IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 1, Ser. 2 (Jan. – Feb. 2026), PP 29-41
www.iosrjournals.org

Feature-Based Analysis Of Machine Learning Models For Hourly Solar Irradiance Prediction In An Urban Region

Ashok S. Sangle, Prapti D. Deshmukh And Rajesh Dhumal
Department Of Computer Science And IT, Dr. Babasaheb Ambedkar, Marathwada University, Chhatrapati Sambhaji Nagar, India
Department MGM's Dr. Pathrikar College Of Computer Science And IT, Chhatrapati Sambhaji Nagar, India, Symbiosis Institute Of Geo-Informatics (SIG) Symbiosis International University, Pune, India.

Abstract:
This study conducts a feature-based analysis to evaluate the predictive performance of set of machine learning (ML) regression models for solar irradiance prediction using multivariate data from Chhatrapati Sambhaji Nagar, India. The dataset, sourced from NASA POWER and NOAA, spans January 2001 to July 2023 and includes various variables such as temperature, pressure, humidity, wind speed, and sunrise/sunset times. After data preprocessing and exploratory analysis, we trained six set of models—Linear Regression, Decision Tree, Random Forest, Gradient Boosting, Extra Trees, and K-Nearest Neighbors—each under two feature conditions: a full nine-feature set and a reduced subset of model-specific relevant features selected via coefficient analysis, permutation importance, and built-in attribute evaluators. Model performance was assessed using R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Maximum Error. Results indicate that ensemble methods—particularly Extra Trees and Random Forest models —outperformed and better than the other simple models in accuracy and robustness. Targeted feature selection not only maintained predictive performance but also improved model interpretability and efficiency. This work offers valuable insights into feature engineering and model selection for solar energy forecasting, supporting enhanced regional energy planning strategies.

Keywords: *Feature Selection, Meteorological Data, Machine Learning (ML) Regression, Model Performance Evaluation, Predictive Analytics, Solar Irradiance Prediction.*

Date of Submission: 21-01-2026 Date of Acceptance: 31-01-2026

I. Introduction

The rooftop solar infrastructure sustainability, design optimization of photovoltaic systems, and efficient integration with the grid and demand-side energy management depend on the solar irradiance conditions as its predictions. The correct prediction of energy systems enables better planning of energy resources and decreases dependence on fossil fuels, while helping achieve worldwide renewable energy targets [1], [2]. Machine learning (ML) techniques have proven effective for short-term solar irradiance prediction because they excel at modeling intricate linear and non-linear patterns between meteorological feature variables and time-dependent factors [3], [4].

The predictive accuracy of ML models depends on both the quality of selected input features and the appropriate choice of algorithms that match regional weather-climate patterns and temporal factors [3], [4]-[8]. The research assessed six collective regression-based ML models including Linear Regression and Decision Tree Regressor and Random Forest Regressor and Gradient Boosting Regressor and Extra Trees Regressor and K-Nearest Neighbors using extensive multivariate hourly interval weather dataset from Chhatrapati Sambhaji Nagar, India. The analysis uses a 22.7-year dataset to examine feature-based model performance.

The Scikit-learn framework in Python provides a versatile and efficient toolkit for implementing a wide range of machine learning algorithms, supporting tasks from preprocessing and model selection to evaluation and deployment [5], [8]-[11]. Here, the data preprocessing steps for datetime handling and normalization, as well as exploratory data analysis and feature engineering techniques, engineered features such as Hours_of_light and Rel_time in our study.

The estimation of required features is a critical preprocessing step that identifies the most informative variables for a model, directly influencing its predictive accuracy and efficiency. Determining the importance of these features provides vital interpretability, revealing the underlying drivers of the model's decisions and ensuring its outputs are both robust and trustworthy [3], [6]-[8], [11],[12]. The evaluation strategy involved assessing each model's performance on two feature sets: the full set and a reduced set comprising only the most relevant features, as identified by algorithm-specific importance methods.

The research aims to achieve three main goals: (i) evaluate and compare model performance through R^2, RMSE, MAE and Maximum Error metrics; (ii) determine the most important meteorological factors that affect solar irradiance levels; and (iii) Visualize the prediction accuracy underscoring the profound relevance of feature selection, illustrating its direct effect on the model's ability to capture the complex dynamics of solar irradiance. To validate this, predictions for a continuous period were plotted against the actual measured values. This research not only develops robust solar estimation approaches but also provides a transferable predictive modeling framework; by explicitly quantifying feature importance, it offers a foundation for additional deep exploration in similar renewable energy context.

II. Methodology

This section describes the study area, data acquisition strategy, preprocessing steps, exploratory data analysis (EDA), feature engineering approaches and ML models employed along with their evaluation protocols and model performance on full vs. selected features (see Figure 1).

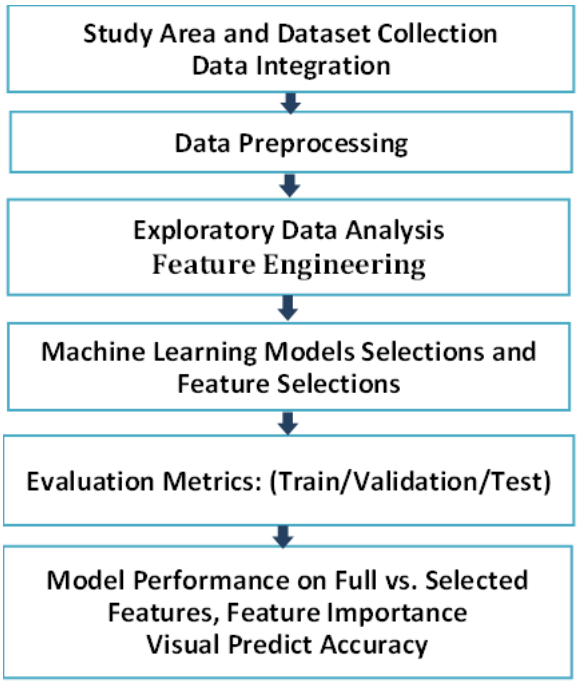

Fig. 1: Shows Methodology Steps

Study Area and Data Collection

Study Area

The research area is located at 19.8797° N latitude and 75.3559° E longitude to study Chhatrapati Sambhaji Nagar's urban area in India. The area has an average height of 560.87 meters according to a 0.5° × 0.625° grid cell resolution [13][14][15][16]. The proposed models require the solar irradiation variability and geographic and climatic characteristics of the region for their design and applicability (see Figure 2).

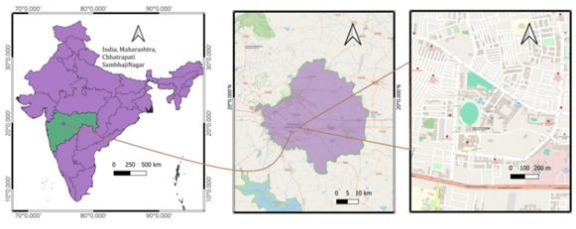

Fig. 2: Indicating Study Location (Chhatrapati Sambhaji Nagar).

