



# SOLAR IRRADIANCE CLASSIFICATION

Dmitrii Kuzimin, Yaroslava Bryukhanova, Timofey Brayko, Artemii Miasoedov, Artur Rakhmetov

# PROJECT OVERVIEW

Weather prediction is an extremely important industry in modern life. Weather forecasts are used not only to decide what clothes to wear each day—they also impact agriculture, construction, transportation, and the energy sector.

We aim to infer solar irradiance using simple, low-cost sensors like thermometers and clocks. This approach replaces expensive, high-maintenance devices such as pyranometers.

Our solution can reduce sensor complexity, maintenance costs, and improve data reliability.

To test this concept, we decided to infer solar irradiance (solar radiation) using other parameters such as temperature, time, and ozone levels. Solar irradiance is the amount of solar energy reaching the Earth's surface and is a key metric in agriculture, construction, and energy production.



# DATASET OVERVIEW

## SIZE & FORMAT





9.29 million rows in CSV format  
(1.5 GB), CC-BY-4.0 license, 2024  
data.

## TARGET VARIABLE

Ground solar radiation categorized  
as low, medium, or high.

## KEY FEATURES

Latitude; Longitude; Ozone, Nitrogen and Carbon monoxide concentration;  
Organic carbon concentration; Surface pressure; Planetary boundary layer  
height; Temperature; Wind speed; Wind direction; Ground solar radiation; Cloud  
fraction; Month; Day; Hours.

<div>AirNOW_03 float64</div> <div></div>	<div>Lat_airnow float64</div> <div></div>	<div>Lon_airnow float64</div> <div></div>	<div>Lat_cmaq float64</div> <div></div>
35	29.489082	-81.276833	
52	40.5802	-74.199402	
41	39.12886	-84.504044	
50	34.63596	-82.810669	
85	34.100132	-117.491982	
52	38.41024	-82.432434	
55	34.66959	-118.130692	
62	42.718601	-109.753098	
45	41.975601	-72.386703	
56	35.054401	-119.4039	
67	35.345608	-118.851822	
63	34.066429	-118.226753	
41	37.2267	-121.9786	
55	35.356613	-119.062614	
18	29.5203	-95.392502	
53	36.9533	-120.034103	
28	30.3503	-95.425003	
48	40.931396		
..	..		

# DATA PIPELINE ARCHITECTURE

1

## DATA LOADING

Load data from HuggingFace for further preprocessing.

2

## INITIAL PREPROCESSING

Analyze the correlation of provided columns, removed redundant features and check values on constraints.

3

## INGEST TO DATABASE

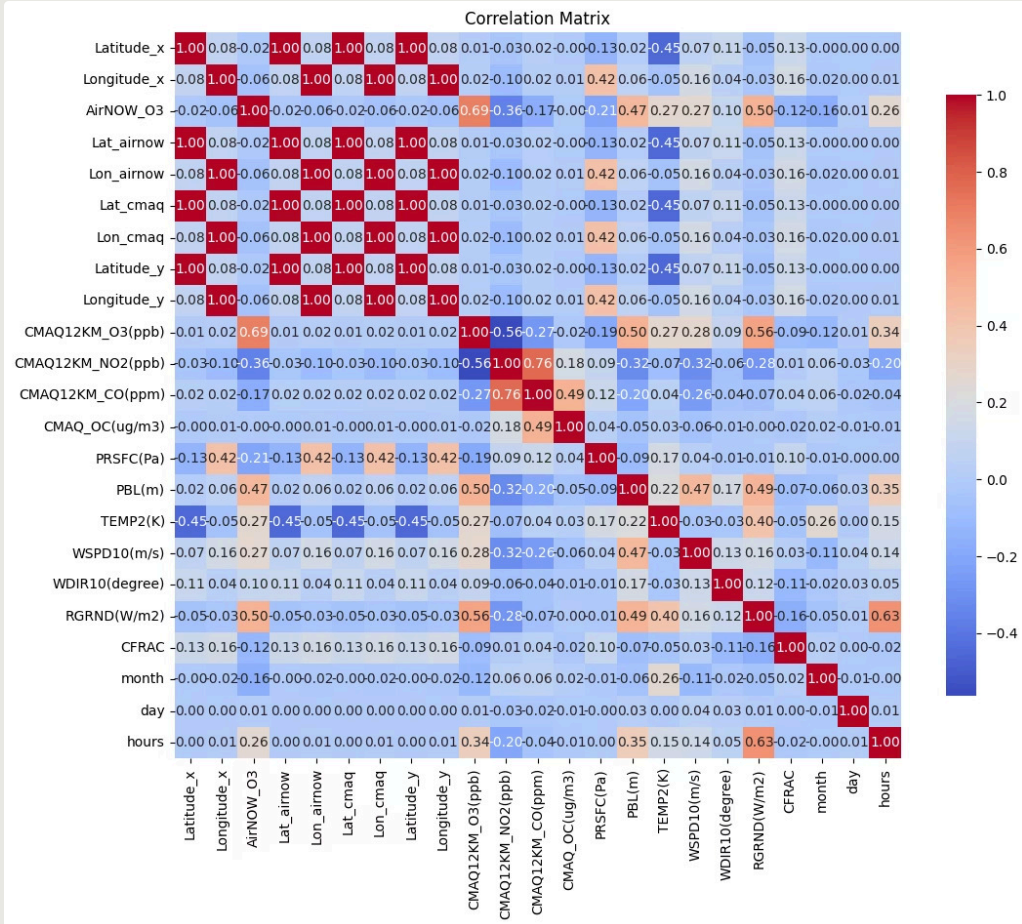
Data ingested into SQL, transferred to HDFS as Parquet, Hive for querying.

