European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

Solar Irradiance Prediction in the Amazon Basin Using Machine Learning: A Sustainable Approach for Renewable Energy Expansion

Shivani Vats [1*], Deepika[2]

[1*,2]Assistant Professor, Jagan Institute of Management Studies (JIMS), Sector- 5, Rohini
Email: drshivanivats@gmail.com, deepikagahlan@gmail.com, (https://orcid.org/0009-0002-7174-4165)
(https://orcid.org/0000-0003-2560-5690)

***Corresponding Author:** Shivani Vats
*Assistant Professor, Jagan Institute of Management Studies (JIMS), Sector- 5, Rohini
Email: drshivanivats@gmail.com, (https://orcid.org/0009-0002-7174-4165)

Abstract
The urgent need for renewable energy sources has spurred global innovation in environmental protection and climate change mitigation. Among the viable options, solar energy stands out despite its intermittent nature. Brazil's predominantly green energy matrix is witnessing substantial solar energy expansion. Harnessing solar power in the Amazon basin offers a pathway to enhance living standards for local communities and cities without resorting to new hydroelectric plants or biomass burning, thereby avoiding significant environmental impacts. This study employs data science and machine learning tools to forecast solar irradiance (W/m²) in four cities within the Amazonas state, utilizing NASA satellite data from 2013 to 2022. We implemented decision-tree-based models and vector autoregressive (time-series) models with daily, weekly, and monthly aggregations. The prediction model achieved a mean absolute error of approximately 0.20 using adaptive boosting and light gradient boosting algorithms, aligning with the accuracy of similar studies. This research highlights the potential of satellite data for solar energy assessment in remote regions, offering a robust framework for sustainable energy planning in the Amazon basin.

Keywords: data science, solar power, Amazon Basin, NASA, Machine Learning, boosting.

1. Introduction

Recent decades have highlighted the necessity for new technologies and innovations to manage the consequences of human activity on the environment [1]. The increasing need to find alternatives to fossil fuels has significantly bolstered the role of renewable energy sources, such as wind, biomass, and solar power, in reducing the generation of GreenHouse Gases (GHG) like Carbon Dioxide (CO2) and Carbon Monoxide (CO), while also meeting the growing demand for electric energy. As a result, the share of renewable options in the energy matrix of many countries has increased. By 2030, Brazil's electric power matrix is projected to be 87% renewable, with hydroelectric power historically dominating due to the country's numerous rivers and lakes. Projections indicate that wind, solar, and biomass sources will expand by 16 GW, 4 GW, and 1 GW, respectively. Conversely, hydroelectric sources may only contribute up to 6 GW in Brazil [2]. This paper focuses on the potential of solar power in the Amazon basin, employing Machine Learning Algorithms (MLAs) as a free and flexible option. With advancements in engineering materials, power storage equipment, digital control technologies, and transmission lines, improved solar systems have emerged within locally used smart electric grids [3].

2. Related work

Predicting the prevalence of solar power is crucial for larger and floating systems [4–6]. Solar panels are particularly suitable for remote areas with limited energy supply alternatives compared to thermal electricity units burning fossil fuels, due to their greener logistic chain. They can be easily installed on mobile units, such as riverboats or small houses, enhancing local living standards. In the Amazon basin, the world's largest rainforest, solar panels offer low-impact energy generation that minimizes environmental harm. The region is a prime candidate for solar power adoption, with benefits including reduced environmental impact, improved living standards, sustainable energy production, and reduced need for complex supply chains. The limited number of local monitoring stations and challenges in data acquisition suggest that satellite technology could provide essential data. Even in less harsh environments, ground stations often face difficulties,

making satellite data crucial [7]. However, satellite data require interpretation and transformation, necessitating additional tasks to generate inputs for economic assessment [8].

In solar energy forecasting, an overview of commonly used MLAs includes support vector machines, Artificial Neural Networks (ANNs), ensemble learning machines, and Decision Trees (DTs) [9, 10]. These algorithms analyze periodic variation and noise in solar energy data due to meteorological and local environmental factors, with time aggregation spanning intervals from hourly to yearly. Studies on MLA performance in various countries, including the USA, India, China, Turkey, Morocco, Algeria, Spain, Australia, and Brazil, utilize different statistical models [11, 12]. Diverse ground conditions and weather characteristics require specific considerations for assessing algorithm performance, including statistical errors. Solar and wind sources are often combined to address intermittence, with meteorological inputs like solar irradiance, wind speed, and direction playing critical roles in forecasting. Michiorri et al. [13] and Alkhayat and Mehmood [14] discuss digital methods, including deep learning, reviewing 125 technical papers and datasets since 2017. Most studies focus on China, Australia, and the USA, with contributions from Spain, the UK, Canada, Brazil, India, Germany, France, and others. These references highlight the interconnectedness of solar and wind energy generation and the need for combined methods to improve forecast performance, given input data variability. Local cloud cover significantly affects solar energy generation, with statistical models incorporating cloud density estimation as a key parameter [15]. Diane et al. [15] address solar irradiance forecasting using statistical methods and cloud imagery, focusing on small-scale urban grids, aligning with our research. Time series (TS) and ANNs are also mentioned, considering various input data scenarios across different terrain types. This research explores these approaches to estimate solar energy potential in four Amazon basin cities, contributing to sustainable energy planning and environmental preservation in this significant region.

3. Materials and Methodologies

3.1 Machine Learning Methods and Metrics

The Data Science (DS) workflow depicted in Fig. 1 serves as a comprehensive guide for predictive modeling in solar energy forecasting [16-21]. This iterative process begins with a thorough examination of the dataset and its metadata, laying the foundation for subsequent data preprocessing steps. The initial phase involves meticulous scrutiny of the data to identify and correct missing values, erroneous data types, and outliers that could skew analytical results. Depending on the nature of these anomalies, appropriate remedial actions are taken, ranging from data imputation techniques such as mean or standard deviation filling to outright exclusion of aberrant data points. This meticulous data cleansing process is crucial for ensuring the integrity and reliability of subsequent analyses and forms the cornerstone of Exploratory Data Analysis (EDA) efforts.