Data Sources and Features

The main dataset features was obtained from the Prediction of Worldwide Energy Resource (POWER) project of NASA [17]. The Global Monitoring Laboratory provided supplementary sunrise-sunset features/ variables through the NOAA Solar Calculator [18].

The dataset includes 22.7 years of data from January 1, 2001 to July 1, 2023 with 197184 hourly entries in CSV format. The following variables cum features were extracted (see Table 1 and Table 2).

Table 1: Detail about Dataset Key Features and Sources

DataSet key Features	Source-Satellite Spatial Resolution, Elevation & Coverage Available Duration of Dataset : 2001-01-01 to 2023-07-01	API link an online platform
a. Solar Irradiance (Wh/m²) –Target variable b. Temperature (°C) c. Surface Pressure (kPa) d. Specific Humidity (g/kg) e. Relative Humidity (%) f. Wind Speed (m/s) g. Wind Direction (degrees) h. Year, Month, Day, and Hour (used to derive Unix Timestamp)	CERES: 1°×1° (~110 km × 110 km). MERRA-2: 0.5°×0.625° (~55 km × 69 km at equator). Elevation from MERRA-2: Average for 0.5 x 0.625 degree lat/lon region = 560.87 meters and Coverage Available- Public – whole World	NASA's "The POWER Project (https://power.larc.nasa.gov/data-access-viewer/) [17].
i. Sunrise Time (hh\:mm) j. Sunset Time (hh\:mm)	lat/lon regions -hours /minutes /seconds and Coverage Available- Public – whole World	NOAA Solar Calculator (https://gml.noaa.gov/grad/solcalc/) [18].

Data Preprocessing

The process of effective preprocessing became essential because it maintained data integrity while preparing the data for machine learning workflows. The original dataset (`Comma Separated Value or.csv file`) underwent cleaning and transformation into DataFrame file for process through the following steps:

1) Data Cleaning and Integrity Checks

The dataset received a scan for missing entries (NaN) and placeholders (e.g., `-999.0`) and infinite values (`inf`) and malformed entries (e.g., `#NAME?`). The dataset contained no major issues that required imputation or removal according to the preprocessing analysis stuff [9] [10] [11] [19].

2) Data Type Conversion and Timestamp Generation

The temporal features (Year, Month, Day, Hour) received separately and then combined datetime object transformation which produced Unix timestamps before being stored as pandas datetime objects (once combined then given column name i.e.`udate`). The DataFrame index received the `udate` column which received chronological sorting for time series analysis purposes [17].

The conversion of `sunrise_time` and `sunset_time` into float values representing hours after midnight occurred to make future feature engineering processes easier [18][19][20].

Table 2: Data Dictionary – Structure Description

<class 'pandas.core.frame.DataFrame'> DatetimeIndex: 197184 entries, 2001-01-01 05:30:00 to 2023-07-01 04:30:00 Data columns (total 12 columns):			
# Column	Non-Null	Count	Dtype
0 date	197184	non-null	object
1 datetime	197184	non-null	object
2 utc_unixtime	197184	non-null	float64
3 Irradiance	197184	non-null	float64
4 Temperature	197184	non-null	float64
5 RHumidity	197184	non-null	float64
6 WindSpeed	197184	non-null	float64
7 WindDirection	197184	non-null	float64
8 Pressure	197184	non-null	float64
9 SHumidity	197184	non-null	float64
10 sunrise_time	197184	non-null	datetime64[ns]
11 sunset_time	197184	non-null	datetime64[ns]
dtypes: datetime64[ns](2), float64(8), object(2) memory usage: 23.6+ MB			

Exploratory Data Analysis (EDA)

The analysis of key features included distribution assessment and interdependency analysis and temporal pattern evaluation [9][10][11][19][20].

1) Descriptive Statistics

The numerical features received summary statistics including mean, median, standard deviation, min, max and quartiles to evaluate their central tendencies and variabilities through `dataframe.describe()`.

2) Distribution Analysis

The distribution of each feature was analyzed through histograms and boxplots to detect skewness and spread and identify outliers see Figure 3.

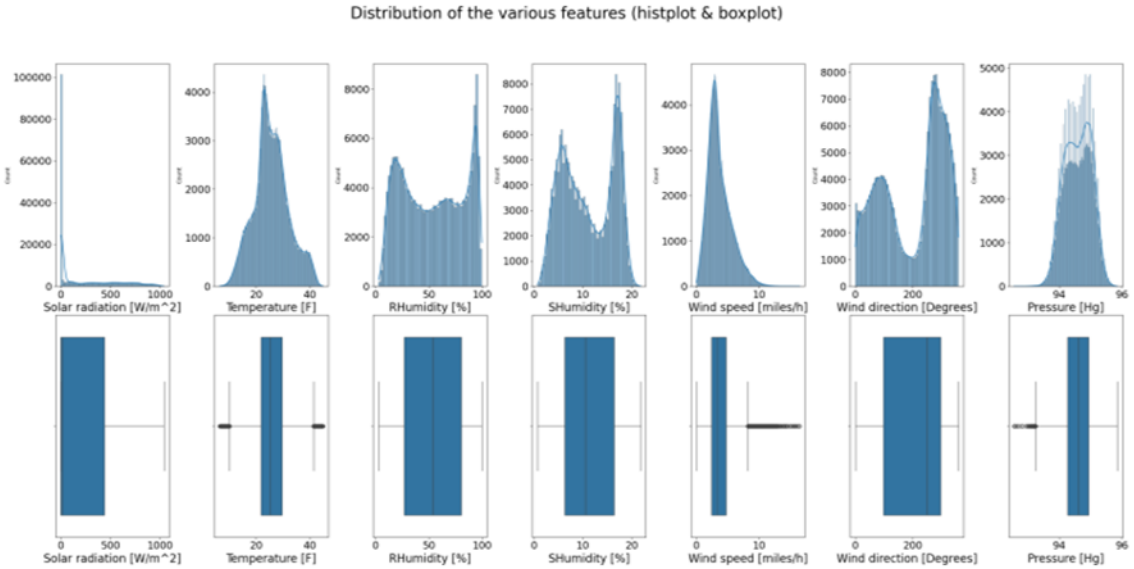

Fig. 3: Histograms & Boxplot of Key Features shows distribution of the dataset to understand the various data is allocated between the lower and upper limits.

3) Correlation Analysis

The analysis used correlation coefficients to detect multicollinearity and strong associations between solar irradiance and other variables/features [5][10][11][21] which were displayed through a heatmap see Figure 4.