4

## FEATURE ENGINEERING

PySpark used for scaling, cyclical and coordinate transformations, label discretization.

# INITIAL PREPROCESSING



During the initial preprocessing phase, we constructed a correlation matrix to identify and remove highly correlated or redundant features. This analysis revealed several duplicate columns related to geographical coordinates, including *Latitude\_x*, *Lat\_airnow*, *Lat\_cmaq*, *Latitude\_y*, as well as *Longitude\_x*, *Lon\_airnow*, *Lon\_cmaq*, and *Longitude\_y*.

We eliminated these redundant coordinate columns, retaining ***Latitude\_x*** and ***Longitude\_x*** as the primary geographical coordinates for the dataset.



# DATABASE SCHEMA & TABLES

## STATIONS TABLE

Contains station ID and coordinates (latitude, longitude).

Identifier	Latitude	Longitude
120350004	29.489082	-81.276833
360850111	40.5802	-74.199402
390610040	39.12886	-84.504044
840450070006	34.63596	-82.810669
60712002	34.100132	-117.491982

## RECORDS TABLE

Stores observations linked to stations with environmental data and timestamps.

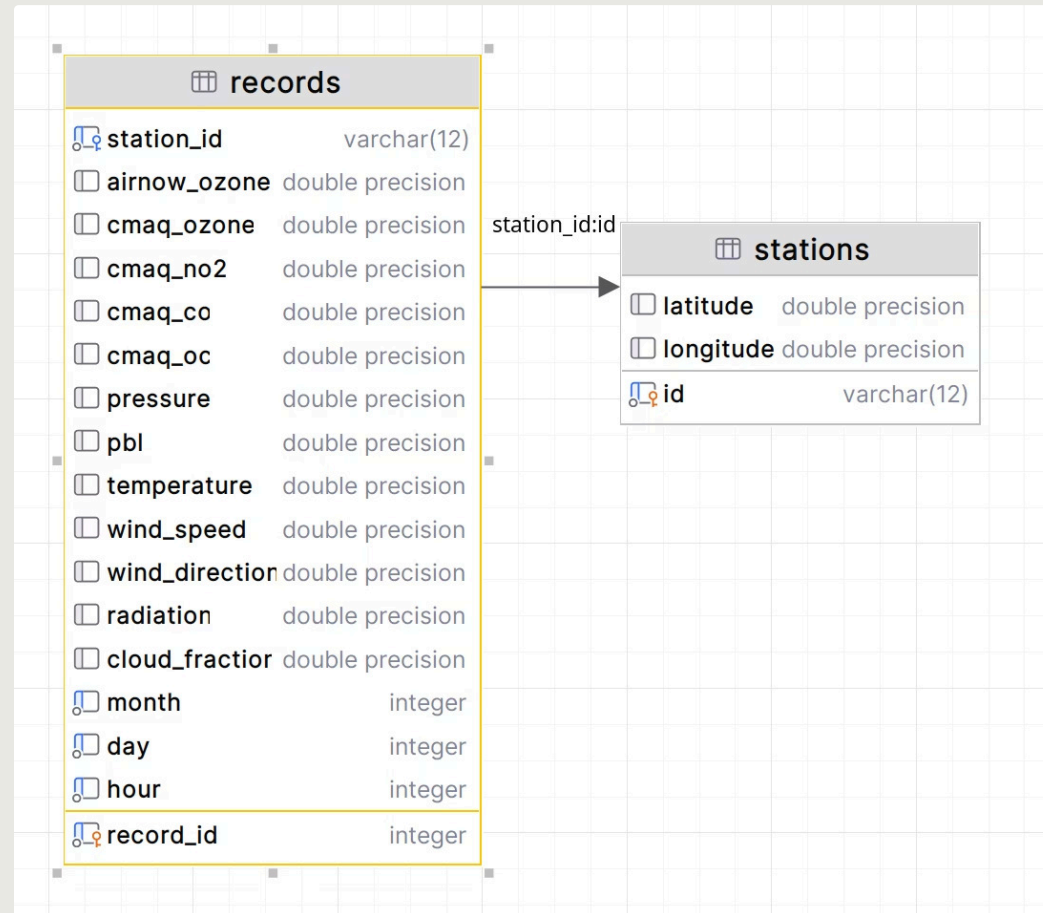
record_id	1	2	3
station_id	120350004	360850111	390610040
airnow ozone	35	52	41
cmaq ozone	25	45	45
cmaq no2	0	2	1
cmaq co	84	147	137
cmaq oc	1	2	5
pressure	101384	100797	98891
pbl	1319	1163	2170
temperature	305	299	302
wind speed	6	4	6
wind direction	242	189	339
radiation	576	416	696
cloud fraction	0	0	0
month	8	8	8
day	1	1	1
hour	21	21	21