The EDA phase utilizes graphical and visualization tools to uncover underlying trends, patterns, and relationships within the dataset. Through descriptive analyses and visualizations, researchers gain valuable insights into general trends and interdependencies among variables, laying the groundwork for feature selection and engineering. Additionally, temporal aggregation of the data at various granularities—ranging from daily to weekly and monthly intervals—provides a nuanced understanding of temporal dynamics and seasonal variations, essential for subsequent modeling efforts.

Before applying predictive algorithms such as Decision Trees (DT) and Time Series (TS) models, a rigorous feature importance evaluation is conducted to identify the most significant predictors that explain the relationship between the target variable (solar energy prevalence) and other covariates. This strategic feature selection process optimizes computational efficiency and enhances the interpretability and predictive performance of the models.

The iterative application of DT and TS algorithms involves fine-tuning model parameters and scoring mechanisms to minimize error and maximize predictive accuracy. This iterative refinement process continues until an optimal convergence point is reached, as depicted by the dashed line in Fig. 1, indicating the achievement of an ideal predictive model configuration.

Importantly, the DS workflow incorporates a diverse array of data sources, including satellite imagery, ground station measurements, and original meteorological variables such as rainfall, wind speed, air moisture, temperature, and seasonal indicators. Additionally, it considers specialized studies that explore correlations between solar radiation and the digital control mechanisms of smart solar panels, energy storage dynamics, and climate change-induced fluctuations.

In essence, the overarching objective of the DS process is to enhance the predictive capabilities of machine learning models, thereby strengthening the viability and efficacy of investments and operations in solar power generation amid evolving climatic conditions. By meticulously integrating domain knowledge, data-driven insights, and advanced

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

analytical techniques, the DS workflow serves as a powerful tool for navigating the complex landscape of solar energy forecasting with precision and confidence.

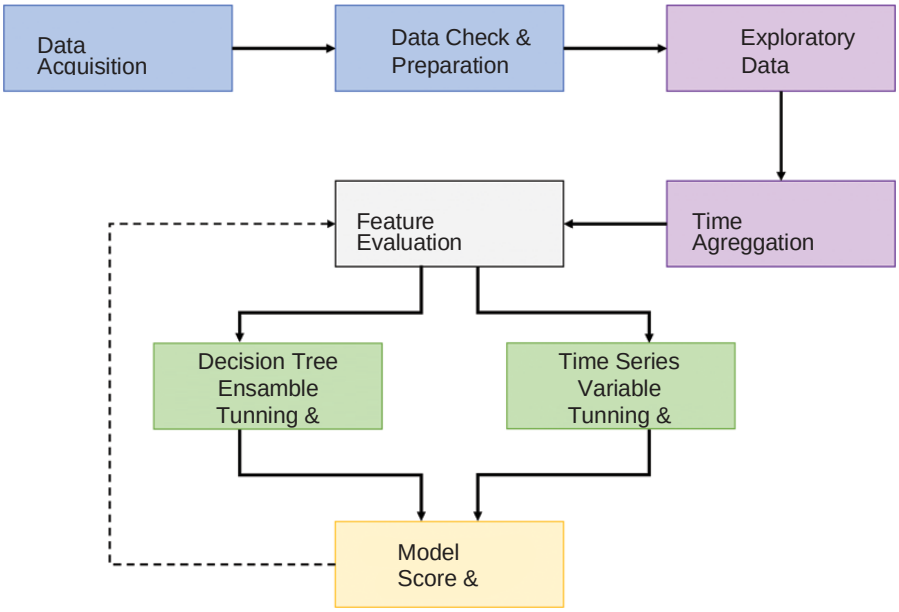

Fig. 1: Process flow for DT and TS models

3.1.1 Decision Trees (DTs)

DTs are powerful tools for elucidating cause-and-effect relationships within complex phenomena, offering a concise and intuitive representation of intricate processes. These supervised learning algorithms operate by posing a series of questions to reach a conclusive decision, effectively navigating through numerous possibilities to arrive at a definitive outcome. Notably, DTs closely mimic human logic, with decision nodes symbolizing questions and branch nodes representing corresponding answers derived from binary choices.

The construction of a DT relies on the judicious selection of conditions and attributes that govern the tree's branching structure, thereby dictating the decision-making process. A key feature of DTs is their iterative nature, allowing for continuous refinement and expansion of decision pathways by analyzing multiple variables simultaneously. However, DTs have a notable drawback: the tendency to overfit. Overfitting occurs when the model becomes excessively tailored to the training data, reducing its ability to generalize to unseen data instances. This issue is typically identified when the performance metrics during training vastly outperform those during testing, necessitating remedial measures to mitigate overfitting.

Ensemble techniques, such as Bootstrap Aggregation (bagging) and Boosting, provide viable solutions to address the challenges posed by overfitting in DTs. Bagging works by aggregating the predictions of multiple weak learners, each trained on a randomly sampled subset of the data, leveraging the collective wisdom of diverse models to achieve a more robust and generalized prediction. Conversely, Boosting focuses on iteratively improving the performance of weak learners by assigning weights to misclassified data points, thereby guiding subsequent model iterations towards areas of deficiency. Notable implementations of boosting algorithms include Adaptive Boosting (AdaBoost), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost), each offering unique optimizations tailored to specific computational constraints and performance objectives.

In summary, DTs and ensemble techniques are versatile and powerful tools for predictive modeling. DTs provide an intuitive framework for representing complex decision-making processes, despite being susceptible to overfitting. Ensemble techniques, such as bagging and boosting, offer effective strategies for mitigating overfitting and enhancing predictive accuracy, facilitating the robust and reliable modeling of complex phenomena across diverse domains.

3.1.2 TS Vector Autoregression

In multivariate Time Series (TS) analysis, the intricate interplay between various variables and their temporal dependencies can significantly influence the measured solar prevalence, the target variable of interest. Recognizing that

the dynamics of each variable may depend not only on its own historical values but also on the interdependencies with other variables within the system is crucial. In this study, the Vector Auto Regressive (VAR) model emerges as a robust forecasting algorithm for TS analysis, offering a flexible framework for capturing the complex temporal relationships inherent in multivariate datasets.

