Solar Forecasting Using Classical Machine Learning Algorithms

Navapallav Borthakur, *Student,* Vedansh Pandey, *Student,* Vaibhav Joshi, *Student*, under Dr. Avirup Maulik, Department of Electrical Engineering, Indian Institute of Technology (BHU), Varanasi.

***Abstract*—**

As the global energy landscape transitions towards renewable sources, solar power has emerged as a significant contributor. Accurate forecasting of solar energy generation is essential for efficiently integrating solar power into the electricity grid. The variability and intermittency of solar radiation, influenced by factors such as weather conditions and time of day, pose challenges for reliable solar forecasting. This paper reviews cutting-edge solutions in solar forecasting, considering the unique characteristics of solar energy generation. By examining existing models and techniques, we aim to assess their strengths and limitations, emphasizing the need for adaptable forecasting tools. In the context of the evolving energy markets and increasing solar capacity, this review explores opportunities and challenges associated with solar forecasting. Leveraging a geolocalized dataset from solar installations in a specific region, we seek to establish a benchmark for solar forecasting models. Our focus on the nuances of solar energy prediction is particularly relevant for regions experiencing rapid growth in solar installations, providing valuable insights for effective energy planning and grid management.

***Index Terms*— Solar forecasting, Renewable Energy, Weather Influences, Machine Learning, Time Series Analysis**

I. INTRODUCTION

The past decade has witnessed a significant surge in the generation of electrical energy from renewable sources (RES), driven by the increasing demand for electricity and the imperative to mitigate environmental impact. Despite the global economic challenges posed by the Covid-19 pandemic, the future trajectory of RES indicates continued growth. In support of this transition towards renewable energy and to meet the ambitious emissions reduction target of 60% from 1990 levels by 2030, endorsed by the European Commission in September 2020, the European Parliament has sanctioned the allocation of 37% of the coronavirus recovery fund specifically for advancing green initiatives. This strategic allocation underscores a commitment to fostering sustainable practices and accelerating the green transition in the wake of global challenges.

The ascent of nonprogrammable renewable energy sources (RES), particularly photovoltaic (PV) and wind, has underscored the imperative for precise and dependable production forecast methodologies essential for their seamless integration into the electricity grid . The generation of PV power is contingent upon meteorological factors, including solar irradiance and cloud coverage, leading to inherent variability in production . This variability is attributable to deterministic components such as the Earth's movements relative to the Sun, alongside stochastic elements, primarily encompassing cloud movements and weather phenomena.

Solar forecasting is an essential component in the dynamic landscape of the energy industry, given the unique characteristics of solar power generation. Unlike conventional commodities, electricity derived from solar energy is not easily stored and must be generated on demand, necessitating accurate predictions. The overarching objective for solar power companies is to ensure a reliable and secure supply of electricity to end-users. Solar Load Forecasting assumes a pivotal role in the planning and operation of the solar energy sector. Precise forecasts for solar energy demand yield cost savings in operations, enhance reliability in power supply, and inform strategic decision-making for future development. These forecasts are typically categorized based on planning horizons: short-term forecasts covering daily or weekly periods, medium-term forecasts spanning one day to a year, and long-term forecasts extending beyond a year. Short-term forecasts aid in scheduling solar electricity generation and transmission, medium-term forecasts assist in fuel procurement planning, and long-term forecasts facilitate the strategic development of solar power supply and delivery systems. The intricate nature of solar demand patterns, influenced by temporal, social, economic, and environmental factors, necessitates sophisticated forecasting methods. This report explores a spectrum of Solar Forecasting methods, including KNN and Random Forest Regression, highlighting their applicability in capturing the complexities of solar energy demand variations.