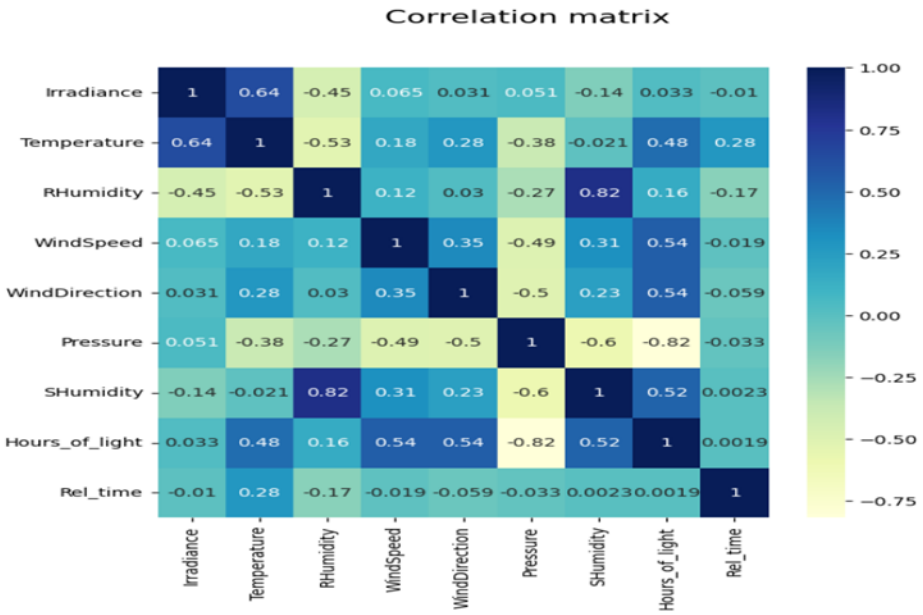

Fig. 4: Heatmap indicating correlation matrix of features.

4) Relationship Visualization

The analysis used scatter plots to study linear and non-linear connections between solar irradiance and temperature and humidity and pressure and time-dependent variables [10] [11] [22] see Figure 5.

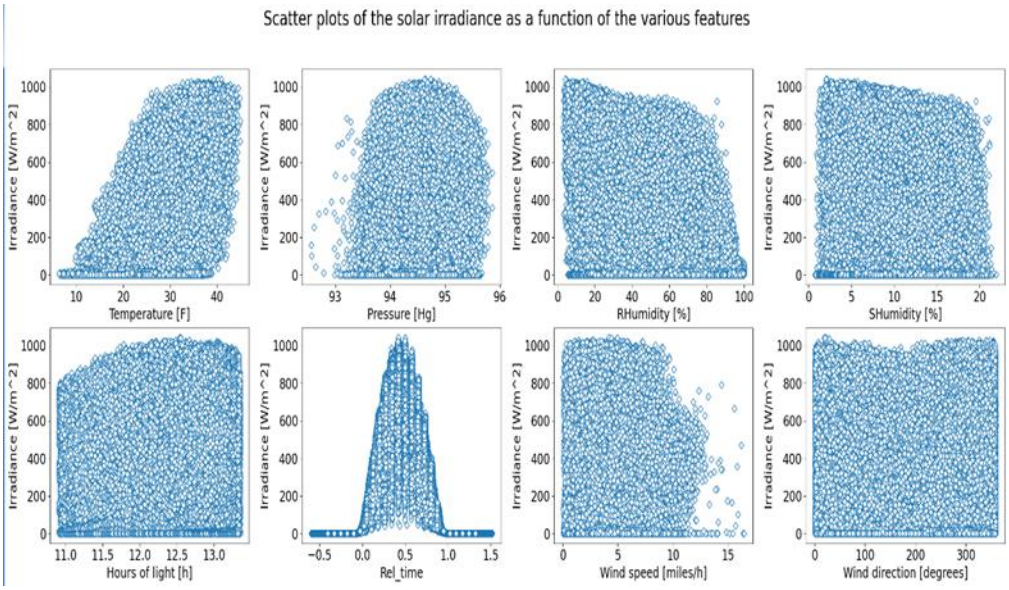

Fig. 5: Scatters plots Irradiance vs. Features allow to identify potential trends linear or non-linear.

5) Time Series Visualization

The time-based plot of solar irradiance included sunrise and sunset indicators to evaluate how daylight affects the data see Figure 6.

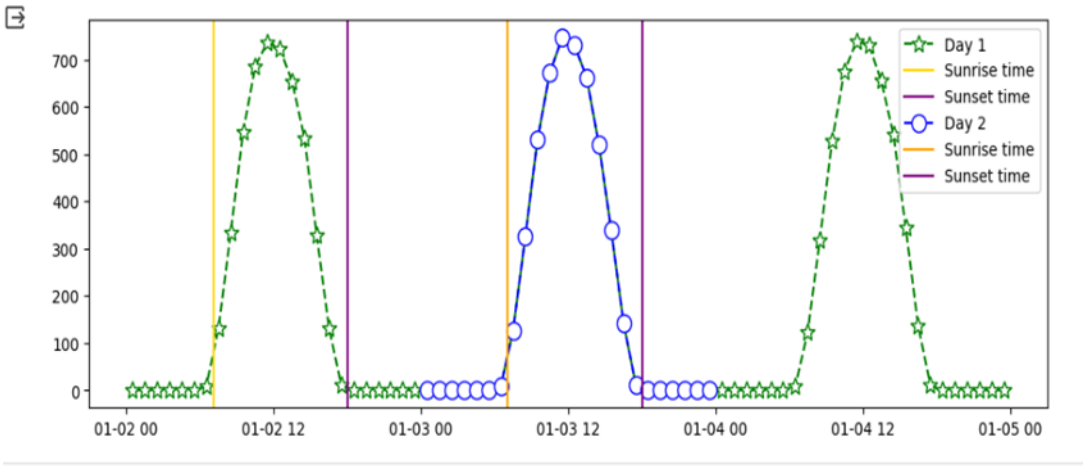

Fig. 6: Solar Irradiance Time Series with Day/Night Markers.

Feature Engineering

The model gained better expressiveness through the development of two temporal features that is 'Hours of Light' and 'Rel_time' [10]:

1) Hours of Light

The feature calculates daylight duration by subtracting sunrise from sunset times in hours because this variable affects solar irradiance variability.

2) Relative Time (Rel_time)

The calculation of relative time involved determining the normalized position of each hourly timestamp relative to the daylight window.

Rel_time = (current_timestamp − sunrise_timestamp) / (sunset_timestamp − sunrise_timestamp)

The values extend from 0 at sunrise to 1 at sunset but nighttime hours fall outside this range. It assumes as like following range:

< 0 before sunrise

= 0 at sunrise

'>' 0 but < 1 between sunrise and sunset
= 1 at sunset
'>' 1 after sunset

Machine Learning Models and Evaluation

1) Data Normalization and Splitting

The input features (X) received standardization through `StandardScaler` to achieve zero mean and unit variance. The target variable (y = solar irradiance) was left unscaled. The dataset received an 80/20 split for training and testing purposes while using `random_state=42` to ensure reproducibility.

