HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 4 (2024)

Machine Learning Prediction of AC Power Output in Solar Cells

Henry Yellin

Frisch School, USA

ABSTRACT

Solar power is used to combat climate change and promote sustainable development. However, the efficiency and reliability of solar energy systems are heavily reliant on accurately forecasting solar output. We aimed to identify the most effective machine learning algorithm from the scikit-learn library for predicting the AC power output of solar cells. Unlike previous models that failed to capture changes in environmental conditions or complex dependencies influencing solar cell performance, we found a model that integrates various input parameters, including daily yield, total yield, ambient temperature, module temperature, irradiation, and month. Our goal was to find an algorithm that would enable accurate forecasts across different seasons and improve long-term prediction capabilities. This robustness is crucial for real-world applications, helping stakeholders in the energy sector make informed decisions, enhance grid reliability, promote renewable energy integration, and expedite the shift toward sustainable energy. Through testing various algorithms, the Random Forest Regressor model demonstrated highest accuracy with an R-Squared score of 0.968. This indicates the model's proficiency in identifying key factors affecting solar energy generation and predicting future solar output with minimal error across a range of renewable energy applications. Grid operators can use the model's predictions to optimize power distribution, while solar companies can enhance module and tracker placement for better efficiency. With the potential to improve financial returns for solar investors and strengthen the bankability for finance partners, this model emerges as a valuable tool for developers, vendors, and energy offtake partners in maximizing the potential of renewable energy installations.

Introduction

As citizens and policymakers have become increasingly concerned with the deleterious environmental impacts of fossil fuels, there has been increasing demand for sustainable and efficient energy sources. This has led to a growing reliance on solar power energy for residential and commercial applications. As of February 2024, solar energy accounted for 3.9% of all power generation in the United States. By 2030, solar energy will constitute 30% of all energy generation in the United States at an approximate investment cost of $120 billion dollars (Solar Energy Industries Association [SEIA], 2021). As solar energy spreads across America, the need for accurate predictions of how solar cells, especially amidst unprecedented weather fluctuations, will behave becomes more essential (U.S. Energy Information Administration [EIA], 2024).

The primary question driving this paper is: How can we predict the output of solar cells to optimize the harvesting of radiant energy? Accurate predictions can lead to better energy planning and better efficiency for solar across the renewable industry sector. This will drive lower total cost of ownership, which will lead to more solar in the United States and less consumption of fossil fuels. This research paper introduces a machine learning model that can forecast the AC power output of solar cells to address this issue.

The models investigated in this research rely on data from the scikit-learn library (Buitinck et al., 2013) as its foundational component. It integrates various inputs including daily yield, total yield, ambient temperature, module temperature, irradiation, and the month of the year. In the context of evaluating artificial intelligence (AI) algorithms used for predicting solar cell output, daily yield and total yield refer to two important metrics that help in understanding and assessing the performance of solar energy systems. Daily yield represents the amount of electricity generated by a solar panel or a solar energy system within a single day. It is typically measured in units such as kilowatt-hours (kWh) or megawatt-hours (MWh). Daily yield is influenced by various factors including solar irradiance, ambient temperature, shading, system efficiency, and tilt angle of the solar panels. AI algorithms can utilize data on these factors, along with historical performance data, weather forecasts, and other relevant information, to predict the daily yield of solar cells for a given day. Predicting daily yield accurately is crucial for estimating daily energy production and optimizing system performance.

Total yield represents the cumulative amount of electricity or energy production efficiency generated by a solar panel or a solar energy system over a certain period, typically since the installation or commissioning of the system. Like daily yield, total yield is also measured in units such as kilowatt-hours (kWh) or megawatt-hours (MWh). AI algorithms can analyze historical data on total yield, along with environmental and operational factors, to identify trends, patterns, and potential areas for optimization. Predicting total yield accurately is essential for assessing the long-term viability and economic viability of solar energy projects.

Ambient temperature and module temperature influence the efficiency of solar panels and overall system performance. High temperatures can reduce the efficiency of solar cells, leading to decreased energy output. By monitoring ambient and module temperatures in real-time and incorporating this data into predictive models, AI algorithms can adjust system parameters such as tilt angle, ventilation, or cooling mechanisms to mitigate temperature-related losses and optimize energy production.