## PARTITIONING & BUCKETING

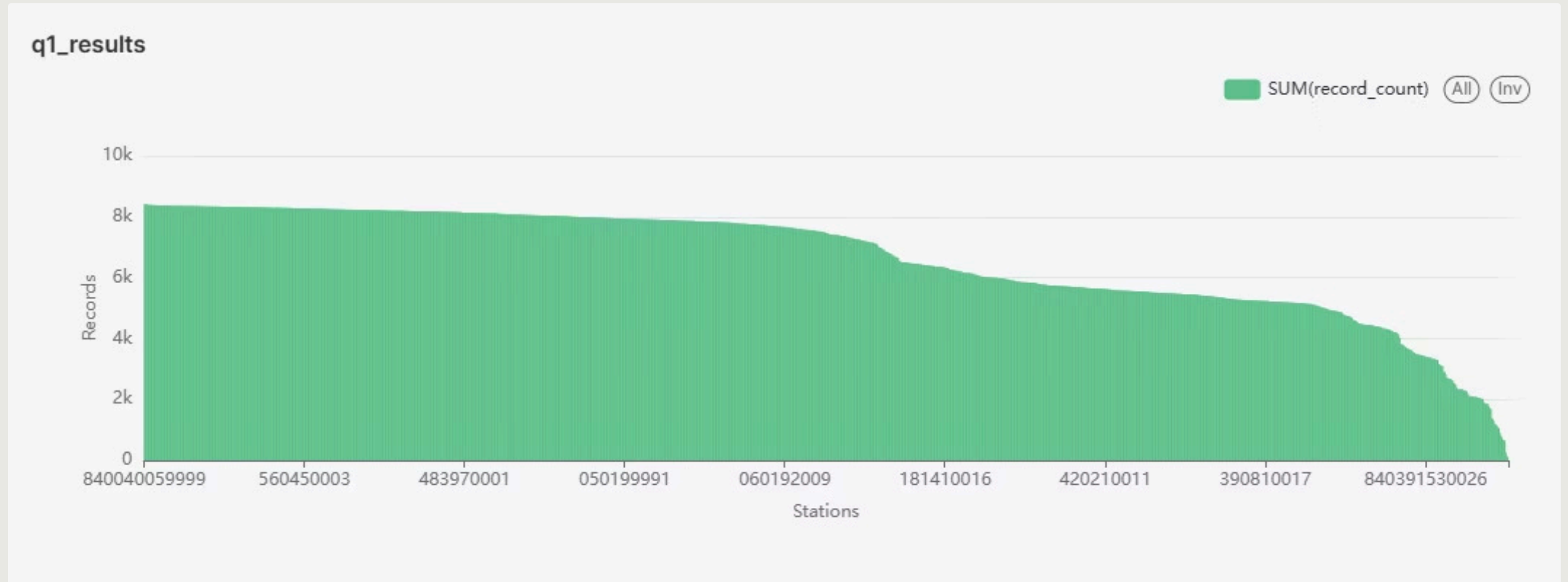
Records partitioned by month/day and bucketed by station ID for efficiency.

```
-- Create the optimized records table with partitioning and bucketing
CREATE EXTERNAL TABLE records (
  record_id INT,
  station_id STRING,
  airnow_ozone DOUBLE,
  cmaq_ozone DOUBLE,
  cmaq_no2 DOUBLE,
  cmaq_co DOUBLE,
  cmaq_oc DOUBLE,
  pressure DOUBLE,
  pbl DOUBLE,
  temperature DOUBLE,
  wind_speed DOUBLE,
  wind_direction DOUBLE,
  radiation DOUBLE,
  cloud_fraction DOUBLE,
  hour INT
)
PARTITIONED BY (month INT, day INT)
CLUSTERED BY (station_id) INTO 2 BUCKETS
STORED AS PARQUET
LOCATION 'project/warehouse/records_optimized'
```

