MACHINE LEARNING FOR SOLAR ENERGY ANALYSIS & PREDICTION

Data-driven Weather Analysis in Aswan

Perihan Yasser (320230065)

Maria Emad (320230063)

REAL-WORLD DATA REQUIRES ROBUST PREPROCESSING AND ANALYSIS

GOAL

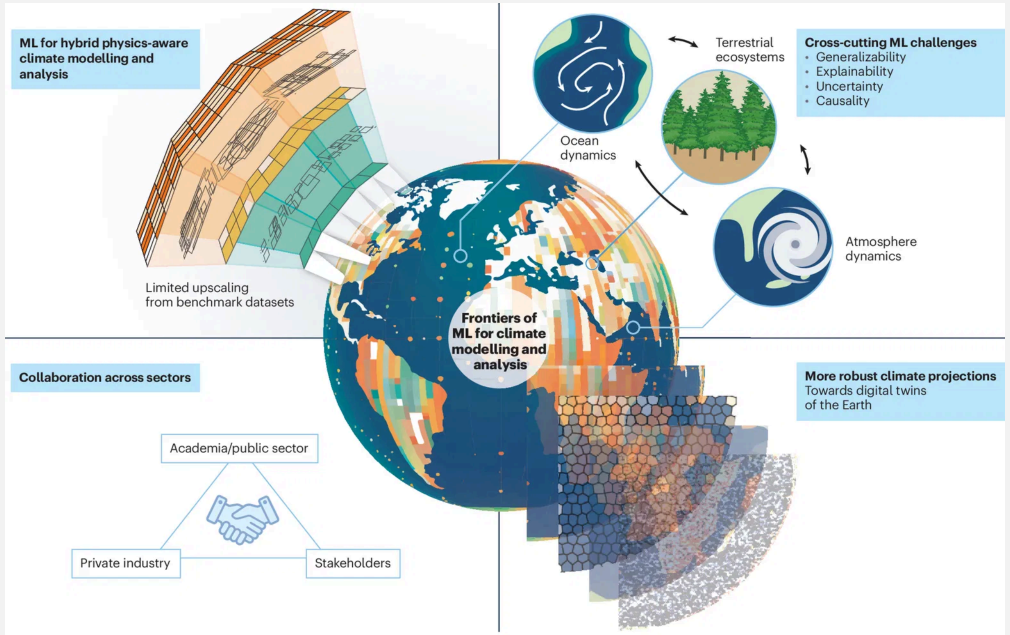
Study the relationship between meteorological features and solar energy output to build reliable predictive models.

CHALLENGES

Addressing missing values, noise, and non-linear relationships that impact model performance.

PROJECT SCOPE

- Data Preprocessing & EDA
- Statistical Hypothesis Testing
- Feature Selection & Reduction
- Model Training and Evaluation


ML for hybrid physics-aware climate modelling and analysis
Limited upscaling from benchmark datasets
Ocean dynamics
Terrestrial ecosystems
Cross-cutting ML challenges
• Generalizability
• Explainability
• Uncertainty
• Causality
Atmosphere dynamics
Frontiers of ML for climate modelling and analysis
Collaboration across sectors
Academia/public sector
Private industry
Stakeholders
More robust climate projections
Towards digital twins of the Earth

DATASET COVERS A FULL YEAR OF METEOROLOGICAL RECORDS IN ASWAN

SOURCE

Aswan, Egypt

TIME RANGE

Apr 2021 - Apr 2022

TOTAL RECORDS

398 Samples

KEY FEATURES

- Avg Temperature
- Humidity
- Wind Speed
- Pressure
- Solar (PV) Output

SYSTEMATIC PREPROCESSING ENSURES DATA CONSISTENCY AND QUALITY

PHASE 01

Data Cleaning & Imputation

Duplicate Removal: 28 redundant records were identified and removed to prevent bias.

Missing Values: 24 missing dates were addressed using a combination of forward fill, backward fill, and linear interpolation.

PHASE 02

Feature Engineering

Extraction of temporal features from the date column to capture seasonal and daily patterns, resulting in a final set of 10 features.

PHASE 03

Dataset Finalization

The final processed dataset contains 394 high-quality records, ensuring temporal consistency for model training.