Solar irradiation is another key determinant of energy generation in solar energy systems. By quantifying the amount of solar radiation reaching the solar panels, AI algorithms can estimate potential energy output under different weather conditions. By analyzing historical irradiation data and using weather forecasting models, AI algorithms can anticipate variations in solar irradiance and adjust system parameters accordingly to maximize energy production efficiency.

Time-related factors such as the time of day, season, and month of the year influence solar energy generation patterns. For example, solar irradiance levels vary throughout the day and across different seasons due to changes in the sun's position and atmospheric conditions. By considering time-related factors in predictive models, AI algorithms can optimize the scheduling of energy generation, storage, and distribution activities to align with periods of peak solar availability and maximize energy production efficiency.

By integrating these diverse inputs and influences on solar cell behavior, such as daily yield, total yield, ambient temperature, module temperature, irradiation, and time-related factors, this study explores which AI algorithms can be used to develop predictive models that identify optimal operating conditions and adjust system parameters to maximize energy production efficiency in solar energy systems. Identifying such a model has the potential to facilitate data-driven decision-making and enable proactive management of solar energy resources, ultimately enhancing the economic viability and sustainability of solar energy projects.

This research addresses a supervised regression problem, where the algorithm is trained on a dataset predicting numeric values. Specifically, the output of the model focuses on the AC power output, a pivotal metric for assessing the performance of solar cells. The dataset comprises two distinct categories of numerical variables: weather-related data and solar-specific data. The overarching objective of this study is to leverage factor dependencies within these datasets to aid in predicting the future behavior of solar cells using Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and Decision Tree Regressor.

Background

Several researchers have demonstrated the potential of using AI applications in solar energy. In Yap's (2022) paper, "Artificial intelligence based MPPT techniques for solar power system: A Review," as well as Kah's (2022) paper "Artificial intelligence techniques in solar energy applications." Yung and Kah explore how AI algorithms enhance the efficiency and performance of solar power systems by optimizing maximum power point tracking (MPPT). Yung's research delves into the development and implementation of AI-based MPPT techniques. By utilizing algorithms such as artificial neural networks (ANNs) or genetic algorithms (GAs), Yap aimed to improve the accuracy and speed of MPPT in solar photovoltaic (PV) systems. These techniques enable the system to dynamically adjust to changing environmental conditions and extract maximum power from the solar panels, thereby enhancing overall energy harvesting efficiency. However, the scalability, adaptability and effectiveness of these techniques may be limited and vary depending on factors such as system size, environmental conditions, and hardware constraints, falling short in scenarios involving shading or unexpected variables. Additionally, the computational complexity and resource requirements associated with AI-based MPPT techniques could pose challenges for real-time implementation in large-scale PV systems or in environments with limited processing capabilities. Furthermore, the performance of AI algorithms may degrade over time due to changes in environmental conditions or system dynamics, requiring continuous monitoring and retraining to maintain optimal performance.

Similar to my work, Yung and Kah's work focused on AI applications in predicting solar irradiance, the amount of solar energy received, particularly in regions like Johannesburg. By leveraging AI algorithms such as support vector machines (SVMs) or deep learning models, Kah aimed to forecast solar irradiance levels with high accuracy. This predictive capability is crucial for optimizing the operation of solar power systems, as it allows for better planning and management of energy generation, storage, and distribution. One limitation of Kah's study is the reliance on specific regional data, such as that from Johannesburg. While focusing on a specific region allows for a detailed analysis of local factors influencing solar irradiance prediction, it may limit the generalizability of the findings to other locations with different climatic conditions, geographic features, and solar energy infrastructures. Additionally, the accuracy of AI models for predicting solar irradiance is highly dependent on the quality and quantity of available data, as well as the performance of the selected algorithms. Therefore, while Kah's work provides valuable insights into the application of AI in solar irradiance prediction for regions like Johannesburg, its findings may not directly translate to other geographic locations without further validation and adaptation to local conditions.