2) Model Selection

As a multivariate regression scenario here six regression models does have the competency were employed from `scikit-learn`: [9][10][11].

a) Linear Regression: Baseline linear model.
b) Decision Tree Regressor: Captures non-linear splits; `random_state=42`.
c) Random Forest: Ensemble of decision trees with`n_estimators=100`,`max_depth=10`
d) Gradient Boosting Regressor: Sequential tree boosting; `random_state=42`.
e) Extra Trees Regressor: More randomized version of Random Forest; `random_state=42`.
f) K-Nearest Neighbors Regressor (KNN): Instance-based learner using k-nearest averaging.

3) Feature Selection / Feature Importance Selection

- Model-specific feature importance techniques were used:
- Tree-based models: `feature_importances /impurity-based feature importances
- Linear Regression: Magnitude of `model.coef_`
- KNN: `permutation-based importance for distance-based models`

The `select_top_features` as by implemented with the following logic:
a) Rank features by importance.
b) Calculate cumulative importance.
c) Select top features that collectively explain ≥70% of total importance.
d) Ensure a minimum of 7 features or ≥70% of total features, whichever is greater.

Models were evaluated using both the full and the reduced feature sets and model was assessed using evaluation metrics such as R^2, MSE, RMSE, MAE and Max Error etc. these considered as in regression/prediction point of view [23][24][25][26][27].

Here for instance as graphical visualization significances so observe findings, following plotted figures for indistinct compare features importance regarding model point of view as for evaluation of performance, see Figure 7, Figure 8, Figure 9 again observe Figure 10 to Figure 13. Hereafter performances summarizes of all models i.e. full sets verses selected the reduced set of feature in Table 3.

[Full Features] Evaluation for Linear Regression
Train Set:
R2: 0.6410, MSE: 31269.8615, RMSE: 176.8329
Test Set:
R2: 0.6416, MSE: 31143.1351, RMSE: 176.4742

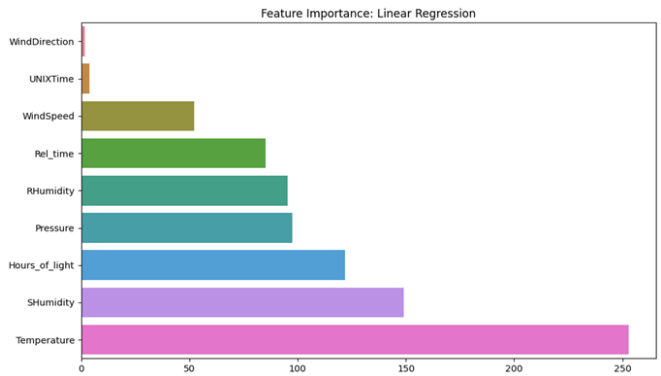

Fig. 7: Indicating features importance of Linear Regression

Selected Top ≥70% Features for Linear Regression: ['Temperature', 'SHumidity', 'Hours_of_light', 'Pressure', 'RHumidity', 'Rel_time', 'WindSpeed']
[Selected Features] Evaluation for Linear Regression
Train Set:
R2: 0.6408, MSE: 31287.3184, RMSE: 176.8822
Test Set:
R2: 0.6413, MSE: 31168.9381, RMSE: 176.5473

[Full Features] Evaluation for Random Forest
Train Set:
R2: 0.9692, MSE: 2683.2870, RMSE: 51.8005
Test Set:
R2: 0.9671, MSE: 2854.6763, RMSE: 53.4292

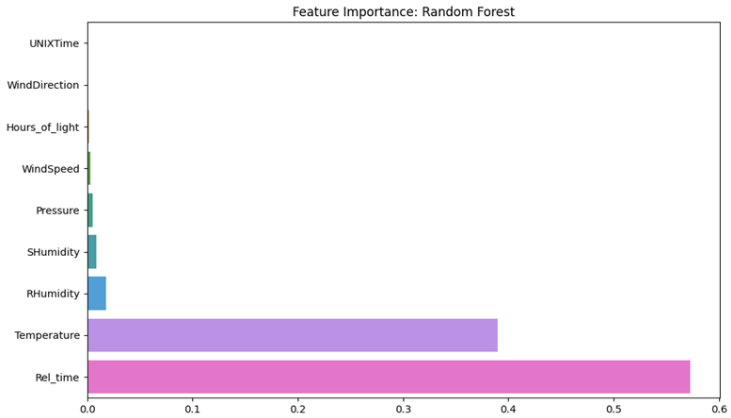

Fig. 8: Indicating features importance of Random Forest

Selected Top ≥70% Features for Random Forest:
['Rel_time', 'Temperature', 'RHumidity', 'SHumidity', 'Pressure', 'WindSpeed', 'Hours_of_light']
[Selected Features] Evaluation for Random Forest
Train Set:
R2: 0.9688, MSE: 2719.0923, RMSE: 52.1449
Test Set:
R2: 0.9669, MSE: 2879.6107, RMSE: 53.6620

[Full Features] Evaluation for Extra Trees
Train Set:
R2: 1.0000, MSE: 0.0000, RMSE: 0.0000
Test Set:
R2: 0.9757, MSE: 2115.7229, RMSE: 45.9970

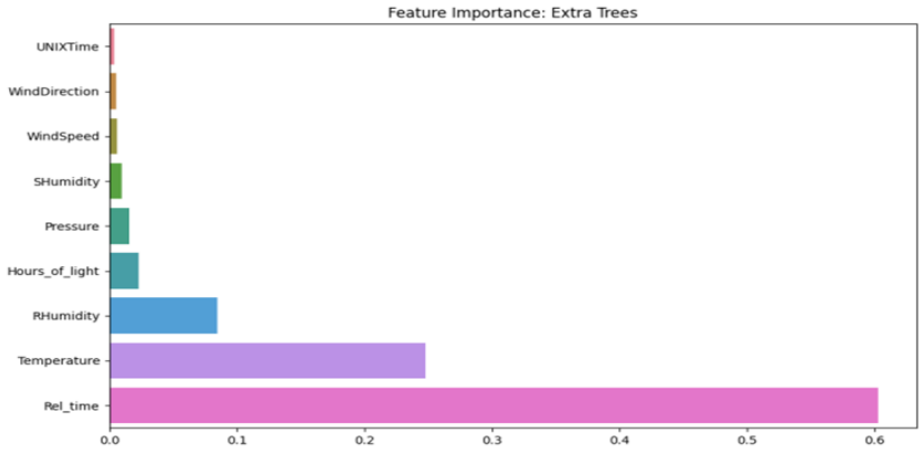

Fig. 9: Indicating features importance of Extra Trees.

