Solar Farm – Datasheet

1. Data source

The original dataset was sourced from Kaggle:
https://www.kaggle.com/code/vinayjshetty/solar-power-generation-forecast-using-different-ml/notebook
It was collected from two solar power plants in India over a period of 34 days. The data includes both sensor and weather readings recorded at 15-minute intervals.
After extensive processing—including merging, cleaning, and enriching with non-linear synthetic variables—a single dataset was produced, containing over **60,000 records**.

2. Motivation

The aim of this dataset is to demonstrate how machine learning techniques can be applied to help a solar farm predict its power output using weather and sensor data.

Due to the number of transformations applied during preprocessing, the dataset is now assumed to be largely synthetic. Despite that, it effectively captures relevant behaviours and is especially useful in highlighting the cost-benefit dynamics of dust cleaning on solar panels.

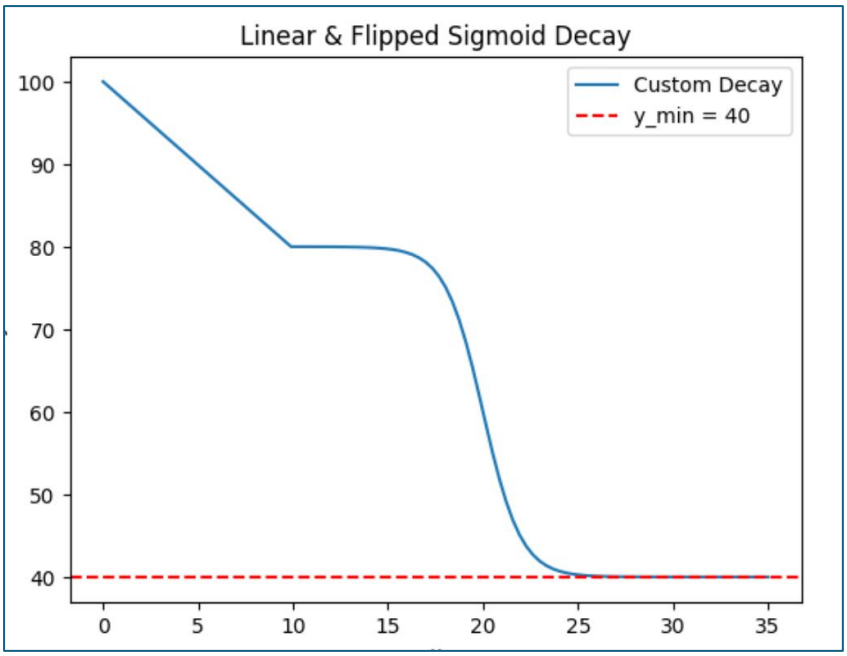
3. Composition

Data Points used in ML training:

DateSolar	Date
Date_time	DateTime
Source_Key	Text
DC_Power	Numeric
AC_Power	Numeric
Total_Yield	Numeric
Yield_Total_Intraday	Numeric
Ambient_Temperature	Numeric
Module_Temperature	Numeric
Irradiation	Numeric
DaySinceLastCleanup	Numeric
PercentageOutput	Numeric
Yield_Total_Intraday_WithDust	Numeric

- Missing values or bad data has been removed.
- Final rowcount: 68,665
- The synthetic column Yield_Total_Intraday_WithDust has been created using a non linear pattern as function of original Total_Yeld and DaySinceLastCleanup.

See below the decay of Yield based on DaysSinceLastCleanup.



There is no confidential data.

4. Collection Process

The data is a subset (34 days) of all data that would normally be collected.
In real life we need to capture all seasons because the temperature may have an impact that is not linear.
With a bit more research this seasonal data can be fabricated and potentially train a model that is satisfactory even before getting all data collected.

5. Uses

The synthetic decay pattern is pure hypothetical. This can be adapted to make it closer to real life patterns.

6. Distribution

Dataset can be distributed for educational purposes and is subject to Kaggle T&C.
The patterns in dataset can be used as scaffolding for synthetic data.

7. Maintenance

Dataset is not maintained.