Kalogirou and Arzu Şencan (2010) also explored the application of various AI techniques in different aspects of solar energy, particularly focusing on optimization. Their study encompasses a range of AI techniques, including but not limited to ANNs, GAs, SVMs, and deep learning models. Since their research primarily focuses on the application of AI techniques in solar energy optimization, it may not fully capture the nuances or challenges specific to different geographical locations, climates, or types of solar energy systems. Additionally, the effectiveness of such AI techniques may vary depending on the availability and quality of data, as well as the specific characteristics of the solar energy infrastructure being optimized. Therefore, while their study provides valuable insights into the potential of AI in solar energy optimization, the applicability of their findings may require further validation across diverse contexts and settings.

While these studies highlight the potential of AI-driven approaches in understanding solar energy systems, ensuring the robustness and reliability of these algorithms in dynamic operating conditions is essential for their practical utility in solar energy systems. AI algorithms must accurately predict the output of solar cells in various conditions to ensure reliable energy generation estimates. In this study we examine whether the inclusion of diverse inputs, such as daily yield, total yield, ambient temperature, module temperature, irradiation and the month of the year can be used in an AI algorithm model to accurately and precisely capture the complex interactions between various environmental factors and solar cell performance, thus improving accuracy and prediction for solar output. Such a model can be used to facilitate the identification of optimal operating conditions, fine-tune system parameters to maximize energy production efficiency and generalize across a range of scenarios to improve robustness to variability.

Data Collection and Methodology

This study used a dataset from Kaggle called "Solar Power Generation Data Solar power generation and sensor data for two power plants." This dataset is derived from a collection effort across two solar power plants in India. Each plant had a power generation dataset and a corresponding sensor readings dataset of weather conditions. The data was collected for AC power, DC power, daily yield, total yield, ambient temperature, module temperature, irradiation, and the time every 15 minutes. The power generation data is measured at the inverter level, which means it measures the amount of power at the place where the DC power gets converted into AC power and at a point of minimal power loss. For reference, solar radiation comes into the module using DC power and gets converted into AC power which the grid uses. The sensor data is gathered at the ground level of the plant by using sensors (Kannal, 2020).

The dataset is structured in four distinct parts, two containing weather-related information such as ambient temperature, module temperature, and irradiation, and the remaining containing power related information such as AC power, DC power, daily yield, and total yield. Before initializing the model, preprocessing steps were conducted to ensure the accuracy and reliability of the data for analysis by the machine learning model. A crucial step involved standardizing the date column format across the dataset. This was achieved using the Pandas library's Datetime module to ensure consistency in the time columns across all files. Subsequently, a merge operation was employed to consolidate the weather and solar datasets for each plant. This merging process facilitated the combination of relevant data into cohesive units. Finally, the same operation was utilized to merge all datasets into a single, comprehensive CSV file, streamlining the data preparation process for model training and evaluation.

To validate the effectiveness of preprocessing, the dataset was examined for null values, which signify missing data within cells. Prior to merging all four files, the dataset exhibited no null values. However, post-merging, a notable challenge surfaced: 23,098 null values were detected. This discrepancy stemmed from asynchronous data retrieval between the weather and solar datasets. Specifically, the solar data contained multiple identical copies of the same line, while the weather data had fewer additional values. To mitigate this issue of unnecessary redundancy, instances with null values were removed. This approach ensured the dataset's integrity and consistency, facilitating more accurate and reliable analysis by the machine learning model.

To ensure comprehensibility for both the model and the reader, certain features required additional processing known as discretization, which involves converting non-numerical information into numerical form. For instance, the plant number feature initially utilized a lengthy array of numbers to denote each plant it referenced. We employed discretization to transform this into easily interpretable numerical representations for each plant (e.g., Plant 1 and Plant 2).

In summary, the dataset utilized in this paper comprises solar power generation and environmental data collected from two plants in India. After undergoing several preprocessing steps, such as datetime standardization, data merging, and resolution of null value issues, the dataset was optimized for effective utilization in machine learning models. Before proceeding with modeling, Figure 1 displays a histogram illustrating the frequency of AC power output. It is notable that a significant portion of the output registers as zero, primarily attributed to the absence of solar generation during nighttime, which spans half of the day.