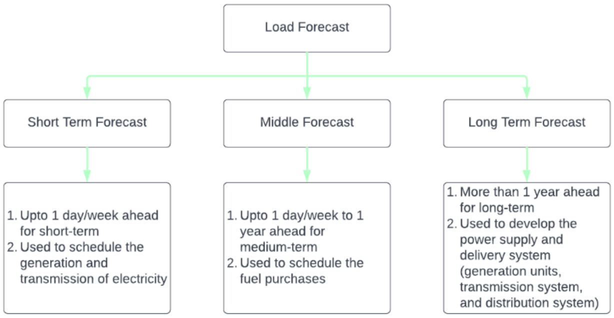

Load Forecast
Short Term Forecast
Middle Forecast
Long Term Forecast
1. Upto 1 day/week ahead for short-term
2. Used to schedule the generation and transmission of electricity
1. Upto 1 day/week to 1 year ahead for medium-term
2. Used to schedule the fuel purchases
1. More than 1 year ahead for long-term
2. Used to develop the power supply and delivery system (generation units, transmission system, and distribution system)

Fig.1. Flow chart of types of Generic Load Forecasting.

Given the inherent non-storability of electricity produced from solar energy, on-demand generation becomes imperative, underscoring the critical importance of accurate Solar Load Forecasting in the planning and operation of the solar energy sector. Similar to electric load forecasting, solar forecasting spans various time periods, including hourly, daily, weekly, monthly, and yearly, stratified into short-term, medium-term, and long-term forecasts based on planning horizons.

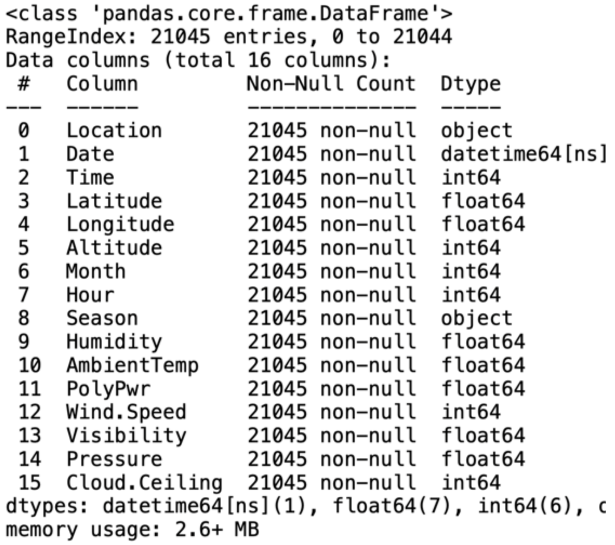
The intricacies within solar demand patterns result from a myriad of factors such as temporal, social, economic, and environmental considerations, necessitating the development of sophisticated forecasting methodologies.

There is no universally applicable method that can effectively handle all cases, especially when multiple factors are considered. Consequently, each solar power output must adopt a customized forecasting approach by modifying general methods to suit their specific circumstances. This paper proposes a practical forecasting methodology that integrates diverse forecasting models for short periods. Our aim is to establish a benchmarking platform for analyzing and advancing algorithms in the field of load forecasting. This file contains power output from horizontal photovoltaic panels located at 12 Northern Hemisphere sites over 14 months. Independent variables in each column include location, date, time sampled, latitude, longitude, altitude, year and month, month, hour, season, humidity, ambient temperature, power output from the solar panel, wind speed, visibility, pressure, and cloud ceiling.

II. Dataset Used for Predictions

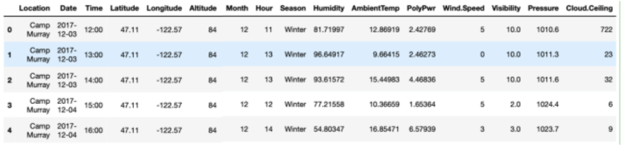
A. 12 Northern Hemisphere sites dataset

The dataset consists of 21,045 rows and 16 columns.
It is a Time Series data - since there's a time attribute to dataset , hence it need to be sorted according to Time.
The target variable is PolyPwr (power output).