The VAR model, embodies the temporal dependencies between the target variable YtY_tYt and its lagged values up to order n, characterized by coefficients f1,f2,…,fnf_1, f_2, \ldots, f_nf1,f2,…,fn, and an error term.

This formulation allows for the incorporation of multiple variables and their respective lagged effects, enabling a comprehensive assessment of the temporal dynamics shaping solar prevalence over time. A crucial prerequisite in TS analysis is ensuring the stationarity of the time series data, where statistical properties such as mean and variance remain constant over time. The Augmented Dickey-Fuller (ADF) test is widely employed for assessing stationarity, providing insights into the temporal stability of the underlying data processes.

Determining the appropriate order of the VAR algorithm is paramount for optimizing model performance. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) offer systematic methodologies for selecting the optimal order of autoregressive models, striking a balance between model complexity and goodness of fit.

In multivariate TS analysis, the Granger causality test is a valuable tool for assessing causal relationships among explanatory variables and their impact on predicting response variables. Additionally, the Johansen cointegration test plays a pivotal role in verifying the existence of significant relationships between multiple time series, elucidating the underlying structural dynamics of the system.

Moreover, the Durbin–Watson statistic provides insights into the presence of autocorrelation in the residual errors of regression analysis, offering valuable diagnostic information for model validation and refinement.

Table 1 provides a comprehensive summary of the diagnostic tests employed in this study to assess stationarity, determine the order of autoregressive models, and elucidate correlations and causal relationships among variables.

These diagnostic tools collectively contribute to the robustness and reliability of the TS analysis framework, facilitating informed decision-making and hypothesis testing within the research domain. Geographic coordinates of the four selected cities are presented in Table 2.

Table 1: Summary of the mathematical tests associated with TS

Test	Purpose	Null Hypothesis	Statistical measure
Augmented Dickey–Fuller (ADF)	To assess stationarity in a time series	The time series has a unit root (i.e. is non-stationary)	Test statistic (e.g. *t*-test, F-test)
Akaike information criterion (AIC) and Bayesian information criterion (BIC)	To select the order of a vector autoregressive (VAR) model	Lower AIC or BIC scores indicate a better fit	N/A
Granger causality	To examine the causality between variables in a multivariate time series	There is no causal relationship between the variables	F-statistic
Johansen cointegration	To verify the existence of a significant relationship between two or more time series and examine the number of independent linear combinations of non-stationary time series that yield a stationary process	Variables are not cointegrated, meaning regression can be performed on multiple variables without falsely assuming that they are correlated	Trace statistic and the maximum eigenvalue statistic
Durbin–Watson	To assess the autocorrelation in the residual errors of regression analysis and determine how past values can influence predicted ones	No first-order autocorrelation in the residuals	A number between 0 and 4. A value of 2 indicates no autocorrelation. As it gets closer to 0 or 4, this indicates positive and negative autocorrelation, respective

Table 2: Geographic coordinates of the four selected cities

City	Latitude (degrees)	Longitude (degrees)	Population (hab)
Manaus	–3.101	–60.025	2341.000
São Gabriel da Cachoeira	+0.130	–67.089	47 031
Tabatinga	–4.253	–69.935	42 400
Humaitá	–7.500	–63.03	43 500

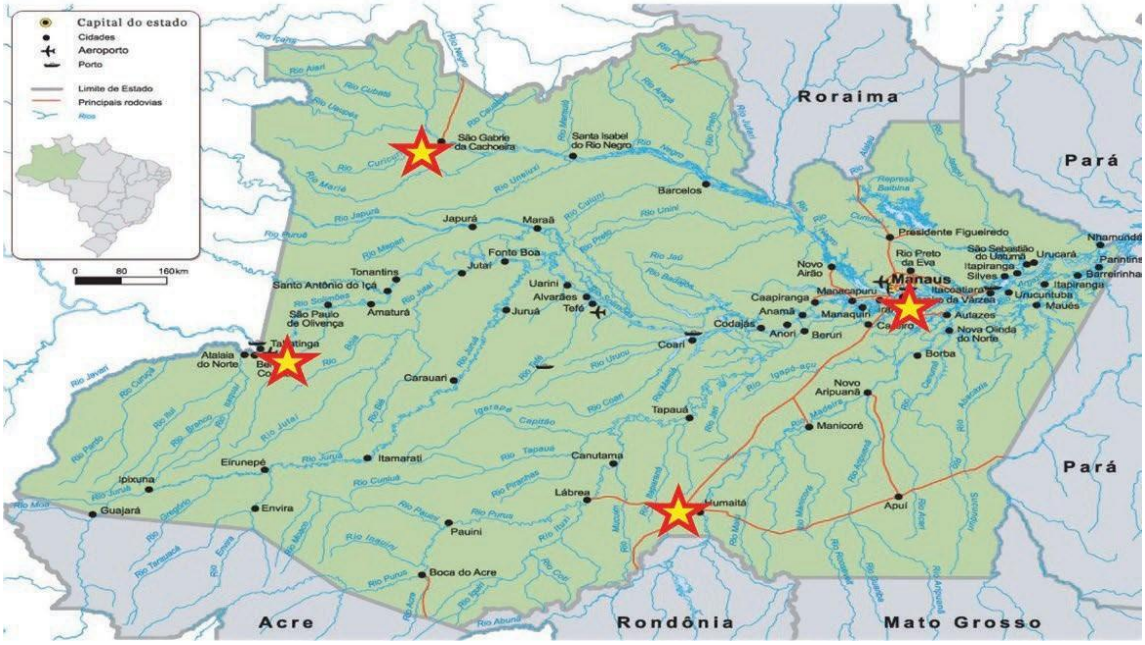

Fig. 2: Location of the four cities in the Amazonas state [22-24]