PHASE 04

Binning Strategy

Continuous variables were categorized to facilitate classification tasks:

Solar: Low, Med, High

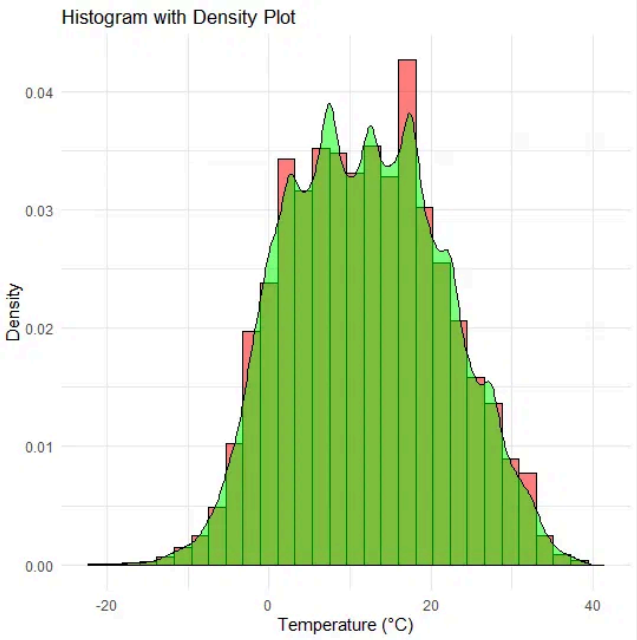
Temp: Cool, Warm, Hot

Humidity: Dry, Mod, Humid

DESCRIPTIVE STATISTICS REVEAL BALANCED SOLAR OUTPUT DISTRIBUTIONS

FEATURE	MIN	MAX	MEAN	STD DEV	SKEW
Avg Temp	51.1	102.7	80.97	14.05	-0.39
Humidity	7.4	47.7	24.16	9.95	0.65
Wind	4.4	17.1	10.35	2.51	0.27
Solar(PV)	8.58	40.04	24.89	7.63	-0.03

Key Insight: Solar output shows near-zero skewness (-0.03), indicating a highly balanced distribution across the Aswan dataset, ideal for robust machine learning modeling.



TEMPERATURE DENSITY ANALYSIS

STATISTICAL TESTS CONFIRM SIGNIFICANT METEOROLOGICAL DEPENDENCIES

CHI-SQUARE TEST

Temperature vs. Solar Category

P-VALUE

0.0186

Confirmed dependency between temperature levels and solar energy output (α=0.05).

T-TEST

High vs. Low Humidity

P-VALUE

0.0000

Solar output differs significantly between humidity levels, indicating strong correlation.

ANOVA TEST

Solar Output Across Months

F-STATISTIC

78.1107

Highly significant seasonal variation confirms solar output is heavily month-dependent.