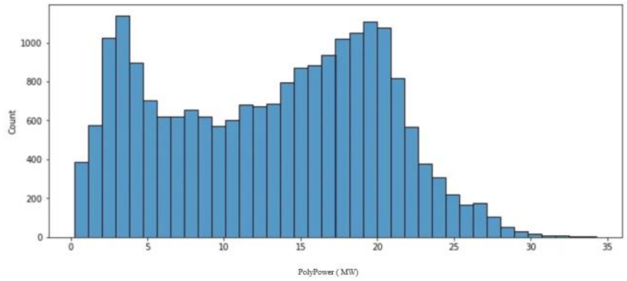
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21045 entries, 0 to 21044
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Location       21045 non-null  object
 1   Date           21045 non-null  datetime64[ns]
 2   Time           21045 non-null  int64
 3   Latitude       21045 non-null  float64
 4   Longitude      21045 non-null  float64
 5   Altitude       21045 non-null  int64
 6   Month          21045 non-null  int64
 7   Hour           21045 non-null  int64
 8   Season         21045 non-null  object
 9   Humidity       21045 non-null  float64
 10  AmbientTemp    21045 non-null  float64
 11  PolyPwr        21045 non-null  float64
 12  Wind.Speed     21045 non-null  int64
 13  Visibility     21045 non-null  float64
 14  Pressure       21045 non-null  float64
 15  Cloud.Ceiling  21045 non-null  int64
dtypes: datetime64[ns](1), float64(7), int64(6), c
memory usage: 2.6+ MB
```

We explored the available columns in the dataset using functions in pandas-Python data analysis library



	Location	Date	Time	Latitude	Longitude	Altitude	Month	Hour	Season	Humidity	AmbientTemp	PolyPwr	Wind.Speed	Visibility	Pressure	Cloud.Ceiling
0	Camp Murray	2017-12-03	12:00	47.11	-122.57	84	12	11	Winter	81.71997	12.86919	2.42769	5	10.0	1010.6	722
1	Camp Murray	2017-12-03	13:00	47.11	-122.57	84	12	13	Winter	96.64917	9.66415	2.46273	0	10.0	1011.3	23
2	Camp Murray	2017-12-03	14:00	47.11	-122.57	84	12	13	Winter	93.61572	15.44983	4.46836	5	10.0	1011.6	32
3	Camp Murray	2017-12-04	15:00	47.11	-122.57	84	12	12	Winter	77.21558	10.36659	1.65364	5	2.0	1024.4	6
4	Camp Murray	2017-12-04	16:00	47.11	-122.57	84	12	14	Winter	54.80347	16.85471	6.57939	3	3.0	1023.7	9

First five data-points



Distribution of Target

B. Data Preprocessing

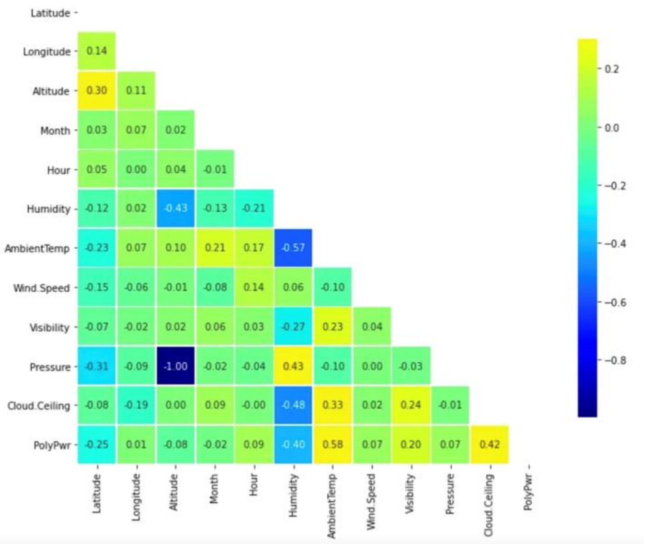
Maintaining the comprehensiveness and consistency of data holds utmost importance when it comes to precise analysis and prediction. Nevertheless, when data is gathered from online sources, the possibility of encountering missing values arises due to factors like website downtime or maintenance. In the specific context of our study, it is plausible that the data extracted from the dataset may contain certain missing values resulting from such circumstances. To address this concern, conventional techniques such as forward fill and backfill were employed to impute the missing values within the dataset. By applying these techniques, the data was rendered more consistent and complete, thereby ensuring the reliability of subsequent analysis and prediction tasks.

C. Correlation Matrix