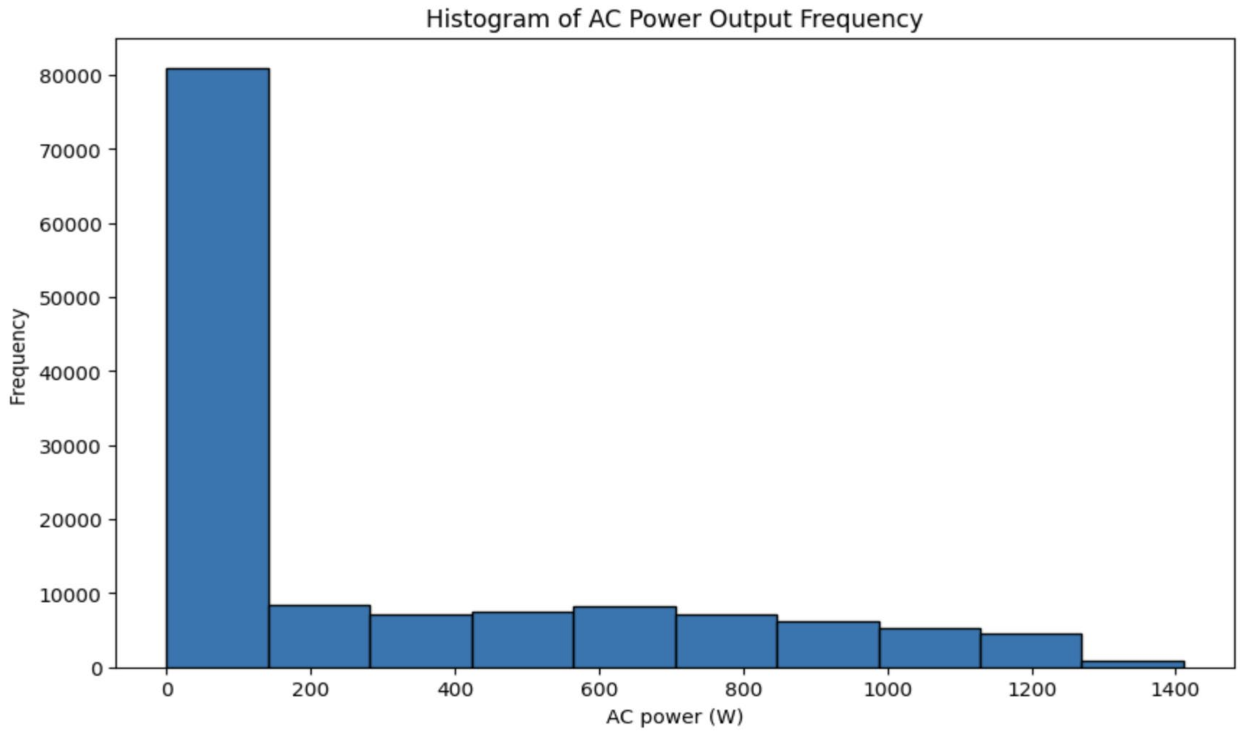

Figure 1. Histogram of AC Power Output Frequency Highlighting Zero Output During Nighttime Hours

Figure 2 shows a positive correlation between irradiation and AC power output. This means that as the amount of sunlight hitting the solar panels increases, the amount of electricity produced also increases. This makes sense, as solar panels convert sunlight into electricity. The graph more importantly shows that there is evident heteroskedasticity between irradiation and AC power output. In other words, the variance of the errors of irradiation and AC power output is not constant across the observations. This is due to the fact there are a number of explanatory factors that can affect the efficiency of solar panels, such as ambient temperature, dust, and shading, which we try to account for in this study.

Figure 2. Evidence of Heteroskedasticity Between Irradiation and AC Power Output

To predict the AC power output of solar cells, various machine learning techniques were utilized from the scikit-learn library in Python, specifically Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and Decision Tree Regressor. The performance of each model was evaluated using various metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and the R^2 score. MAE measures the average difference between the predicted and actual values. Lower MAE values mean a smaller average difference between predicted and actual values. MSE is the average of the squared differences between predicted and actual values. Squaring the differences means larger errors and a lower MSE means less overall error; it may be influenced by outliers because one far off outlier subtracted by the actual value squared can heavily impact results. RMSE is the square root of the MSE. Similar to MSE, a lower RMSE suggests better model accuracy. R^2, also known as the coefficient of determination, is a commonly used metric to determine the performance of a model. The R^2 Score ranges from 0 to 1, where 1 means our model perfectly predicts the data, and 0 means it doesn't explain any variation. So, the closer the R^2 Score is to 1, the better the model is at capturing patterns in the data. Once different metrics had been identified, the data was split into training, validation, and test data sets (Chugh, 2020).

