0.88308 | $19.64s$ | $0.01798s$ |
| XGBRegressor | 0.93011 | 0.91368 | $32.44s$ | $0.03163s$ |