The correlation matrix serves as a valuable tool for understanding the relationships between different features and the power output in a solar energy system. Upon visualizing the correlation matrix, it becomes evident that certain features exhibit stronger correlations with solar power output than others. Notably, ambient temperature, cloud ceiling, and humidity emerge as the top three most correlated features. This insight implies that fluctuations in these meteorological factors significantly impact the efficiency and generation of solar power. Understanding these correlations is crucial for accurate solar forecasting and system optimization, allowing for informed decisions in the planning and operation of solar energy installations.

Furthermore, an interesting observation is the significance of latitude in relation to power output. The correlation plot reveals that latitude exhibits a notable correlation with solar power output, emphasizing the role of geographic location in influencing solar energy generation. On the contrary, longitude does not demonstrate the same behavior and, as a result, was excluded from the modeling process. This decision underscores the importance of selectively choosing relevant features based on their actual impact on solar power output, enhancing the precision and efficiency of predictive models.

In the context of feature selection, altitude was also excluded from the modeling process. While altitude presents a perfect correlation with pressure, it was deemed less pertinent due to its lack of variation for a given location. This decision streamlines the modeling process by focusing on features that contribute meaningfully to the variability in solar power output, thus optimizing the accuracy and applicability of the forecasting model.



III. Methods

This paper explores 2 Machine Learning Models K Nearest Neighbours and Random Forest Regression, which have been applied to the dataset discussed in the previous section.

A. Random Forest Regression

Random forest regression is an ensemble learning method that combines multiple decision trees to make a prediction.

Each decision tree is trained on a subset of the training data and with a subset of the features. Random forest regression is a powerful algorithm that can handle large datasets with a large number of features. It can also handle missing data and outliers well.

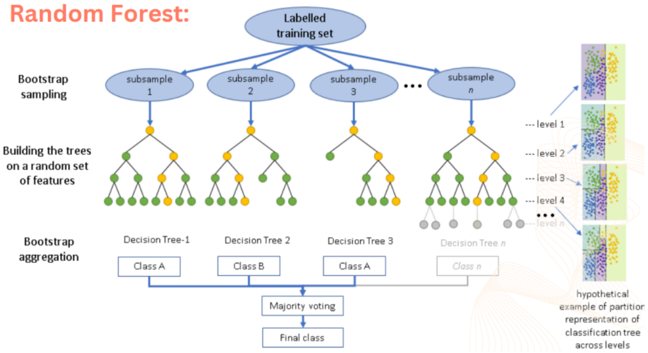
Random forest regression is robust to overfitting, which is a common problem in decision tree algorithms. The algorithm uses a technique called bagging, which involves training multiple trees on different subsets of the data and features.

We have not used Decision Tree Regression as Random Forest Regression is an Improvement upon Decision Tree, Thus Random Forest Regression is used.

Random forest regression is relatively easy to implement and requires minimal tuning of hyperparameters. Random forest regression can be used for both continuous and categorical variables, making it a versatile algorithm for a wide range of applications. Random forest regression is relatively easy to implement and requires minimal tuning of hyperparameters.

