# Smart Solar Scheduling: Machine Learning for efficient Solar powered Electric Vehicle charging and Home energy use

DATA245 Machine Learning Technology

Riddhi Kamleshkumar Vyas  Raveena Kagne  Priya Govindarajulu  Soujanya Sankathala  Tanvi Pagrut

***Abstract*—Our project, "Smart Solar Scheduling: Machine Learning for Efficient Solar-Powered EV Charging and Home Energy Use," aims to optimize the use of solar energy for electric vehicle (EV) charging and home energy consumption using machine learning techniques. This involves a comprehensive process that begins with the collection and integration of weather data and solar energy production and consumption data from home solar panels and electric vehicle charging. The research focuses on data cleaning, preprocessing, and exploratory data analysis using Python to prepare the data for model development. The project employed a variety of machine learning models, including linear regression, decision trees, random forests, support vector machines, and gradient boosting machines, to identify the most effective method for predicting solar availability and planning energy usage. The project includes performance evaluation metrics to select the optimal model. This best model is used for the development of a user-friendly Tableau dashboard that helps homeowners make informed decisions about the best times to charge Electric Vehicles and use energy based on solar generation forecasts. The project aims to enhance the sustainability and efficiency of solar energy in homes**

***Index Terms*—linear regression, decision trees, random forests, support vector machines, gradient boosting machines,RMSE, $R^2$,MAE**

## I. Introduction

In an era of increasing environmental awareness and a significant push towards renewable energy sources, finding methods to increase the use of solar power emerges as a critical challenge. The transition to green energy, particularly solar power, presents a unique opportunity to revolutionize how we consume energy in our daily lives, from powering homes to charging electric vehicles (EVs). Traditional methods of solar energy utilization often rely on static scheduling, where energy usage and EV charging are planned based on average production rates rather than actual solar output. This approach may result in inefficiencies such as unnecessary reliance on the grid during off-peak hours and underuse of solar energy during peak production hours. Because solar energy production is dynamic and subject to weather and seasonal variations, optimizing its potential requires a more advanced strategy. The idea is to bridge the energy gap between producing and using solar energy by using machine learning to create a model that can reliably forecast solar energy production based on several parameters such as weather patterns, historical and real-time solar data, and the actual performance of solar panels. These forecasts can then be used to create smart scheduling systems that adapt to changing conditions, ensuring that solar energy is used optimally throughout the day.This Smart Solar Scheduling is an innovative method for usage of power to promote sustainability. The developed machine learning model will also find the best suitable times to use solar energy that will help in decreasing the dependence on the electricity grid and reducing the energy bill costs. Smart Solar Scheduling is a cutting-edge approach to energy use that encourages sustainability. Enhancing the accessibility and efficiency of renewable energy, contributes to global efforts toward a sustainable and environmentally friendly future.

## II. MOTIVATION

This project was inspired by an experience one of our team members had last summer when they charged their electric vehicle using their home solar panels without knowing the optimal times for solar energy generation. They charged their EV between 9 am and 2 pm, which was the peak solar production hours. This incident underscores a growing environmental awareness and the urgent need to transition from fossil fuels to sustainable energy sources, such as solar power, to mitigate the effects of climate change. Also it highlighted a widespread issue among solar energy users: a lack of knowledge about the best times to use solar energy. By addressing this problem, we can significantly increase the use of renewable resources, reduce costs, and improve energy efficiency. It highlights the necessity for an advanced system that can predict solar energy peaks and recommend the best times to charge electric cars and control household energy consumption, enhancing overall energy efficiency and sustainability.

## III. PROBLEM DEFINITION

The project addresses the problem of inefficient utilization of solar energy for charging electric vehicles and powering homes due to a lack of predictive insights into solar production peaks. Due to inadequate prediction and utilization of solar energy based on weather and production data in real-time, current methods result in inefficient energy use and a greater reliance on non-renewable energy sources during periods of

low solar production. Inadequate scheduling leads to an over-reliance on the grid, causing unnecessary grid use even when there is adequate solar power available. This inefficiency not only leads to inappropriate use of grid electricity but also contributes to higher energy costs. Homes fail to use solar energy when it is most accessible, resulting in inefficient scheduling that drives up energy bills. This cycle of reliance and inefficiency underscores the need for better energy management and optimization strategies.