The dataset, encompassing parameters such as daily yield, total yield, ambient temperature, module temperature, and irradiation, underwent partitioning into features denoted by a capital X and the target variable denoted by a lowercase 'y,' with AC_POWER representing the output to be predicted. Each model independently predicted the 'y' when provided with the X in various manners, yielding diverse results.

The dataset comprises numerical data pertaining to power generation and sensor readings, totaling 136,476 samples. To facilitate model development, the dataset was partitioned into training and testing sets, with 80% allocated for training and the remaining 20% reserved for evaluating model effectiveness. This 80/20 split methodology ensures a balance between providing sufficient data for the model to learn from and retaining a representative portion for accurate testing. This separation allowed for robust model training on the majority of the dataset while enabling an unbiased evaluation of its predictive capabilities on unseen data. Different algorithms then were applied to the model.

For the first predictive model, the Linear Regression algorithm was used. Linear Regression is a method that finds the best-fitting line through data points. It considers the relationship between input and output as a straight line. During training, the algorithm makes the line that best fits the data. When a new input is provided, the model uses this line to make predictions about the output (Kanade, 2022). Next, Ridge Regression was used, an extension of Linear Regression. Like Linear Regression, Ridge regression aims to fit a line that best captures the data, but it incorporates a penalty term to prevent overfitting, allowing the model to generalize well with new data. Essentially, Ridge regression strikes a balance between fitting the data accurately and avoiding excessive complexity in the model (Ashok, 2024). Then the Lasso Regression was implemented, which bears resemblance to Ridge regression in its use of regularization to counteract overfitting. The key distinction lies in the mathematical formula of the regularization penalty. While Ridge employs the sum of squared values of the coefficients, known as L2 regularization, Lasso adopts the sum of absolute values of the coefficients, referred to as L1 regularization (Jain, 2023).

Next, Random Forest Regressor algorithm was explored, a machine learning model that constructs numerous decision trees, each trained on different subsets of the data. During training, these decision trees are built using random portions of the dataset, resulting in diverse and random trees. When making predictions, each individual tree contributes to the final output, thereby enhancing prediction accuracy. During testing, each tree generates its own prediction, and the final prediction is derived from averaging the outputs of all trees in the forest (Beheshti, 2022).

Lastly, the Decision Tree Regressor was used, a system that partitions the data into smaller subsets and makes decisions at each step based on features such as daily yield, total yield, ambient temperature, module temperature, irradiation, and the month. It constructs a tree-like structure of decisions to arrive at a final prediction for the target variable, in this case, the AC power output of solar cells. During training, it learns to partition the data based on certain conditions to improve prediction accuracy (Verma, 2023).

In the scikit-learn library, modifying hyperparameters for models such as Linear Regression, Ridge, Lasso, Random Forest Regressor, and Decision Tree Regressor can play a crucial role in fine-tuning machine learning models to achieve optimal performance and generalization on unseen data. They enable the customization of the behavior and complexity of models according to the specific characteristics of the dataset and the requirements of the prediction task.

For Linear Regression, the intercept term parameter can be adjusted to preprocess variables before the data is subjected to regression. In the case of Ridge and Lasso, altering the α parameter enables control over the level of regularization, with higher values implying more stringent regularization. Regarding Random Forest Regressor, the maximum tree depth parameter allows selection of the number and depth of trees in the forest. Similarly, Decision Tree Regressor incorporates a maximum tree depth variable, which not only influences the number of trees but also impacts the model's complexity. By tuning these hyperparameters, we can manually improve the models' efficiency and performance. The best approach for determining the optimal hyperparameter for each model is through trial and error and plotting these on a graph. In this case, adjusting the hyperparameters of Linear Regression, Ridge, and Lasso would be inefficient and ineffective. Given that the data does not conform neatly to a linear chart, attempting to overly complicate or regularize these models would yield minimal benefits. However, adjusting the maximum tree depth parameter with respect to the Random Forest Regressor and Decision Tree Regressor models helped control for the complexity of decision trees, struck a balance between bias and variance, and optimized the model's performance and prediction task.