Table 3: Technical variables considered for the solar incidence forecast

Variable	*Unit*	*Description*
ALLSKY_SFC_ SW_DWN	*kW.h/ m² (day)*	*Total solar irradiance incident (direct plus diffuse) on a horizontal plane*
ALLSKY_KT	*Dimensionless*	*A fraction representing clearness of the atmosphere and all-sky insolation*
PRECTOTCORR	mm/day	Precipitation-corrected—the bias-corrected average of total precipitation at the surface of Earth in water mass (including water content in snow)
WS10M	m/s	The average wind speed at 10 m above the surface of the Earth
WS10M_Max	m/s	The maximum hourly wind speed at 10 m above the surface of the Earth
WS10M_Min	m/s	Minimum hourly wind speed at 10 m above the surface of the Earth
WD10M	Degrees	The average wind direction at 10 m above the surface of the Earth
PS	kPa	The average surface pressure at the surface of the Earth
RH2M	Dimensionless	The ratio of the actual partial pressure of water vapor to the partial *pressure at saturation expressed as a percentage (%)*
T2M_Min	*Celsius*	*The minimum hourly air (dry bulb) temperature at 2 m above the ground surface in the period of interest*
T2M_Max	*Celsius*	*The maximum hourly air (dry bulb) temperature at 2 m above the surface of Earth in the period of interest*
T2M	*Celsius*	*The average air (dry bulb) temperature at 2 m above the surface of the Earth*

3.1.3 Model Performance Metrics

The metric chosen to evaluate the predictions was the mean absolute error (MAE), defined by Equation (1). Its dimension is the same as the variables in this study, kW.h/ m2(day), from the target variable ALLSKY_SFC_SW_DWN, defined in

Table 3.

Mean Absolute Error (MAE) = (1/ n) ∗ Σ 1ypred-yactual1 (1)

where ypred represents the value predicted by an algorithm, yactual represents the real value taken from the data source and n represents the number of compliances.

Another performance index used was the mean value of a data group, which represents the simplest way to predict the value of that data group. The farther the rate (ypred/ymean) stays lower than 1.0, the predictor model shows to be more descriptive. The mean rate (dimensionless) is defined by Equation (2) as follows:

Mean rate = (ypred/ymean) (2)

3.2 Data Analysis and Processing

3.2.1 Data Acquisition

This study focuses on the Amazon basin, a vast region in northern Brazil. Particularly, the state of Amazonas, the largest in Brazil, faces increasing deforestation due to expanding agriculture and livestock activities, especially from the south to the north. The state's vibrant nucleus is Manaus, a megacity known for its electronics, home appliances, and motorcycle manufacturing sectors.

In this research, four medium and large cities—Manaus, Tabatinga, Humaitá, and São Gabriel da Cachoeira—were strategically selected as reference points. These cities span the state's four cardinal directions and are near major waterways and rainforest areas. However, their landscapes have changed significantly due to deforestation. Table 2 provides demographic details, while Fig. 2 shows their geographical locations.

Geographical coordinates following the ISO 6709 standard were used for precise spatial delineation, with geospatial mapping techniques connecting these cities. Initially, data were obtained from the Brazilian National Institute of Meteorology (INMET) website, but availability issues, especially beyond Manaus, led to the use of satellite-derived datasets from the NASA POWER Project [25-27]. These datasets, utilizing models like CERES and MERRA2, were validated with ground station data. The study period spans from January 2013 to November 2022, aligning with other research initiatives in the region.

The dataset, in comma-separated values format, has diurnal temporal resolution and includes 12 variables detailed in Table 3. Each city's dataset underwent thorough scrutiny, preprocessing, and refinement for Exploratory Data Analysis (EDA), covering daily (3,621), diurnal (518), and annual (119) profiles. Table 4 summarizes the statistical distribution of the target variable ALLSKY_SFC_SW_DWN across the four cities.

Technological infrastructure using Amazon Web Services (AWS) SageMaker notebooks and Python libraries facilitated computational analyses, allocating 4 GB of memory and two vCPUs (ml.t3.medium instance type) over 20 hours. Significant efforts addressed missing data, ensuring data integrity. Custom functions within Jupyter notebooks imputed missing values with parameter-specific standard values. Notably, the variable ALLSKY_KT had the highest incidence of missing values, with 243 instances documented in the diurnal aggregation scheme.

This methodological framework ensures rigorous and systematic data acquisition, preprocessing, and analysis, forming a robust foundation for scientific inquiry into the Amazonian region's environmental and climatological dynamics.

3.2.2 Exploratory Data Analysis (EDA)

EDA was a crucial step in understanding the datasets from the four selected cities. This phase aimed to uncover overarching patterns, trends, and characteristics within the data, facilitating informed decision-making for model development and analysis.

EDA involved a meticulous examination of the datasets to understand their magnitude and variability. This provided essential insights into the data's distributional properties, temporal dynamics, and potential anomalies, setting the stage for further analysis.

Fig. 3 illustrates the variation in the target variable ALLSKY_SFC_SW_DWN over time for Manaus under a daily aggregation scheme. This visual representation offers a nuanced understanding of the temporal evolution of solar surface downward shortwave radiation, a key climatological variable in Manaus. By plotting this variable over time, discernible patterns, fluctuations, and trends emerge, providing insights into diurnal and seasonal variations and potential long-term trends. Analyzing Fig. 3 helps identify recurring patterns and anomalies and contextualizes the data within a broader temporal framework. It also helps researchers identify potential relationships and dependencies between the target variable and other factors, enriching the analytical framework and guiding subsequent modeling efforts.

Overall, EDA represents a pivotal phase in the research, serving as the cornerstone for informed decision-making and hypothesis generation. By systematically exploring the datasets' intricacies, researchers gain deeper insights into the underlying phenomena, paving the way for robust scientific inquiry and evidence-based decision-making.

Fig, 4 offers a cohesive visual representation of the Time Series (TS) data, illustrating the distribution of values associated with the target variable. This visual encapsulation provides a comprehensive overview of its frequency distribution and inherent variability, showing characteristics reminiscent of a Gaussian distribution, essential for subsequent modeling efforts, particularly with Decision Trees (DT).