## IV. OBJECTIVE

The project's objective is to create machine learning models that can be used to predict solar energy production accurately based on various factors like weather conditions, historical solar data, and real time performance of solar panels by utilizing real-time data from household solar panels and available weather data. Then making informed decisions about energy consumption at homes and charging schedules for electric vehicles, aiming to reduce electricity grid dependence and energy bills.

## V. LITERATURE REVIEW

[1] Almughram, Abdullah ben Slama, and Zafar (2023) the researchers integrate both Vehicle-to-Home (V2H) units and an intelligent Home Energy Management System (HEMS) to achieve maximum generation optimization, particularly during peak hours of electricity consumption. They propose to use a reinforcement learning strategy between the solar photovoltaic (PV) system, grid storage and electrical vehicles (EVs) to optimize operations during the asynchronous nature of renewable energy output as well as the uncertainties in microgrids of the future. Using deep learning algorithms to forecast solar energy generation and V2H technology to immerge appliance load profiles, they would be trying to improve users power consumption costs and also encourage sustainable power production. This study illustrated that reinforcement learning could dynamically react to diffuse solar radiation fluctuations and efficiently control the demand-response using V2H technology that could function as smart building storage.

[2] Aguilar-Dominguez et al. (2021) designed a machine learning (ML) model for predicting electric vehicle (EV) availability for Vehicle-to-Home (V2H) services emphasizing the inadequate research on how vehicle availability affects V2H capacity Employing the car use patterns and optimization models, the research shows that EVs not always employed for travel are more available for V2H services and the has the biggest implications across all customer groups when implementing V2H. The study showed that the ML algorithms achieved the accuracy of over 85% for locating EVs and even further by 46% in reducing electricity cost. Besides that, it has demonstrated the efficiency of ML in providing V2H services. The results recommend further investigation on the topic of dynamic vehicle usage and its effect on power management as well as cost benefits.

[3] Forootan, M. M., Larki, I., Zahedi, R., and Ahmadi, A. (2022) categorize ML and DL techniques used for improving performance in diverse energy systems eacht game has different goals, rules, and types of moves. It is this security integration that will help individuals and small groups towards modern technology adoption. The review emphasizes the evolvement and importance of ML and DL in tackling the intricacies of the energy dealing through the advancing demand for energy and exigency for sustainable choices. Such areas being swinging of high consumers, forecasting of solar and wind power, implementation of optimization strategies and detection of faults in energy systems are crucial. The paper gives a good picture of how ML and DL technologies can be very helpful in prediction of energy, optimizing energy use, and in case of renewables, the reliability of energy sources. The authors support the model that survey other algorithms or hybrid approaches that may improve the accuracy of forecasting and to reduce the operational costs in energy systems. They indicate that such work is necessary for new technologies that would be needed to address current and future energy demands.

[4] In their study, Sharma, N., Sharma, P., Irwin, D., and Shenoy, P. (2011) Renewable problem was forecasting weather which is essential to integrate power coming from the Sun into the grid. Previous models had fully concentrated on manual development and conventional metrics like sky condition; besides they had become obsolete and unable to be adapted to distributed generation across different plants. Authors achieved it by implementing machine learning methods encompassing linear regression and a collection of kernels for SVM. These techniques were informative enough for weather forecasts and gave rise to site-specific prediction models establishing relationships between a set of weather forecast parameters. They did so by leading to realization of 27% accuracy improvement as against existing models; hence, the machine learning approach proved to be a very helpful technique for more accurate solar generation predictions. To be continued, the study recommends next steps such as using advanced machine learning models and their capability in adapting themselves with the changing weather patterns and site-specific conditions, ultimately leading to development of a more cost effective and dependable solar energy technology integrated in the smart grid.