# ENTITY RELATION DIAGRAM



# RECORDS PER STATION





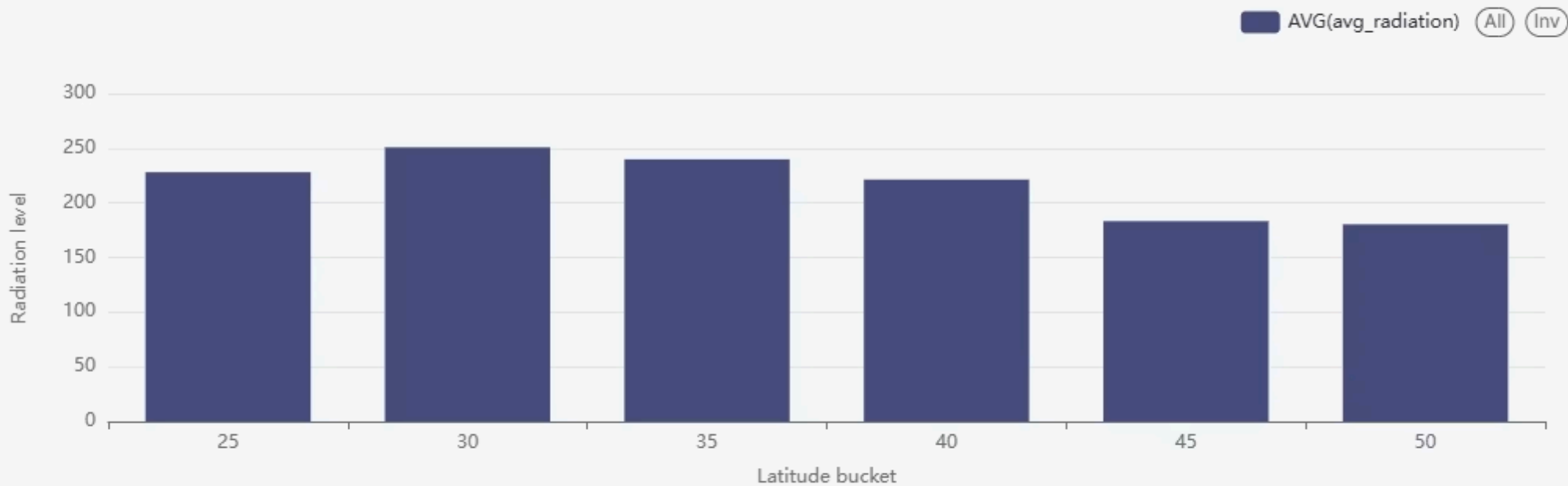
# STATIONS DISTRIBUTION

q2\_results



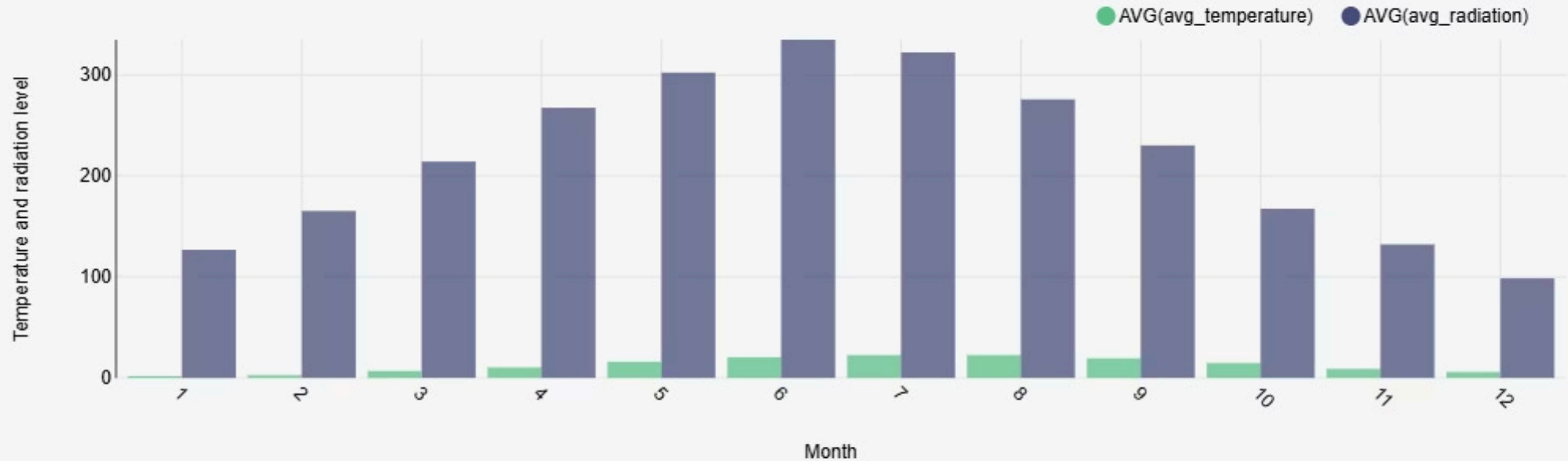
# RADIATION BY LATITUDE

q3\_results

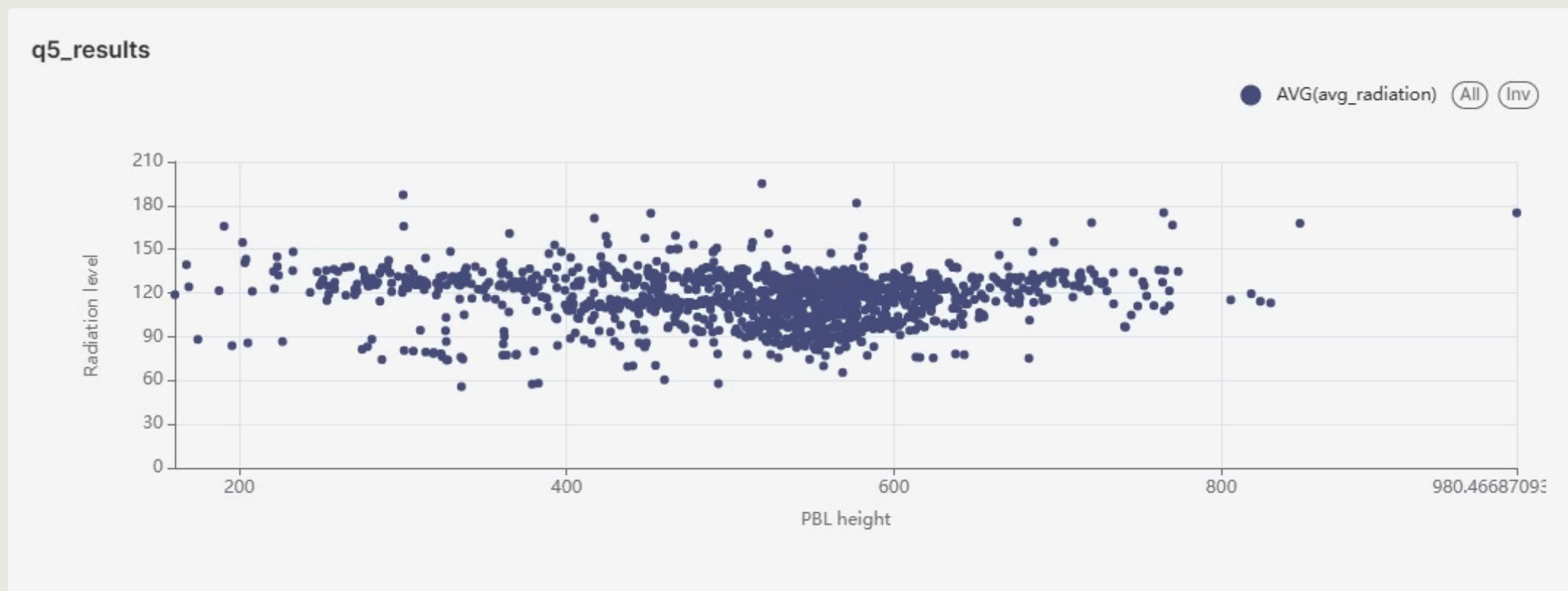


# TEMPERATURE AND RADIATION ACROSS MONTHS

q4\_results

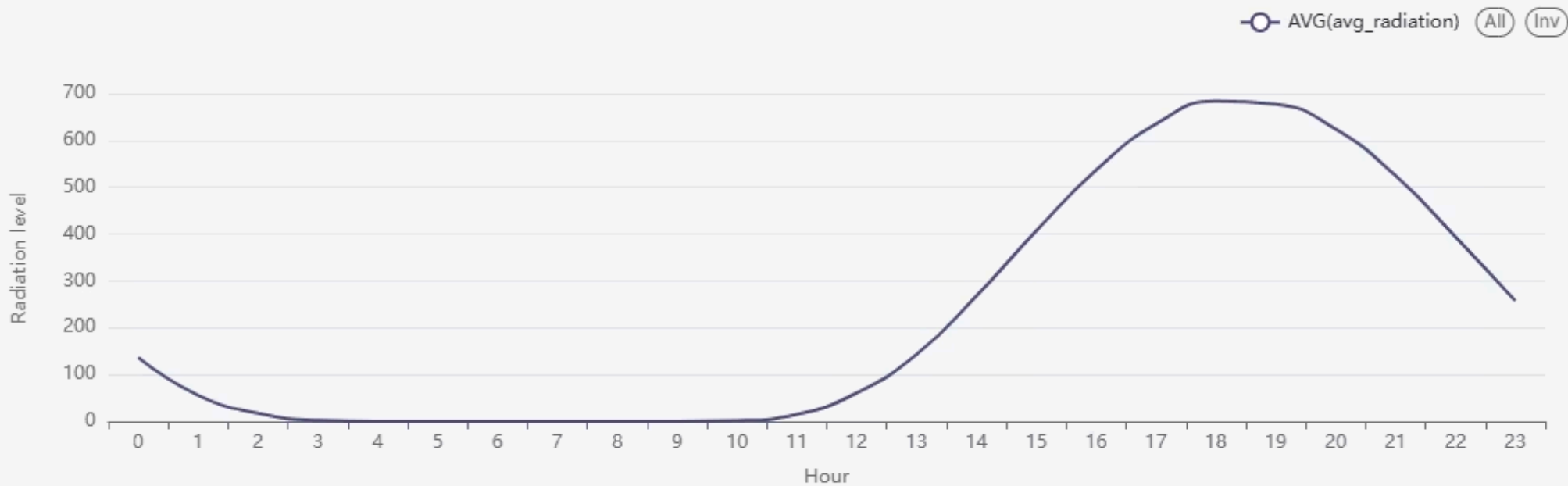


# AVERAGE PLANETARY BOUNDARY LAYER (PBL) HEIGHT AND AVERAGE RADIATION FOR EACH STATION



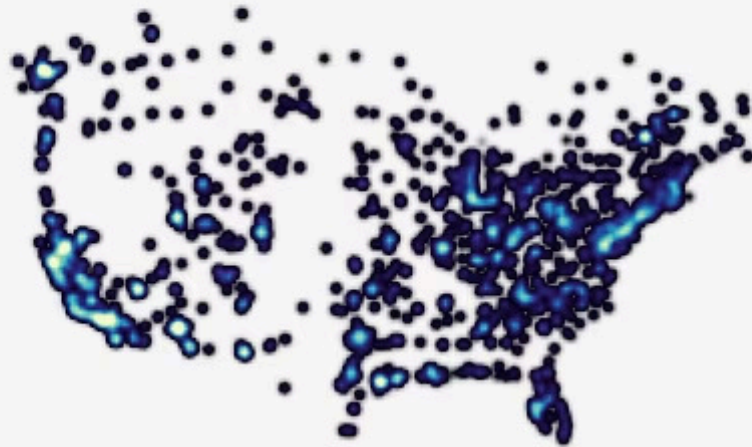
# RADIATION BY HOUR

q6\_results