Selected Top ≥70% Features for Extra Trees:
['Rel_time', 'Temperature', 'RHumidity', 'Hours_of_light', 'Pressure', 'SHumidity', 'WindSpeed']

[Selected Features] Evaluation for Extra Trees
Train Set:
R2: 1.0000, MSE: 0.0000, RMSE: 0.0000
Test Set:
R2: 0.9715, MSE: 2475.5886, RMSE: 49.7553

4) Evaluation Metrics

Each model was assessed on both train and test sets using the following metrics:
R^2 (Coefficient of Determination):

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - y^-)^2} \quad \ldots\ldots. (1)$$

Mean Squared Error (MSE) :

$$\text{MSE} = \frac{1}{\text{N}} \sum_{\text{I}=1}^{\text{N}} (y_i - \hat{y_i})^2 \quad \ldots\ldots\ldots (2)$$

Root Mean Squared Error (RMSE) :

$$\text{RMSE} = \sqrt{\text{MSE}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Mean Absolute Error (MAE) :

$$\text{MAE} = \frac{1}{\text{N}} \sum_{\text{I}=1}^{\text{N}} \left| y_i - \hat{y_i} \right| \quad \ldots\ldots\ldots\ldots (4)$$

Maximum Error (Max Absolute Error) :
Largest observed absolute prediction error:

$$\text{ME} = \text{MAX}\left(\left| y_i - \hat{y_i} \right| \right) \quad \ldots\ldots\ldots (5)$$

Where:
- yi is the actual observed value
- y^i is the predicted value from the model
- y^- is the mean of the observed values
- n is the total number of observations.

III. Results

This section presents the results of the exploratory data analysis (EDA), model performance evaluation, feature importance analysis, and visualization of predicted versus actual irradiance values for selected five-day period. The outputs correspond to the designed pipeline involving six regression models evaluated under two feature scenarios: full feature set and a selected reduced set derived from feature importance selection.

EDA Findings Summary

The EDA revealed several key patterns in the data. Solar irradiance exhibited distinct *diurnal and seasonal trends*, with peak values typically observed during midday and higher irradiance levels occurring in the summer months.

A correlation heatmap (Figure 4) showed that *Temperature and Relative Humidity* had strong linear relationships with irradiance— *positive* in the case of temperature and *negative* for humidity. Wind-related variables, such as Wind Direction, demonstrated weaker linear associations. Among the engineered features, Rel_time showed a pronounced positive correlation with irradiance during daylight hours, as it effectively captures the normalized position of the hour within the solar day.

Scatter plots (Figure 5) further confirmed these associations, particularly the parabolic trend of irradiance with Rel_time, suggesting its importance as a non-linear predictor.

Model Performance

Model performance was assessed using multiple evaluation metrics on both the training and testing subsets. Each model was trained using both the complete feature set and the top ≥70% of features selected based on importance scores. See Figure 7, Figure 8, Figure 9 and Table 3.

1) Performance on Full vs. Selected Features

Comparative analysis revealed that using selected features often maintained and significantly improving interpretability while reducing model complexity. These features align well with domain knowledge, making the model more transparent without sacrificing much performance. For example, the Extra Trees, when trained with the *selected features*, achieved an *R^2 of 97.150956* and *RMSE of 49.755287,* compared to *R^2 of 97.565109* and *RMSE of 45.996988* when trained on all full features as shown in Table 3.

This trend was consistent across most models, with slight gains or stability in RMSE and MAE values observed. Feature importance selection assisted in eliminating redundant or noisy features, contributing to better generalization and interpretability.

Table 3: Performance Summary of All Models With Full Vs. Selected Features Sets Using Evaluation Metrics

Models	r2	mse	rmse	mae	me
Linear Regression (Full Features)	64.158761	31143.13514	176.474177	141.502404	708.447444
Linear Regression (Selected Features)	64.129066	31168.93808	176.547269	141.604613	709.189437
Decision Tree (Full Features)	95.184413	4184.355427	64.686594	28.677778	718.29
Decision Tree (Selected Features)	94.868315	4459.018559	66.775883	29.468884	691.22
Random Forest (Full Features)	96.714681	2854.67631	53.429171	25.081649	575.286554
Random Forest (Selected Features)	96.685985	2879.610732	53.662005	25.23155	575.636427
Gradient Boosting (Full Features)	96.33756	3182.363907	56.412445	30.316349	587.312399
Gradient Boosting (Selected Features)	96.310607	3205.783906	56.619642	30.273141	588.235547
Extra Trees (Full Features)	97.565109	2115.722912	45.996988	20.931618	539.492
Extra Trees (Selected Features)	97.150956	2475.588565	49.755287	22.803364	586.9368
K Neighbors (Full Features)	95.003423	4341.621107	65.890979	36.294211	695.632
K Neighbors (Selected Features)	95.447335	3955.897053	62.895922	33.513757	559.272

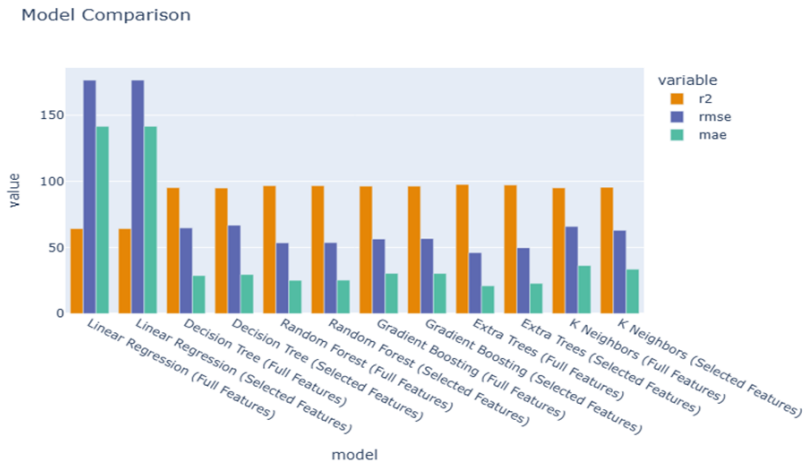

Fig. 10: Comparing R^2, RMSE, MAE across all models as using feature configurations.

Feature Importance

Feature importance analysis provided insights into the influence of each predictor variable across different model types: See Figure 7 to Figure 13, specifically Figure 14.