By detailing the frequency distribution of the target variable, Fig. 4 provides valuable insights into its statistical properties and structural tendencies, forming a foundation for DT methodologies. Fig. 5 presents a correlation matrix, highlighting the relationships between the target variable and other covariates. Notably, variables such as ALLSKY_KT and PRECTOTCORR are significant correlates, with correlation coefficients of 0.887 and 0.645, respectively. These findings underscore the importance of atmospheric clarity and precipitation dynamics in shaping the target variable's variability, informing feature selection in DT and TS modeling.

Fig. 6 provides a temporal perspective on the target variable's variability compared to ALLSKY_KT fluctuations. This dual-axis representation aids in understanding temporal dynamics, allowing researchers to discern potential dependencies and patterns that inform model selection and parameterization.

The visualizations in Fig. 4, 5, and 6 collectively serve as a foundation for data-driven inquiry and hypothesis generation. They help researchers understand statistical regularities, interrelationships, and temporal dynamics within the dataset, informing decisions on model selection, feature engineering, and parameterization, enhancing subsequent analytical endeavors.

The relationship between the ALLSKY_KT variable and the target variable, ALLSKY_SFC_SW_DWN, reveals intriguing insights into the temporal dynamics of solar irradiance. The ALLSKY_KT variable shows a temporal profile similar to the target variable but with a discernible delay, highlighting the interplay between atmospheric clarity and solar surface downward shortwave radiation.

A notable observation is the constant value of ALLSKY_KT in recent weeks, represented by a vertical line. This is due to imputing the median value when discrete data points from NASA were unavailable. Despite this, the correlation between ALLSKY_KT and ALLSKY_SFC_SW_DWN underscores atmospheric clarity's role in solar irradiance dynamics, informing predictive modeling efforts.

The strong correlation between ALLSKY_KT and ALLSKY_SFC_SW_DWN is due to sophisticated modeling techniques using satellite imagery to estimate atmospheric constituents and solar irradiance levels. These models enhance the fidelity and predictive power of the analytical framework.

Graphical representations of ALLSKY_KT's temporal evolution help identify patterns and trends, particularly in high-order variations. This aids in feature engineering and model parameterization, enhancing predictive performance.

Using regularized values allows researchers to quantify mean and standard deviation variations across multiple temporal dimensions, providing comprehensive insights into the dataset's variability and stability. This multifaceted data analysis approach empowers researchers to derive actionable insights, facilitating informed decision-making and hypothesis generation.

Table 4: Data summary

ALLSKY_SFC_SW_DWN	Manaus	Humaitá	SG Cachoeira	Tabatinga
Observed counts	3621	3621	3621	3621
Mean	4.69	4.86	4.61	4.57
Standard deviation	1.20	1.04	1.12	1.08
Minimum	0.35	0.87	0.84	0.74
25%	3.96	4.26	3.94	3.92
50%	4.91	5.04	4.73	4.68
75%	5.59	5.62	5.43	5.36
Maximum	7.11	7.23	7.11	7.17

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

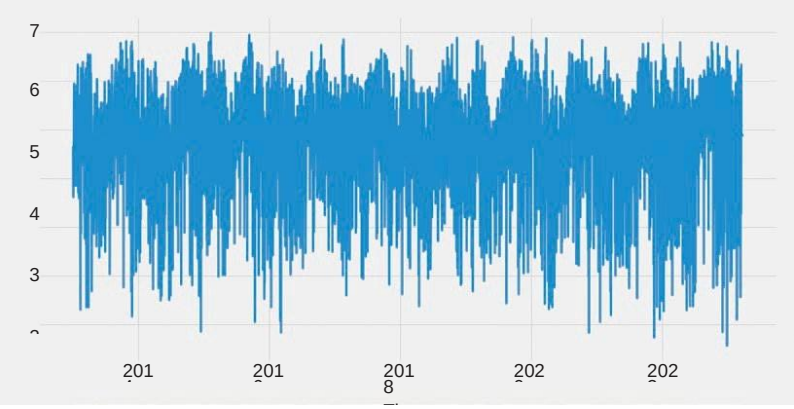

Fig. 3: Variation of ALLSKY_SFC_SW_DWN as a function of time for Manaus

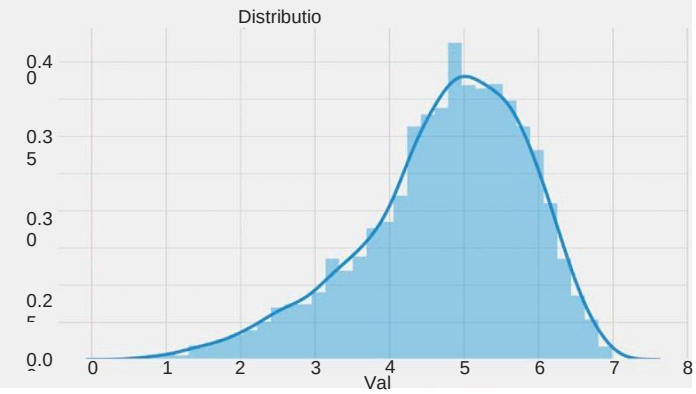

Fig. 4 The distribution of ALLSKY_SFC_SW_DWN for Manaus with daily aggregation

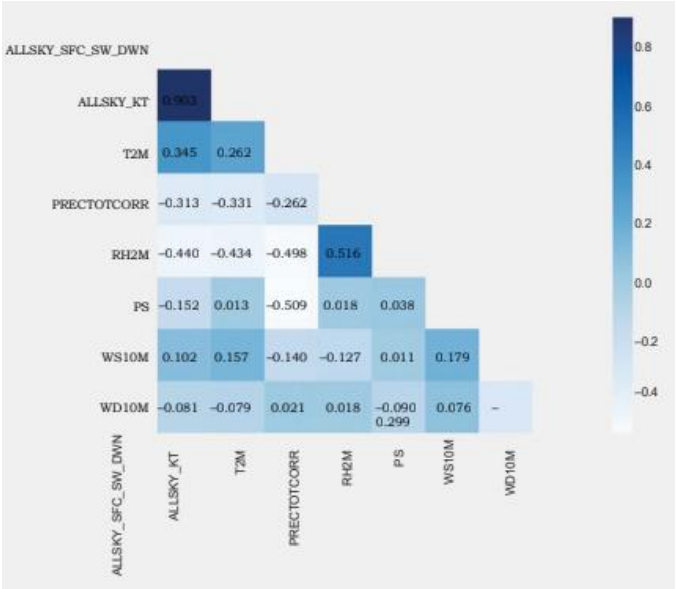

Fig. 5: The correlation matrix of the variables related to the city of Manaus with daily aggregation