Results

The effectiveness of each model was assessed using different metrics to give insights into their ability to forecast solar cell power output. In the initial phase of the analysis, baseline versions of each model were used to gauge their predictive capabilities. These baseline models served as a reference point for subsequent adjustments through hyperparameter tuning, allowing us to assess the raw performance of each algorithm before optimization.

The Linear Regression model exhibited poor performance, with a high MSE of 28898.71, a similarly high MAE of 85.49 and relatively low R^2 score, 0.798. Its inadequacy can be attributed to its simplistic assumption of linear relationships, which fails to account for the nonlinear factors inherent in solar generation, such as weather conditions and time of day.

Subsequently, the Ridge model yielded comparable results, with similar MSE, MAE, RMSE and R^2 to Linear Regression, as shown in Table 1. This indicates that Ridge did not significantly outperform Linear Regression, suggesting that its regularization penalties did not enhance its accuracy. Similarly, the Lasso model performed slightly worse than Ridge and Linear Regression, with a marginally higher MSE (29005.62) and MAE (85.79). This suggests that the variance in regularization formulas did not necessarily improve the model's output. Conversely, the Decision Tree Regressor demonstrated superior performance, with notably lower MSE (9372.14), MAE scores (26.23) and R^2 score of 0.934. Its ability to capture the complex dependencies of solar generation within its decision trees allowed it to effectively predict outliers and rare occurrences. Finally, the Random Forest Regressor emerged as the top performer, boasting the lowest MSE (4618.42) and MAE (24.76), with the highest R2 score of 0.968 among all models. Its robustness in handling intricate dependencies and relationships within the data makes it particularly well-suited for solar output prediction.

In summary, while Linear Regression and its variants struggled due to their oversimplified assumptions about solar generation, the Decision Tree Regressor showcased better performance, and the Random Forest Regressor emerged as the most effective model, highlighting its capability to account for numerous outliers and dependencies that affect AC power output for solar cells.

Table 1. Final model metrics after hyperparameter tuning

Model	**MAE**	**MSE**	**RMSE**	**R-squared (R^2) Score**
Linear Regression	85.49	28898.71	169.996	0.798
Ridge Regression	85.49	28897.93	169.994	0.798
Lasso Regression	85.79	29005.62	170.310	0.797
Random Forest Regressor	21.40	4618.42	67.959	0.968
Decision Tree Regressor	21.03	9372.14	96.810	0.934

After finalizing our model structures, we find the metrics outlined in Table 1. In our study, adjusting the hyperparameters of Linear Regression, Ridge Regression, and Lasso Regression were inefficient and ineffective. Given that the data does not conform neatly to a linear chart, attempting to overly complicate or regularize these models would yield minimal benefits. However, adjusting the maximum tree depth parameter with respect to the

Random Forest Regressor and Decision Tree Regressor models helped control for the complexity of decision trees, struck a balance between bias and variance, and optimized the model's performance and prediction task.