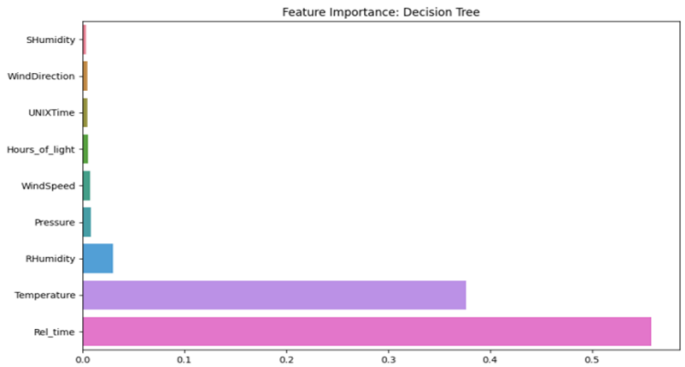

Fig. 11: Top ≥70% Features for Decision Tree: ['Rel_time', 'Temperature', 'RHumidity', 'Pressure', 'WindSpeed', 'Hours_of_light', 'UNIXTime']

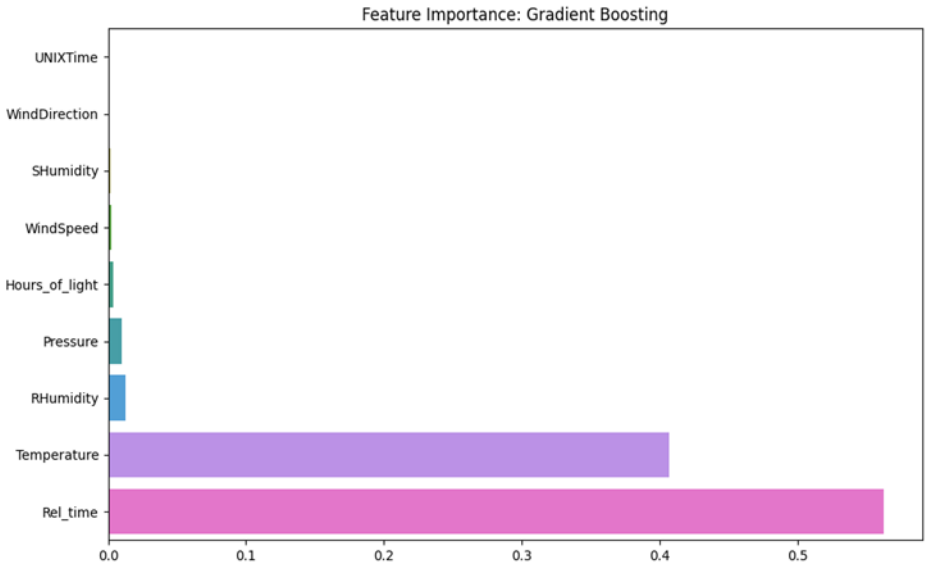

Fig. 12: Top ≥70% Features for Gradient Boosting: ['Rel_time', 'Temperature', 'RHumidity', 'Pressure', 'Hours_of_light', 'WindSpeed', 'SHumidity']

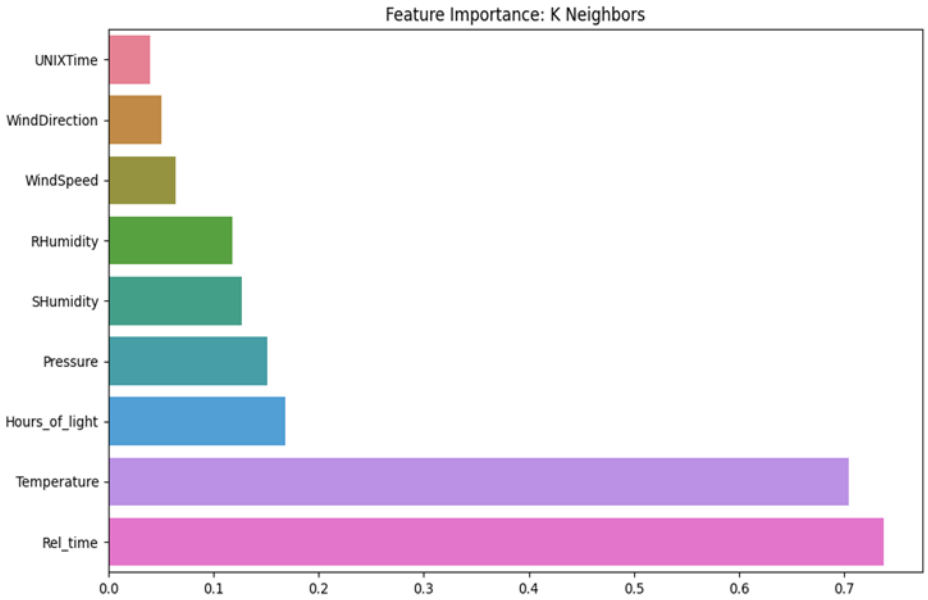

Fig. 13: Top 7 or ≥70% Features for K Neighbors: ['Rel_time', 'Temperature', 'Hours_of_light', 'Pressure', 'SHumidity', 'RHumidity', 'WindSpeed']

Visual Predict Accuracy

To assess how well models captured the dynamics of solar irradiance, predictions for a continuous period, for instance 5-day period were plotted against actual irradiance values.

The ensemble models—particularly *Random Forest, Extra Trees, and Gradient Boosting*—demonstrated high alignment with the actual irradiance curve, capturing both peak and trough patterns effectively as pronounced while using selected features.

Conversely, simpler models like Multiple *Linear Regression* struggled to capture the non-linear variability, often under- or over-estimating during midday peaks.

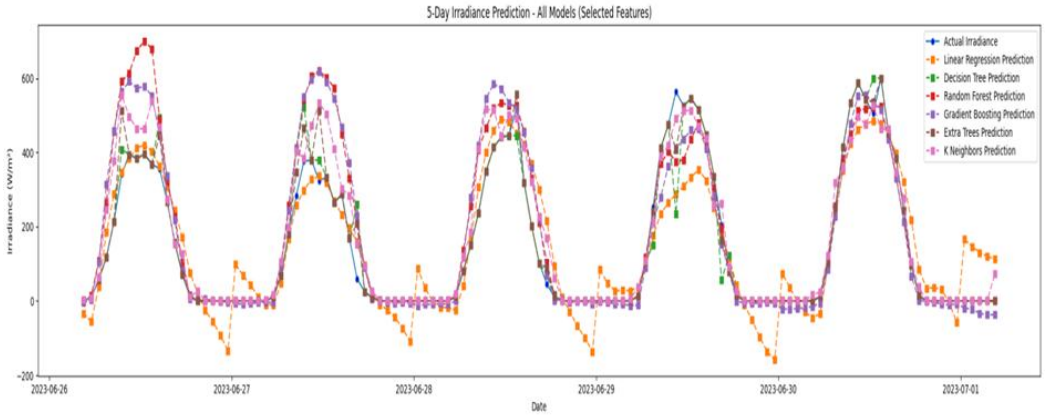

Fig. 14: Actual vs. Predicted Irradiance (over a 5-day Period) of All Models Using Top Selected Features.

IV. Outputs

In this study the Python-based approaches utilized to produces complete findings from the implementation of six machine learning regression models for predicting hourly solar irradiance in geographical area "Chhatrapati Sambhaji Nagar, India". The study has three main findings which are:

The R^2, RMSE, MAE, MSE and Maximum Error values obtained during model evaluation enable researchers to assess performance differences between models. This is based on both training and testing datasets. The evaluation metrics used to assess model performance and to compare the performance of among the models observe Figure 10 and in Table 3.