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

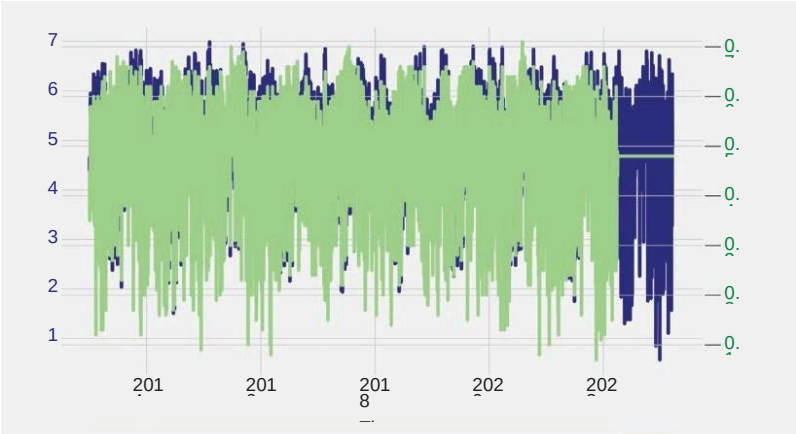

Fig. 6: Combined presentation of ALLSKY_SFC_SW_DWN and ALLSKY_KT

3.3 Feature Engineering

In the realm of feature engineering, the initial focus was directed towards the analysis of 12 input variables outlined in Table 1, serving as the foundational elements for constructing predictive models. This initial feature set can be metaphorically likened to the 'birth' of the predictive models. Subsequently, a comprehensive assessment of feature significance was conducted utilizing a redundant tree regressor algorithm, elucidating the relative importance of each feature for prediction. Fig. 7 provides a visual representation of the feature significance analysis, showcasing the six most impactful features for Manaus.

Among the identified features, WS10M, T2M, and ALLSKY_KT emerged as the top contributors to the target variable, each exhibiting a substantial impact, with a magnitude of approximately 20. Other variables demonstrated comparatively lower impacts on prediction. To enhance the predictive performance of the algorithms beyond the initial feature set, additional features were explored. These included variations such as original mean and standard deviation, both at -1, at, and at 1, in comparison with the overall mean and standard deviation.

Furthermore, novel features were engineered to augment the predictive capabilities of the models. These included:

- **local_mean_comp**: The ratio between the original variable value at 1 and the original mean, calculated as (at 2 at) / 2.
- **mean_comp_2var**: The ratio between the features above of two variables, one of which being the ALLSKY_SFC_SW_DWN.
- **overall_mean_comp**: The ratio between the original mean and the overall mean up to the specific value at 1.
- **overall_std_comp**: The ratio between the original and overall standard deviations up to the specific value at 1.

In terms of feature selection, three distinct options were considered for each megacity and time aggregation:

- **Option A**: Utilizing the original 11 input variables as listed in Table 2.
- **Option B**: Incorporating six input variables with the highest significance alongside local_mean, presentation, mean_comp_2var, overall_mean_comp, and overall_std_comp.
- **Option C**: Introducing additional time-related features, totaling 21 input variables, including those listed in Table 2 along with local_mean_comp, mean_comp_2var, overall_mean_comp, overall_std_comp, day, month, day of the week, week of the time, quarter of the time, and semester.

Fig. 8 depicts a schematic representation of the feature engineering process along with the subsequent execution of decision trees (DT) and time series (TS) algorithms. Following feature engineering, DTs and TS were independently executed, and the model with the smallest Mean Absolute Error (MAE) was selected for each megacity[28].

In the context of the TS model, data stationarity was verified, and if necessary, series modification was undertaken to achieve stationarity. Subsequently, the training-testing split was performed, and the model order was determined based on the Akaike Information Criterion (AIC) indicator. The model was then fitted using the training dataset, followed by diagnostic tests including Durbin–Watson indicator analysis, Granger's causality test, and cointegration tests.

Evaluation metrics such as MAE and mean rate were computed to assess the predictive performance of the models, ensuring adherence to predefined acceptance criteria. Moreover, the number of predictions to be generated played a crucial role, impacting the training-testing split, with longer forecast periods necessitating adjustments in data partitioning. To mitigate overfitting, the ratio between MAE_test and MAE_train was monitored, with a threshold of <1.35 indicating acceptable model generalization.

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

In essence, the feature engineering process and subsequent model evaluation framework outlined herein provide a structured approach towards developing robust predictive models for megacity crime forecasting, thereby facilitating informed decision-making and resource allocation in urban security management.