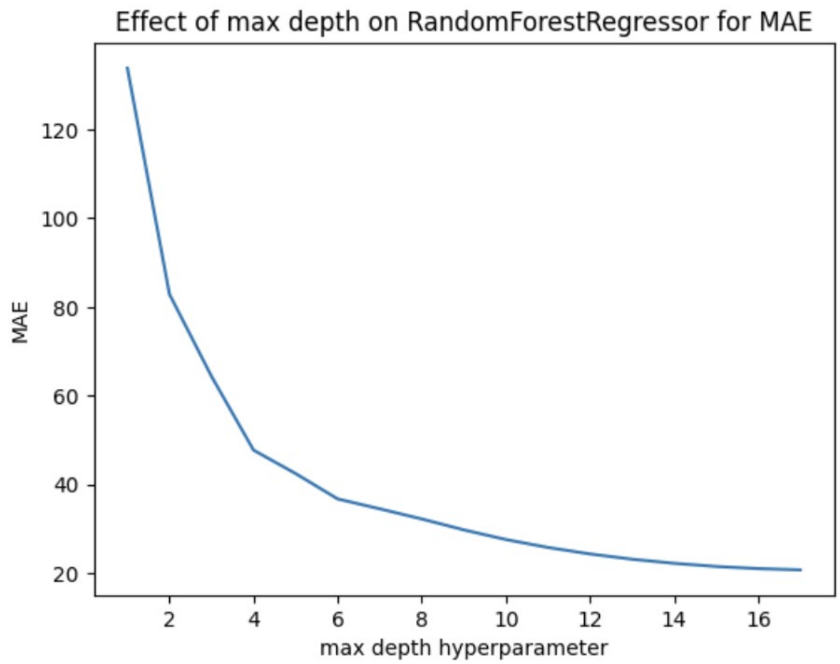

Figure 3. Effects of maximum tree depth parameter for MAE on Random Forest Regressor
For instance, the graph in Figure 4 demonstrates that the optimal maximum tree depth for the Random Forest Regressor is 14, resulting in a MAE of 21.40, as compared to the initial 24.76 MAE before adjusting the hyperparameters.

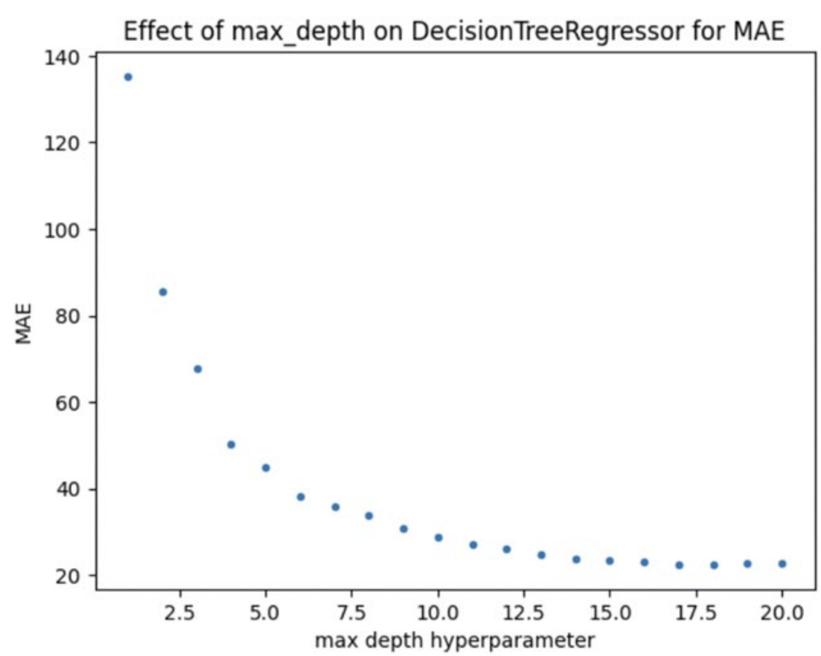

Figure 4. Effects of maximum tree depth parameter for MAE on Decision Tree Regressor

Similarly, the graph in Figure 5 shows that the optimal maximum tree depth for the Decision Tree Regressor is 19.5, producing a 21.03 MAE compared to the 26.23 MAE before changing the hyperparameters. Although the linear regression, ridge, and Lasso models struggled to accurately predict the intricate nature of solar generation data due to their lack of model complexity, the Decision Tree Regressor and Random Forest Regressor models exhibited more promising results.

Discussion

This research paper employs a machine learning model to predict the AC power output of solar cells, catering to the growing demand for precise predictions in the renewable energy sector. This study underscores that while Linear Regression, Ridge, and Lasso models struggled to capture the complexities of solar generation due to their simplicity, the Decision Tree Regressor and Random Forest Regressor exhibited more promising outcomes. Utilizing the Decision Tree Regressor algorithm alongside various techniques from the scikit-learn library, the model leverages multiple input parameters such as daily yield, total yield, ambient temperature, module temperature, irradiation, and the month of the year to accurately forecast solar output.

Through evaluation metrics including MAE, MSE, RMSE, and the R^2 score, the Random Forest Regressor outperformed other models. Post-hyperparameter optimization, the Random Forest Regressor's MAE marginally decreased. Furthermore, the Decision Tree Regressor also demonstrated improvement, underscoring the significance of hyperparameter tuning in enhancing model accuracy.