The second output is the feature importance rankings which are obtained through different importance measures such as regression coefficients for linear models, impurity-based feature importance for tree-based models and permutation-based importance for distance-based models observe table 4.

The third findings are the visualizations which includes: The correlation heatmaps and scatter plots display the EDA results in the first output. Bar plots are also used to assess model performance in the study and also the performance between the full and selected feature sets. Each model type has its feature importance plots.

Optimal feature subsets are determined through cumulative importance-based selection for model complexity reduction and accuracy improvement.

Time-series plots are used to compare actual and predicted irradiance values during a 5-day continuous test window with use of top selected importance features.

In addition, the study output includes a modular and reproducible machine learning pipeline that combines preprocessing with training and evaluation and visualization stages.

V. Discussion

The findings demonstrate that various regression models have different strengths and weaknesses in terms of irradiance prediction. The Tree-based ensemble models were found to be the best among all models in the study. The *Extra Trees and Random Forest models performed better than the simple models* such as Linear Regression and K-Nearest Neighbors based on all the evaluation metrics. These models performed better than the other models in capturing non-linear interactions between the meteorological and temporal features as evidenced by higher R^2 values and lower error scores on the test set.

Feature selection was key components of the modeling process its direct effect on the model's ability to capture the complex dynamics of solar irradiance. This study found that the selection of the top 70% most important features so selecting features strategically most of the time improved model interpretability and computational efficiency while retaining predictive accuracy. The model-specific top most feature importance across all models listed in following Table 4.

Table 4: Model-Specific Top Most Feature Importance Across All Models.

All Models	Model-specific Top Most Features Importance		
Linear Regression	'Temperature'	'SHumidity'	'Hours_of_light'
Random Forest	'Rel_time'	'Temperature'	'RHumidity'
Extra Trees	'Rel_time'	'Temperature'	'RHumidity'
Decision Tree	'Rel_time'	'Temperature'	'RHumidity'
Gradient Boosting	'Rel_time'	'Temperature'	'RHumidity'
K Neighbors	'Rel_time'	'Temperature'	'Hours_of_light'

A strategic, model-specific approach to feature selection enhances interpretability and efficiency while largely maintaining accuracy. As per table 4 For example, optimal features differed between models suppose Linear Regression ('Temperature', 'SHumidity') and Random Forest ('Rel_time', 'RHumidity'). This confirms that while no single feature set is universally optimal, the process of strategically selecting features per model consistently yields superior results over using an uncurated set of variables.

Temporal engineering is crucial to the success of the models. The 'Rel_time' feature which is the normalized time within the solar day was the most important feature in all the models and it is a good indicator of the solar position which affects solar irradiance levels. Similarly, the 'Hours_of_light' feature which is based on the sunrise and sunset data significantly improved the ability of the model to predict diurnal irradiance variations. The meteorological factors i.e. 'Temperature' which determine the most affect solar irradiance levels. The removal of less informative or noisy predictors in Decision Tree and Gradient Boosting helped to reduce overfitting in these models.

The visual comparison of the predicted and actual irradiance over a 5-day period further highlighted the strengths and weaknesses of the models. The ensemble methods were able to predict the peaks and troughs of irradiance well but the simpler models produced smoother and less responsive outputs which show that they could not capture the complex temporal dynamics. The visualization of prediction accuracy underscores the profound relevance of feature selection, illustrating its direct effect on the model's ability to capture the complex dynamics of solar irradiance. To validate this, predictions for a continuous period were plotted against actual measured values see Figure 14. While our model is built upon standard meteorological data (e.g., tempretue, irradiance), we acknowledge that real-world solar yield is also influenced by other factors—such as particulate matter, cloud cover, aerosol concentration, panel soiling, and subtle shading—which are often impractical to measure at scale. Consequently, this study demonstrates that strategic feature selection from readily available weather parameters can still yield a highly effective and deployable prediction model, providing a robust and practical framework for predicting under common data constraints.

VI. Conclusion

This research created and evaluated various machine learning models which predict solar irradiance per hour in Chhatrapati Sambhaji Nagar, India. The feature importance methods revealed the fundamental meteorological factors which affect solar irradiance levels. The research confirmed that:

The predictive results of Tree-based ensemble models including Extra Trees and Random Forest Regressor surpassed those of Linear Regression and K-Nearest Neighbors.

The temporal features 'Rel_time' and 'Hours_of_light' demonstrate high significance in irradiance prediction because they effectively represent solar time patterns.

The evaluation of feature importance through cumulative importance thresholds leads to better model efficiency without harming predictive accuracy. Targeted feature selection not only maintained predictive performance but also improved model computational efficiency, interpretability and efficiency.

A multivariate regression framework proposed for solar irradiance prediction delivers an optimal solution for the studied geographic area. The study establishes a robust framework for feature-driven irradiance modeling although findings from this research can help to explore feature importance while predicting and forecasting solar irradiation dynamics for solar power potential estimation while working in mostly environments with similar weather/climate or geographic conditions.

Author Contributions: Ashok Sangle (AS) and Prapti Deshmukh (PD) collaboratively contributed to the planning and implementation of this research. AS: Conceptualization, Methodology, Algorithm Development, Data Curation, Analysis, Visualization, Investigation, Validation, and Writing (Original Draft, Review, and Editing). PD: Conceptualization, Development of the Original Concept, Supervision, and Guidance throughout the study.

Funding Statement: The authors declare the following financial interests, which are not potential competing interests: This research was funded by the Mahatma Jyotiba Phule Research & Training Institute (MAHAJYOTI), Nagpur, through the Mahatma Jyotiba Phule Research Fellowship (MJPRF-2021). The funding agency had no role in the study design, data collection, analysis, or manuscript preparation. The authors affirm that there are no personal relationships or other affiliations that could influence the outcomes of this work.

Conflict of Interest: The authors declare no conflict of interest regarding the publication of this manuscript.

Acknowledgment: We sincerely thank the Computer Science and IT Department for their invaluable support, resources, and guidance throughout this investigation. We also gratefully acknowledge MAHAJYOTI for their

financial support and encouragement, which significantly contributed to the advancement of our study. The assistance from both institutions was contributory in achieving our research goals.

Ethical Approval: This paper does not include any experiments involving human or animal subjects conducted by the authors. The study data are publicly available and do not require ethical committee approval.

Data Availability Statement: The data supporting the findings of this study are publicly accessible from open sources. Solar energy data were obtained from NASA's Prediction of Worldwide Energy Resource (POWER) Data Access Viewer, available at https://power.larc.nasa.gov/data-access-viewer, while solar sunrise-sunset data were acquired from the National Oceanic and Atmospheric Administration (NOAA) Solar Calculator at https://gml.noaa.gov/grad/solcalc/. Both datasets are openly accessible without restrictions for research and educational purposes, in accordance with NASA's and NOAA's open data policies.