One potential disadvantage of random forest regression is that it can be slower to train than other regression algorithms, especially on large datasets.

Random forest regression can also be difficult to interpret compared to some other regression algorithms, such as linear regression.

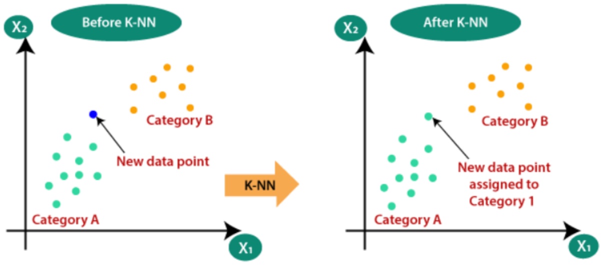


D. K-Nearest Neighbours

K-Nearest Neighbors (KNN) stands as a powerful supervised machine learning algorithm renowned for its simplicity and effectiveness. Operating as a non-parametric and instance-based algorithm, KNN relies on the fundamental concept of finding the K nearest neighbors of a new data point within the training set. The determination of proximity is typically achieved through a distance metric, with Euclidean distance being a commonly employed measure. Once the K neighbors are identified, the algorithm predicts the value for the new data point based on the majority or average response from its neighbours. This mechanism makes KNN particularly adept at handling situations where the decision boundary is highly irregular or when the data distribution lacks a discernible pattern.

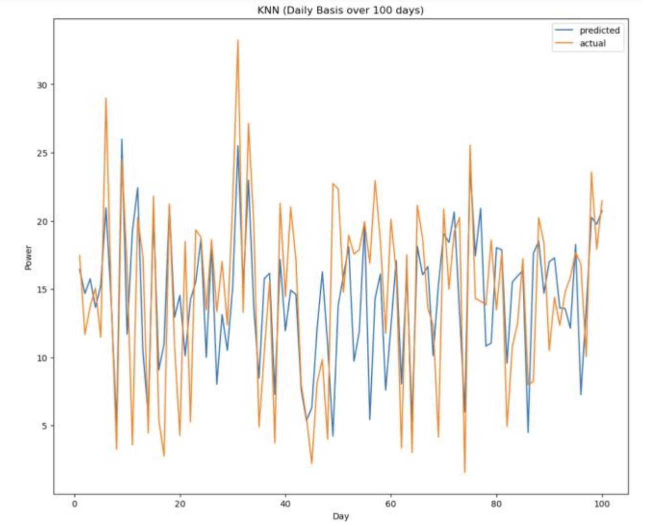
One of the prominent strengths of KNN lies in its simplicity of implementation, making it an attractive choice for practitioners seeking an uncomplicated yet effective algorithm. Its straightforward nature facilitates quick adoption and application across various domains, especially in scenarios involving small to medium-sized datasets. Additionally, KNN exhibits robustness in the face of noisy data, demonstrating an ability to navigate and make accurate predictions even when confronted with imperfect or erratic data points. This resilience to noise contributes to the algorithm's versatility, making it well-suited for real-world datasets that may contain uncertainties or irregularities.

While KNN offers notable advantages, it is essential to acknowledge its limitations. The algorithm's computational cost can escalate with larger datasets, as the need to calculate distances to all data points in the training set becomes more resource-intensive. Consequently, practitioners often weigh the simplicity and robustness of KNN against considerations of scalability, opting for this algorithm in scenarios where its strengths align with the characteristics of the dataset at hand.

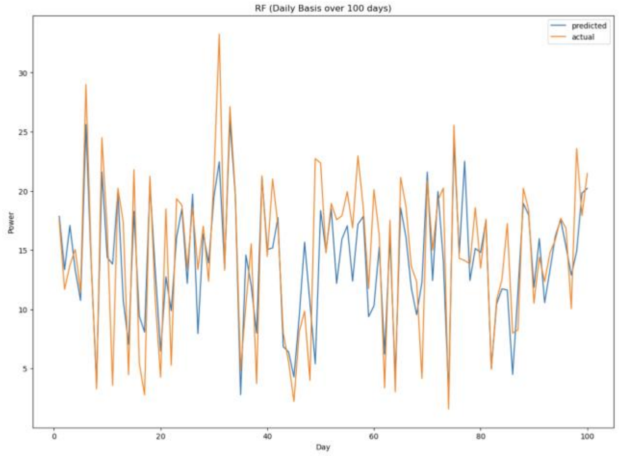


IV. RESULTS

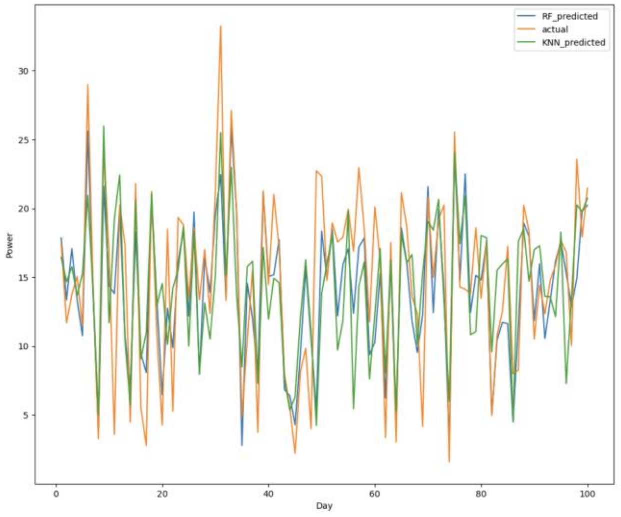
The following section presents the outcomes of the aforementioned methodologies applied to the load data for different dates. In all the plots below, the actual load data obtained from the dataset is displayed alongside the load curve forecasted by the aforementioned methods. The real load curve predominantly exhibits the characteristic day-to-day variations.