WIND AND HUMIDITY ARE THE PRIMARY DRIVERS FOR SOLAR PREDICTION

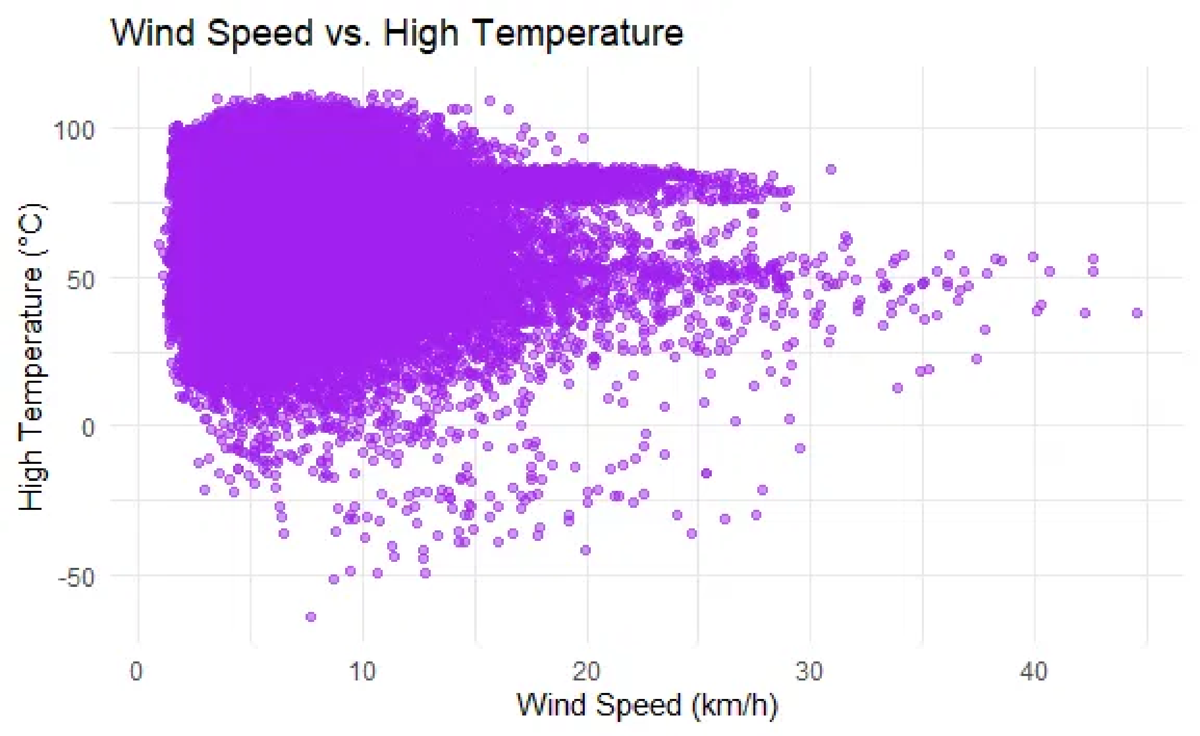
ANOVA F-TEST IMPORTANCE

#	Feature	F-Score
1	Wind Speed	13.35
2	Humidity	5.88
3	Average Dew Point	5.45

DIMENSIONALITY REDUCTION

PCA Component 1 explains 51.44% of variance.

First 3 components capture 94.38% of total information.



Visualization of Feature Relationships: Wind Speed vs. High Temperature

DIVERSE MACHINE LEARNING MODELS WERE EVALUATED FOR PERFORMANCE

CLASSIFICATION SUITE

- Naive Bayes & Decision Tree
- K-Nearest Neighbors (Multiple Distances)
- Logistic Regression
- Linear Discriminant Analysis (LDA)

REGRESSION ANALYSIS

- Linear Regression

Methodology: 80/20 Train-Test Split
Validation: 5-Fold Cross-Validation

DECISION TREE OUTPERFORMS OTHER CLASSIFICATION MODELS

MODEL ARCHITECTURE	TRAIN ACC	TEST ACC	PRECISION	F1-SCORE
Decision Tree	0.7333	0.7342	0.7136	0.7168
K-NN (Euclidean)	0.7270	0.6835	0.6748	0.6739
PCA + Decision Tree	0.7397	0.6456	0.6334	0.6352
LDA	0.5524	0.5570	0.5548	0.5433
Naive Bayes	0.5238	0.4684	0.4572	0.4617

PERFORMANCE LEADER

The Decision Tree model achieves the highest test accuracy and F1-score, showing superior capture of non-linear weather patterns in Aswan.

DIMENSIONALITY IMPACT

PCA combined with Decision Tree reduced performance, indicating original features contain critical information PCA components may miss.

DECISION TREE ACHIEVES HIGH ACCURACY WITH MINIMAL OVERFITTING

TEST ACCURACY

73.42%

> The model demonstrates exceptional generalization, outperforming all other tested architectures on unseen weather data.

OVERFITTING GAP

-0.0008

> A near-zero gap between training and testing accuracy confirms the model is perfectly fitted and highly stable.

CLASS-WISE F1-SCORES

High Solar Output	0.81
Medium Solar Output	0.74
Low Solar Output	0.18

PERFORMANCE INSIGHT

While the model excels at identifying High and Medium solar conditions, the low F1-score for the 'Low' category suggests a class imbalance in the dataset that requires future attention.

NON-LINEAR RELATIONSHIPS LIMIT LINEAR REGRESSION EFFECTIVENESS

REGRESSION PERFORMANCE

0.0473

R-SQUARED (R^2) SCORE

6.8255

MEAN ABSOLUTE ERROR (MAE)

> *The low R^2 score confirms that linear models fail to capture the complex, non-linear dependencies inherent in Aswan's weather data.*

PROBABILISTIC INSIGHTS

100%

Probability of High Solar Output when conditions are Cool and Dry.

66.7%

Probability of Medium Solar Output during Hot and Humid periods.

*Based on Empirical Conditional Probabilities from Bayesian Analysis.

DECISION TREE IS THE OPTIMAL MODEL FOR ASWAN SOLAR PREDICTION

BEST MODEL PERFORMANCE

The Decision Tree classifier is the most reliable, achieving 73.42% accuracy with near-zero overfitting.

KEY WEATHER DRIVERS

Wind speed and Humidity are statistically more significant predictors than temperature alone in the Aswan region.

FUTURE DIRECTIONS

Addressing class imbalance for 'Low' solar energy categories and exploring non-linear regression models to improve continuous value prediction.