References

[1] M. J. Samma Et Al., "Illuminating The Future: A Comprehensive Review Of Ai-Based Solar Irradiance Prediction Models," Ieee Access, Vol. 12, Pp. 114394–114415, 2024, Doi: 10.1109/Access.2024.3402096
[2] A. Javed, B. Kasi, And F. Khan, "Predicting Solar Irradiance Using Machine Learning Techniques," In International Conference On Wireless Communications And Mobile Computing, Jun. 2019. Doi: 10.1109/Iwcmc.2019.8766480
[3] E. Abrahamsen, O. M. Brastein, And B. Lie, "Machine Learning In Python For Weather Forecast Based On Freely Available Weather Data," The 59th Conference On Imulation And Modelling (Sims 59), Pp. 169–176, Sep. 2018, Doi: 10.3384/Ecp18153169. Available: Https://Ep.Liu.Se/En/Conference-Article.Aspx?Series=Ecp&Issue=153&Article_No=24
[4] A. H. M. Jakaria, M. M. Hossain, And M. Rahman, "Smart Weather Forecasting Using Machine Learning: A Case Study In Tennessee," 2018.
[5] M. Holmstrom, D. Liu, And C. Vo, "Machine Learning Applied To Weather Forecasting," Course Project Report, Dec. 2016, Available: Https://Cs229.Stanford.Edu/Proj2016/Report/Holmstromliuvo-Machinelearningappliedtoweatherforecasting-Report.Pdf
[6] S.-G. Kim, J.-Y. Jung, And M. K. Sim, "A Two-Step Approach To Solar Power Generation Prediction Based On Weather Data Using Machine Learning," Sustainability, Mar. 2019, Doi: 10.3390/Su11051501
[7] R. Hossain, A. M. T. Oo, And A. B. M. S. Ali, "The Effectiveness Of Feature Selection Method In Solar Power Prediction," Aug. 2013, Doi: 10.1155/2013/952613
[8] A. Tandon, A. Awasthi, K. C. Pattnayak, A. Tandon, T. Choudhury, And K. Kotecha, "Machine Learning-Driven Solar Irradiance Prediction: Advancing Renewable Energy In Rajasthan".
[9] "Scikit-Learn: Machine Learning In Python — Scikit-Learn 1.6.1 Documentation.," Scikit-Learn 1.6.1 Documentation, Available: Https://Scikit-Learn.Org/Stable/
[10] E. Baldasso, "Prediction Of Solar Radiation Data," Kaggle, Vol. 26, Nov. 2020, Available: Https://Kaggle.Com/Code/Enricobaldasso/Prediction-Of-Solar-Radiation-Data
[11] "Learning Model Building In Scikit-Learn," Geeksforgeeks, Feb. 2017, Available: Https://Www.Geeksforgeeks.Org/Learning-Model-Building-Scikit-Learn-Python-Machine-Learning-Library/
[12] M. Komorowski, D. C. Marshall, J. D. Salciccioli, And Y. Crutain, "Exploratory Data Analysis," Secondary Analysis Of Electronic Health Records, Pp. 185–203, 2016, Doi: 10.1007/978-3-319-43742-2_15. Available: Https://Doi.Org/10.1007/978-3-319-43742-2_15
[13] G. Global Modeling And Assimilation Office, "Merra-2," Global Modeling And Assimilation Office, Available: Https://Gmao.Gsfc.Nasa.Gov/Reanalysis/Merra-2/
[14] K. Saha, "Smart Solutions For A Smart City: A Gis Approach".
[15] Y. Choi, J. Suh, And S.-M. Kim, "Gis-Based Solar Radiation Mapping, Site Evaluation, And Potential Assessment: A Review," Applied Sciences, Vol. 9, No. 9, P. 1960, May 2019, Doi: 10.3390/App9091960. Available: Https://Www.Mdpi.Com/2076-3417/9/9/1960
[16] K. Saha, "A Remote Sensing Approach To Smart City Development In India: Case Of Bhopal City, Madhya Pradesh".
[17] "Power Data Access Viewer," Nasa Prediction Of Worldwide Energy Resources (Power), Jun. 2022, Available: Https://Power.Larc.Nasa.Gov/Data-Access-Viewer/
[18] "Solar Position And Intensity Calculator," Noaa Global Radiation And Aerosols Division (Grad), Jun. 2022, Available: Https://Gml.Noaa.Gov/Grad/Solcalc/
[19] N. Rafsan, "Solar Radiation," Kaggle, Vol. 2, Jan. 2022, Available: Https://Kaggle.Com/Code/Rafu01/Solar-Radiation
[20] L. E. Ordoñez Palacios, V. A. Bucheli Guerrero, And E. F. Caicedo Bravo, "Assessment Of Solar Irradiation Data Sources And Prediction Models For Rural Villages In The Colombian Amazon Region," Ieee Latin America Transactions, Vol. 22, No. 12, Pp. 1019–1025, Dec. 2024, Doi: 10.1109/Tla.2024.10789635
[21] T. Hai, "Global Solar Radiation Estimation And Climatic Variability Analysis Using Extreme Learning Machine Based Predictive Model," Ieee Access, Vol. 8, Pp. 12026–12042, 2020, Doi: 10.1109/Access.2020.2965303
[22] A. Javed, B. K. Kasi, And F. A. Khan, "Predicting Solar Irradiance Using Machine Learning Techniques," 2019 15th International Wireless Communications & Mobile Computing Conference (Iwcmc), Pp. 1458–1462, 2019, Doi: 10.1109/Iwcmc.2019.8766480
[23] D. Chicco, M. J. Warrens, And G. Jurman, "The Coefficient Of Determination R-Squared Is More Informative Than Smape, Mae, Mape, Mse And Rmse In Regression Analysis Evaluation.," Peerj, Jul. 2021, Doi: 10.7717/Peerj-Cs.623
[24] V. Plevris, G. Solorzano, N. Bakas, And M. E. A. B. Seghier, "Investigation Of Performance Metrics In Regression Analysis And Machine Learning-Based Prediction Models," 8th European Congress On Computational Methods In Applied Sciences And Engineering, Jan. 2022, Doi: 10.23967/Eccomas.2022.155
[25] M. Westphal And W. Brannath, "Evaluation Of Multiple Prediction Models: A Novel View On Model Selection And Performance Assessment.," Statistical Methods In Medical Research, Jun. 2020, Doi: 10.1177/0962280219854487
[26] J. A. Troncoso, Á. T. Quijije, B. Oviedo, And C. Zambrano-Vega, "Solar Radiation Prediction In The Uteq Based On Machine Learning Models," Arxiv, Dec. 2023, Doi: 10.48550/Arxiv.2312.17659. Available: Http://Arxiv.Org/Abs/2312.17659
[27] Y. Tian, G. Nearing, C. D. Peters-Lidard, K. W. Harrison, And L. Tang, "Performance Metrics, Error Modeling, And Uncertainty Quantification," Monthly Weather Review, Feb. 2016, Doi: 10.1175/Mwr-D-15-0087.1