The Decision Tree Regressor and Random Forest Regressor algorithms are significant for predicting solar output due to several key reasons. They are better equipped for handling nonlinear relationships. Solar energy generation is influenced by various nonlinear factors such as weather conditions, time of day, and seasonal variations. Decision trees, as well as random forests composed of multiple decision trees, are capable of capturing nonlinear relationships in the data, making them well-suited for modeling the complex interactions involved in solar energy production. Additionally, these algorithms can provide insights into the importance of different features (e.g., daily yield, ambient temperature, irradiation) in predicting solar output. By examining feature importance scores, analysts can identify which factors have the greatest impact on solar energy generation, aiding in system optimization and resource allocation. Moreover, decision trees and random forests are inherently robust to outliers and noise in the data. This robustness allows them to handle irregularities or anomalies in solar generation patterns, ensuring reliable predictions even in the presence of unexpected events or data discrepancies.

Decision trees also offer interpretability, as the decision-making process is represented in a tree-like structure that can be easily visualized and understood. This interpretability is valuable for stakeholders seeking to comprehend the factors influencing solar output and make informed decisions based on model insights. This ensemble based method, composed of multiple decision trees, combines the predictions of individual trees to improve overall performance and generalization. By leveraging the diversity leverage of constituent trees, random forest algorithms can mitigate overfitting and enhance predictive accuracy, making them particularly effective for solar output prediction tasks. With the growing availability of data from solar energy installations and weather monitoring stations, the scalability of Decision Tree Regressor and Random Forest Regressor enables the analysis of extensive datasets to extract valuable insights and optimize solar energy systems.

This study demonstrates that Decision Tree Regressor and Random Forest Regressor algorithms are powerful tools for predicting solar output, thanks to their ability to model nonlinear relationships, handle outliers, provide interpretability, leverage ensemble learning, and scale to accommodate large datasets. These attributes make them invaluable assets in the pursuit of efficient and reliable solar energy forecasting and optimization. Future research endeavors should focus on gathering more diverse datasets, exploring the integration of real-time data, or incorporating other hyperparameters that could further refine solar output predictions and propel advancements in renewable energy technologies with these algorithms. Overall, this study highlights the potential of machine learning in revolutionizing solar energy forecasting and facilitating the transition to a sustainable energy future using Decision Tree Regressor and Random Forest Regressor algorithms.

As shown in this study, the integration of diverse inputs in AI algorithms used to predict the output of solar cells is essential for improving accuracy, adaptability, robustness, and optimization of energy harvesting processes. These factors enable the model to develop a comprehensive understanding of solar cell performance and provide reliable predictions under varying environmental conditions, ultimately facilitating efficient utilization of solar energy resources.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Ashok, P. (2024). What is ridge regression?. Great Learning Team. https://www.mygreatlearning.com/blog/what-is-ridge-regression/

Beheshti, N. (2022). Random forest regression: A basic explanation and use case in 7 minutes. Towards Data Science. https://towardsdatascience.com/random-forest-classification-678e551462f5

Buitinck, L. et al. (2013). API design for machine learning software: Experiences from the scikit-learn project.

Chugh, A. (2020). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared: Which metric is better?. Medium. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Jain, A. (2023). Ridge and lasso regression in Python.

Kalogirou, S., & Senc, A. (2010). Artificial intelligence techniques in solar energy applications. Sciyo.

Kanade, V. (2022). What is linear regression? Types, equation, examples, and best practices for 2022. AI Researcher.

Kannal, A. (2020). Solar power generation data: Solar power generation and sensor data for two power plants.

Kaggle. https://www.kaggle.com/datasets/anikannal/solar-power-generation-data

Solar Energy Industries Association. (2021). 30% by 2030: A new target for the solar+ decade. www.seia.org/research-resources/30-2030-new-target-solar-decade

U.S. Energy Information Administration. (9024). Frequently asked questions (FAQs). www.eia.gov/tools/faqs/faq.php?id=427&t=3

Verma, N. (2023). Simple decision tree regression in Python with scikit-learn.

Yung Yap, K. et al. (2020). Artificial intelligence based MPPT techniques for solar power system: A review. Journal of Modern Power Systems and Clean Energy, 8(6), 1043–1059.