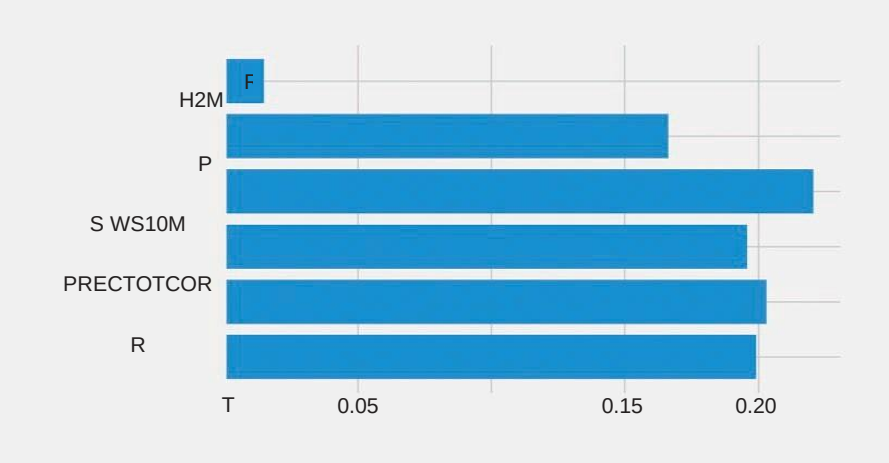

Fig. 7: Feature importance analysis for the city of Manaus

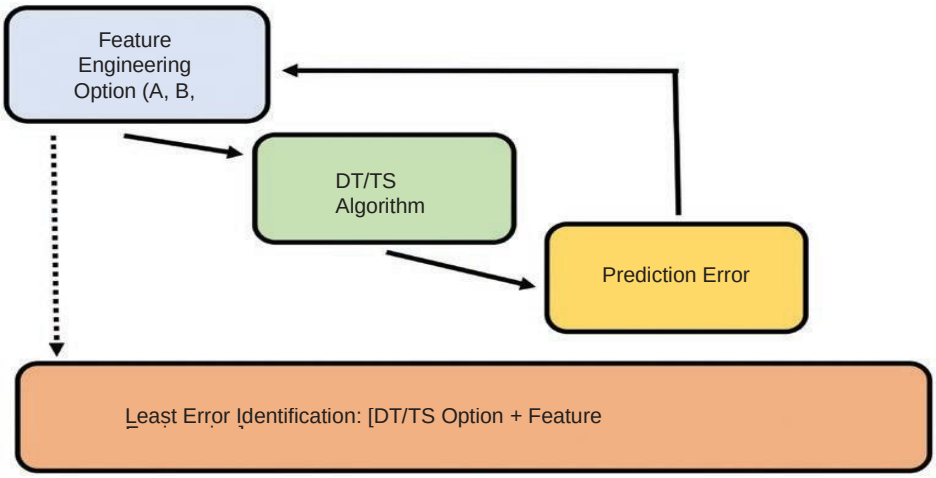

Fig. 8: Feature engineering and DT and TS task flow

4. Results and Discussion

4.1 Baseline Setup:

11 input variables across daily, diurnal, and yearly temporal aggregations. Mean MAE of 0.300 and mean rate of 0.373, indicating good performance without overfitting.

4.2 Increased Variables:

Smallest MAE improved to 0.221 with a mean rate of 0.273. Mean MAE increased by 30 (to 0.330), despite a reasonable mean rate (<0.5). Adding more features led to performance degradation, suggesting restraint in feature addition.

4.3 Decision Trees (DT) Implementations:

São Gabriel da Cachoeira: Lowest MAE: 0.191 under daily aggregation with adaptive boost.

Mean Rate: 0.769. Lowest Mean Rate: 0.585 under diurnal aggregation with adaptive boost. Mean MAE across all instances was 0.470, indicating consistent performance.

4.4 Comparative Analysis:

MAE magnitudes aligned with select references, though environmental factors caused disparities. To explore greenhouse gas influence on solar energy forecasting, considering their environmental impact.Manaus: Novel features reduced MAE and mean rate, showing potential efficacy. Third-law configuration also yielded improvements. Emphasizes the importance of feature engineering in improving predictive models.

4.5 Time Series Analysis:

VAR Algorithm: Provided valuable performance insights for further refinement.
Mean Rate Values: Mostly below 1.0, indicating effective forecasting.
Outliers: Excluding them significantly reduced mean MAE, aligning with research trends.

4.6 Overall Findings:

Highlights the importance of feature engineering and model refinement for optimizing predictive algorithms. Significant performance enhancement in Manaus underscores the efficacy of data-driven approaches for informed decision-making and resource management in dynamic environments. Table 5 summarizes overall results in terms of MAE and mean rate.

Table 5 : DT Results (wrt lowest value of MAE and mean ratio)

City	Time aggregation	MAE	Mean ratio	Algorithm (best-performing)
	Day	0.327	0.585	Adapt Boost
Manaus	Week	0.233	0.612	Light Grad Boost
	Month	0.440	0.787	Light Grad Boost
	Day	0.761	0.703	Light Grad Boost
Humaitá	Week	0.397	0.788	Light Grad Boost
	Month	0.195	1.007	Gradient Boost
	Day	1.097	0.922	Gradient Boost
SG Cachoeira	Week	0.191	0.769	Adapt Boost
	Month	0.328	1.000	Gradient Boost
	Day	0.921	0.831	Extreme GBoost
Tabatinga	Week	0.433	0.818	Adapt Boost
	Month	0.418	0.834	Light Grad Boost

5. Conclusion

Upon comparing the two refined approaches, it became apparent that both yielded Mean Absolute Errors (MAEs) of similar magnitudes, although the Decision Tree (DT) method showed a modest advantage with a 30-unit lower MAE. The most notable MAE value (0.167) was achieved through light-grade boosting combined with feature engineering from Option B. This approach also demonstrated resilience against overfitting, emphasizing its suitability for predictive modeling tasks. The corresponding mean rate stood at 0.206, indicating the DT methodology's ability to accurately predict around one-fifth of the forecasted outcomes based solely on the mean of the overall target variable, highlighting its efficacy as a reliable predictive tool. On the other hand, within the Time Series (TS) framework, the optimal model configuration involved a VAR model of order 2, resulting in a minimal MAE of 0.188 for daily aggregation and four forecast horizons. Lower-order models (order < 3) outperformed higher-order ones (order ≥ 4), emphasizing the importance of model simplicity in forecasting accuracy. Despite achieving commendable MAE values across various TS configurations, some configurations yielded 'negative' forecasted values, particularly in the cases of Humaitá, Tabatinga, and São Gabriel da Cachoeira. This underscores the need for careful attention when deploying TS methodologies, especially in regions with distinct environmental dynamics.

The workflow described aligns with existing references and can be replicated in similar environmental contexts worldwide, such as those in India, Indonesia, and Africa, by leveraging satellite-derived data to overcome limitations posed by sparse ground station coverage and measurement biases. In future research, it's essential to incorporate considerations of greenhouse gas (GHG) concentrations, especially concerning the escalating incidence of rainforest fires linked primarily to deforestation activities. Additionally, the outlined methodology holds promise for estimating the potential for thermal electricity generation from renewable energy sources within the region. Augmenting the input dataset with environmental monitoring data, particularly focusing on CO2 and methane concentrations, and leveraging other ground-based data sources can enhance the accuracy and robustness of predictive models. However, inherent limitations stemming from data sparsity and quality issues must be acknowledged and addressed in subsequent research endeavors.