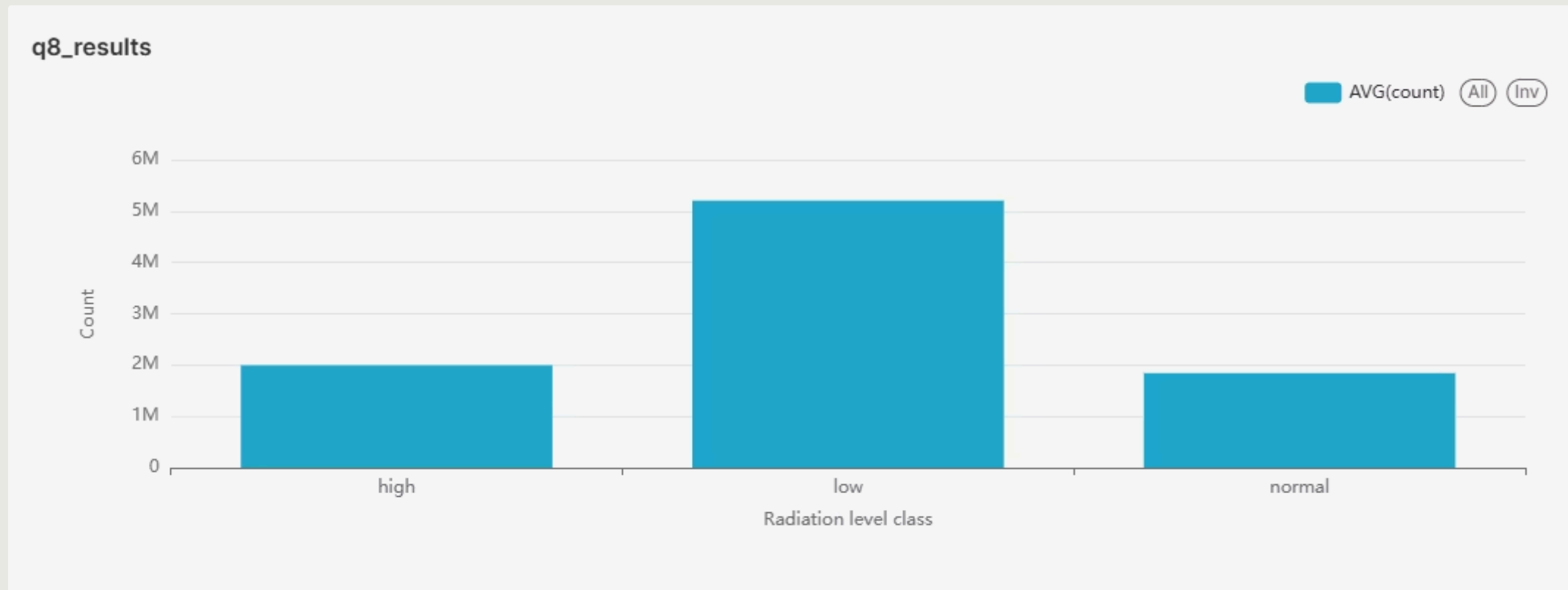
# PEARSON CORRELATION BETWEEN DAILY AVERAGE TEMPERATURE AND SOLAR RADIATION

q7\_results





# RADIATION CLASS DISTRIBUTION



# FEATURE ENGINEERING

$$\begin{aligned}X &= (N + h) \cos \phi \cos \lambda \\Y &= (N + h) \cos \phi \sin \lambda \\Z &= \left( \frac{b^2}{a^2} N + h \right) \sin \phi\end{aligned}$$

## COORDINATE CONVERSION

Latitude and longitude transformed to Earth-Centered, Earth-Fixed system.

$$\begin{aligned}Month_{sin} &= \sin \frac{2 \times \pi \times x}{12} \\Month_{cos} &= \cos \frac{2 \times \pi \times x}{12}\end{aligned}$$

## CYCLICAL ENCODING

Month, day, hour encoded with sine and cosine functions.

$$Label = \begin{cases} 0, & \text{Radiation} \leq 100 \\ 1, & 100 < \text{Radiation} \leq 500 \\ 2, & \text{Radiation} > 500 \end{cases}$$

## LABEL DISCRETIZATION

Solar radiation binned into low, medium, and high categories.

$$z = \frac{x - \mu}{\sigma}$$

## SCALING

StandardScaler applied to numerical features like ozone, temperature, wind speed.