Results of KNN



Results of RFR



Comparison between RFR and KNN

Evaluation Metrics Used

- R-Squared

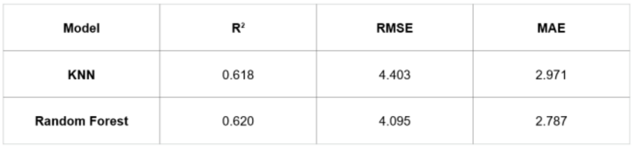
$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- Root-mean Square error.

$$RMSE = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}$$

- Mean Absolute error

$$MAE = \frac{1}{n}\sum_i \left|y_i - \hat{y}_i\right|$$



Model	R²	RMSE	MAE
KNN	0.618	4.403	2.971
Random Forest	0.620	4.095	2.787

Comparison between different Models

IV. Conclusion and future prospects

In conclusion, Solar Forecasting stands as a critical element in solar engineering, given the intricacies involved in managing transmission and load networks. This research undertook an extensive literature review and implemented cutting-edge techniques for solar load forecasting. Contrasting with traditional statistical methods, our focus shifted to machine learning approaches, specifically Random Forest Regression and K-Nearest Neighbors (KNN), to address the challenges inherent in capturing intricate temporal patterns within solar time series analysis.

Our findings reveal that Random Forest Regression and KNN models exhibit promising performance in short-term solar load forecasting. These machine learning techniques, with their ability to adapt to complex patterns and dependencies, showcase potential in enhancing the precision of solar forecasting, especially when considering factors like weather conditions and humidity. Unlike traditional statistical methods, the adaptability of Random Forest and KNN to diverse patterns positions them as promising tools for the nuanced nature of solar energy generation.

Looking ahead, the research aims to prototype these algorithms using real-time data from solar installations. The intention is to benchmark their performance, providing valuable insights into the practical relevance of Random Forest and KNN models in solar forecasting. Furthermore, considering the success of hybrid models in load forecasting, future investigations may explore combining the strengths of Random Forest and KNN with other techniques to potentially enhance the accuracy and robustness of solar load forecasting. This iterative approach, coupled with the continual exploration of novel techniques in machine learning, presents exciting opportunities for advancing the field of solar forecasting.

References

[1] Dynamic Forecasting of Solar Energy Microgrid Systems Using Feature Engineering
Author: Muamar Mohamed , Farhad E. Mahmood , Mehmmood A. Abd, Ambrish Chandra , Fellow, IEEE,

[2] A New Probabilistic Ensemble Method for an Enhanced Day-Ahead PV Power Forecastand Alfredo
Author: Silvia Pretto, Emanuele Ogliari , Member, IEEE, Alessandro Niccolai