References:

1. World Economic Forum. (2022). Four Innovations Preparing Cities for Climate Change. Retrieved from https://www.weforum.org/agenda/2022/10/innovations-protect-cities-climate-change/

2. Empresa de Pesquisa Energética. (n.d.). Ten-Year Energy Expansion Plan. Retrieved from https://www.epe.gov.br/sites-en/publicacoes-dados-abertos/publicacoes/Paginas/PDE-2031---English-Version.aspx
3. Sousa, S. R. O., da Silva, W. V., Kaczam, F., et al. (2022). The relationship between socioeconomic development, renewable energies and the innovative process. International Journal of Energy Sector Management, 16(6), 1037–1063. https://doi.org/10.1108/IJESM-05-2021-0020
4. Herman, S. (2022). Something New Under the Sun: Floating Solar Panels. Retrieved from https://www.voanews.com/a/something-new-under-the-sun-floating-solar-panels-/6794529.html
5. Pouran, H. M., Lopes, M. P. C., Nogueira, T., et al. (2022). Environmental and technical impacts of floating photovoltaic plants as an emerging clean energy technology. iScience, 25, 105253. https://doi.org/10.1016/j.isci.2022.105253
6. Maka, A. O., Alabid, J. M. (2022). Solar energy technology and its roles in sustainable development. Clean Energy, 6, 476–483. https://doi.org/10.1093/ce/zkac023
7. Frackiewicz, M. (n.d.). The Challenges of Satellite Communications in Remote Areas. Retrieved from https://ts2.space/en/the-challenges-of-satellite-communication-in-remote-areas/
8. Deepshikha Aggarwal, Deepti Sharma, & Archana B. Saxena. (2023). Adoption of Artificial Intelligence (AI) For Development of Smart Education as the Future of a Sustainable Education System. *Journal of Artificial Intelligence,Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172*, *3*(06), 23–28. https://doi.org/10.55529/jaimlnn.36.23.28
9. De Freitas Viscondi, G., Alves-Souza, S. N. (2018). A systematic literature review on big data for solar photovoltaic electricity generation forecasting. Sustainable Energy Technology and Assessments, 31, 54–63. https://doi.org/10.1016/j.seta.2018.11.008
10. Tina, G. M., Ventura, C., Ferlito, S., et al. (2021). A state-of-art review on machine-learning based methods for PV. Applied Sciences, 11, 7550. https://doi.org/10.3390/app11167550
11. Bamisile, O., Cai, D., Oluwasanmi, A., et al. (2022). Comprehensive assessment, review, and comparison of AI models for solar irradiance prediction based on different time/estimation intervals. Scientific Reports, 12, 9644. https://doi.org/10.1038/s41598-022-13652-w
12. Gürel, A. E., Agbulut, U., Bakir, H., et al. (2023). A state-of-the-art review on estimation of solar radiation with various models. Heliyon, 9, e13167. https://doi.org/10.1016/j.heliyon.2023.e13167
13. Michiorri, A., Sempreviva, A. M., Philipp, S., et al. (2022). Topic taxonomy and metadata to support renewable energy digitalisation. Energies, 15, 9531. https://doi.org/10.3390/en15249531
14. Alkhayat, G., Mehmood, R. (2021). A review and taxonomy of wind and solar energy forecasting methods based on deep learning. Energy and AI, 4, 100060. https://doi.org/10.1016/j.egyai.2021.100060
15. Diagne, M., David, M., Boland, J., et al. (2014). Post-processing of solar irradiance forecasts from WRF model at Reunion Island. Solar Energy, 105, 99–108. https://doi.org/10.1016/j.solener.2014.03.016
16. Iung, A. M., Oliveira, F. L. C., Marcato, A. L. M. (2023). A review on modeling variable renewable energy: complementarity and spatial-temporal dependence. Energies, 16, 1013. https://doi.org/10.3390/en16031013
17. Joseph, V. R. (2022). Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15, 531–538. https://doi.org/10.1002/sam.11583
18. Freund, Y., Schapire, R., Abe, N. (1999). A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence, 14, 771–780.
19. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29, 1189–1232. https://doi.org/10.1214/aos/1013203451
20. Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16) (pp. 785–794). San Francisco, CA, USA. https://doi.org/10.1145/2939672.2939785
21. Singhal, A., Madan, J., & Madan, S. (2023). HCS: A hybrid data security enhancing model based on cryptography algorithms. In V. Goar, M. Kuri, R. Kumar, & T. Senjyu (Eds.), *Advances in information communication technology and computing* (Lecture Notes in Networks and Systems, Vol. 628). Springer, Singapore. https://doi.org/10.1007/978-981-19-9888-1_39

22. Oliveira, F., Rocha, A. P. (2020). Filling missing values in spatial-temporal data collected from traffic sensors. In: 2020 IEEE International Smart Cities Conference (ISC2) (pp. 1–7). Piscataway, NJ, USA. https://doi.org/10.1109/ISC251055.2020.9239016
23. Guia Geográfico. (n.d.). Amazon Map. Retrieved January 10, 2023, from https://www.guiageo.com/amazonas.htm
24. Tabela de Dados das Estações. Instituto Nacional de Meteorologia (INMET). (n.d.). Retrieved January 12, 2023, from https://portal.inmet.gov.br/
25. NASA. (n.d.). The POWER Project. Retrieved January 15, 2023, from https://power.larc.nasa.gov/
26. NASA. (n.d.). TERRA The EOS Flagship. Retrieved May 12, 2023, from https://terra.nasa.gov/about/terra-instruments/ceres
27. NASA. (n.d.). Global Modeling and Assimilation Office. Retrieved May 10, 2023, from https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/
28. Arora, A., & Gupta, P. K. (2021). Data science and its relation to big data and machine learning. *International Research Journal of Modernization in Engineering Technology and Science*, *3*(5), 61. https://www.irjmets.com