# MACHINE LEARNING MODELS & TRAINING

## MODEL SELECTION

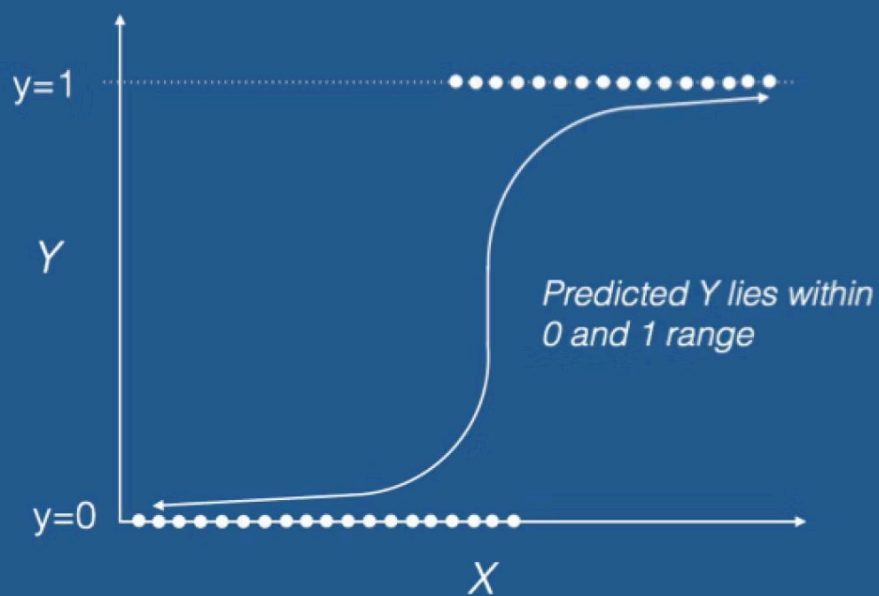
We tested multiple classification algorithms to predict solar irradiance levels.

- Logistic regression
- Multilayer perceptron
- Naive bayes
- Random forest
- One-vs-Rest Linear SVC

Cross-validation with 3 folds.



## Logistic Regression



# LOGISTIC REGRESSION

Grid search:

- `regParam = [0.01, 0.1, 1.0]`
- `elasticNetParam = [0.0, 0.5, 1.0]`
- `threshold = [0.3, 0.5, 0.7]`

Grid size: 27

K-fold: 3

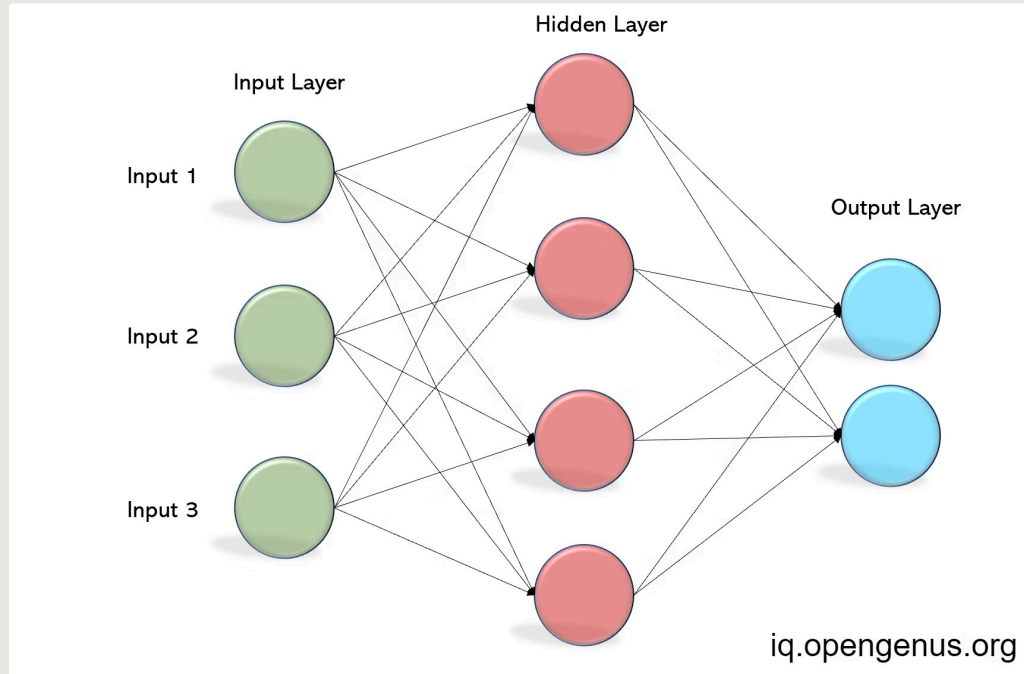
# MULTILAYER PERCEPTRON

Grid search:

- layers = [
  - [features, 8, labels]
  - [features, 16, labels],
  - [features, 24, labels]
- stepSize = [0.01, 0.05, 0.1]
- solver = ["l-bfgs", "gd"]

Grid size: 18

K-fold: 3



# GAUSSIAN NAIVE BAYES CLASSIFIER

"Gaussian" because this is a normal distribution

This is our prior belief

$$P(\text{class} | \text{data}) = \frac{P(\text{data} | \text{class}) \times P(\text{class})}{P(\text{data})}$$

We don't calculate this in naive bayes classifiers

ChrisAlbon

## NAIVE BAYES

Grid search:

- smoothing = [0.5, 1.0, 1.5]
- threshold
  - [1.0, 1.0, 1.0], # Neutral (no weighting)
  - [1.5, 1.0, 0.5], # Favor class 0
  - [0.5, 1.0, 1.5], # Favor class 2

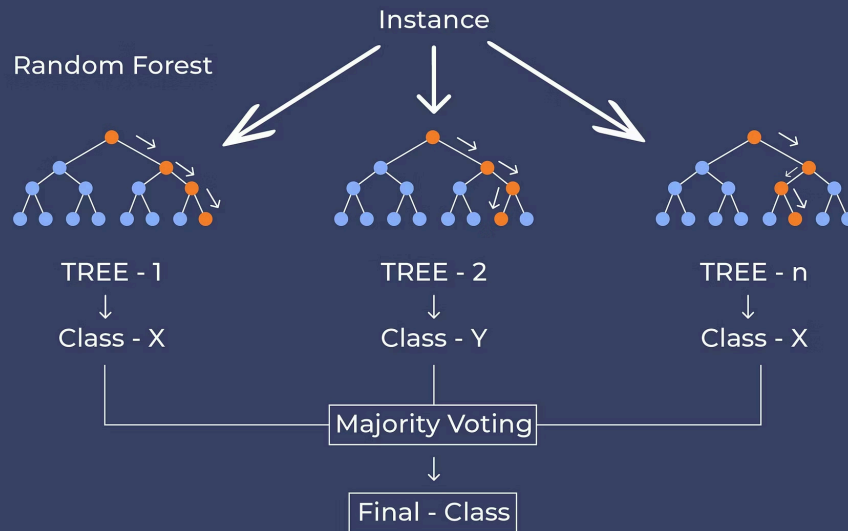
Grid size: 9

K-fold: 3



# RANDOM FOREST

## CLASSIFICATION



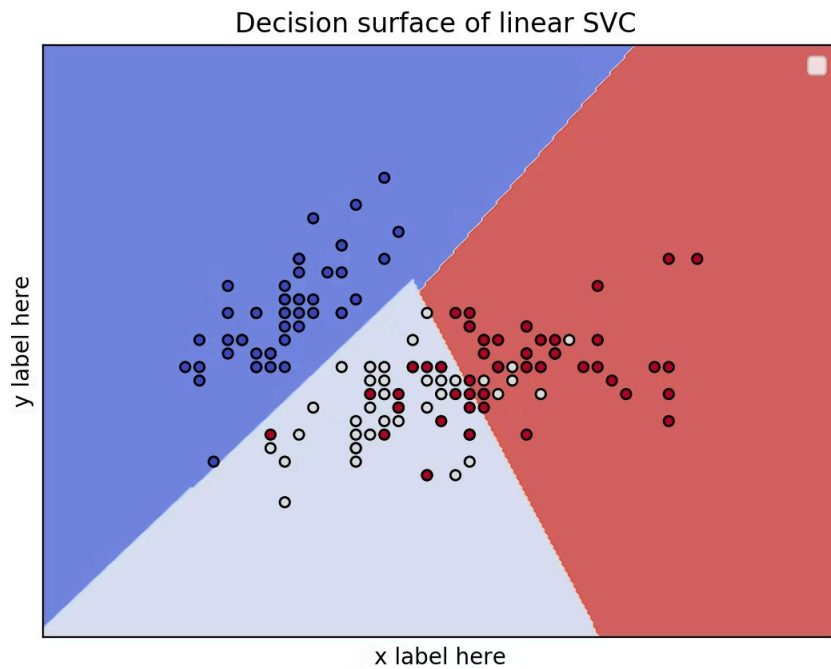
## RANDOM FOREST

Grid search:

- numTrees = [16, 32, 64]
- maxDepth = [3, 5, 8]
- featureSubsetStrategy = ["auto", "sqrt", "log2"]

Grid size: 27

K-fold: 3



## LINEAR SVC

Grid Search

- `regParam = [0.01, 0.1, 1.0]`
- `aggregationDepth = [1, 2]`

Grid size: 6

K-fold: 3

# EVALUATION RESULTS



Model	Accuracy	F1 Score
Random Forest	0.85	0.84
Logistic Regression	0.71	0.67
Linear SVC	0.70	0.62
Naive Bayes	0.62	0.56
Multilayer Perceptron	0.57	0.42



# PROJECT SUMMARY & REFLECTION

## SUMMARY

Compact stations can infer solar irradiance using simple sensors and ML.

## CHALLENGES

Handling large data, feature engineering, and model tuning were complex tasks.

## RECOMMENDATIONS

Further tuning of Random Forest and expanding inference to other parameters.