[5] Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., and Fouilloy, A. (2017) articulated an exhaustive survey on machine learning methods for the solar radiation transmission in the Renewable Energy magazine, emphasizing the contribution of appropriate prediction models to the successful incorporation of renewable energy to grids. The paper investigates some machine learning examples, such as artificial neural networks, support vector machines, and ensembles and it scrutinizes their potential to work with the irregular factor of the solar radiation, during the forecasting. According to the authors, new study lines are needed, including the development of fast and adaptive real time forecasting systems enabling more precise, reliable, and flexible solar irradiance forecasts be produced.

## VI. HYBRID CRISP-DM

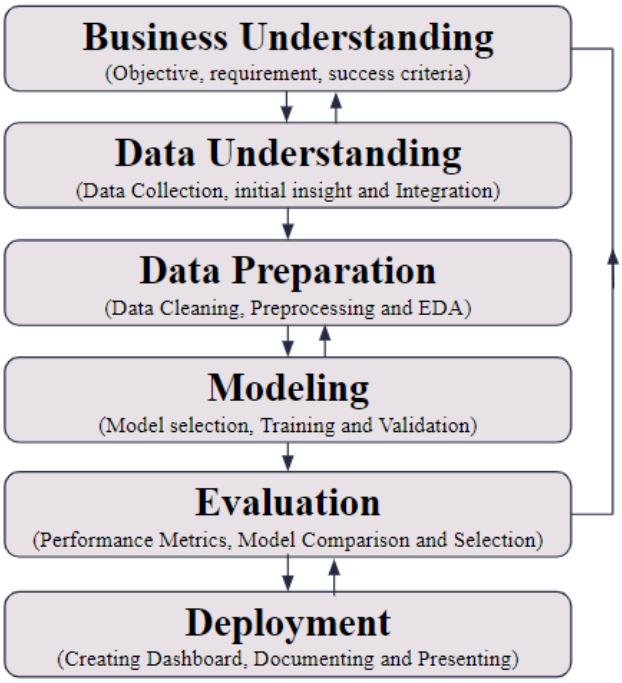


Fig. 1. Hybrid crisp-dm

### A. Business Understanding

This phase involves determining the project goals and specifications. For the project, the goal is to develop machine learning algorithms to forecast solar energy production and optimize the use of electric vehicles and residential energy, improving the effectiveness and availability of renewable energy.

### B. Data Understanding

This is done by gathering and properly integrating the relevant data. The data comes from solar panels at homes, electric vehicle charging stations, and weather variables such as temperature, humidity, solar radiation, etc. Understanding these variables helps analyze their effects on energy generation.

### C. Data Preparation

This phase involves cleaning and preprocessing the data. We have merged weather and solar generation datasets, dealt with null values, cleaned up duplicates, and applied transformations that involve one-hot encoding. This facilitates accurate modeling.

### D. Modeling

This is the time when some models are selected and trained. This project involves Linear Regression, Decision Tree, Random Forest, SVM and Gradient Boosting Machines to be used to cover different aspects of the data and to ultimately increase the accuracy of the prediction.

### E. Evaluation

After modeling, performance of these models is evaluated using metrics like RMSE, $R^2$, and MAE. Model evaluation is a very useful tool for comparing different models and choosing the most suitable for deployment. For example, the Random Forest model showed the high accuracy and good balance of training and validation performance.

### F. Deployment

The last step creating the dashboards for visualizing predictions and writing a project outcome report for stakeholders as well.

The use of a Bloom filter and differential privacy serves two key purposes: ensuring the uniqueness of data and protecting the privacy of individual transactions. These techniques are crucial for handling sensitive data in real-time streaming applications, particularly in fields like finance where data security and privacy are paramount.

## VII. DATA PROCESS FLOW

The project's data process flow is meticulously designed to ensure the effective utilization of solar energy for electric vehicle charging and home energy use. The process starts with data collection, which is the step of collecting real-time and historic data from home solar panels and electric vehicle charging stations, along with comprehensive weather data, including temperature, humidity, solar radiation, and other relevant variables. Not only data is systematically being integrated but also the times being compared to make sure each record of solar energy production is correlated with respective weather conditions. Following the gathering of data, it goes through the way of cleaning and preprocessing, concentrating on the data quality and compatibility and making it suitable for modeling, such as missing values handling, reducing the duplicates, and transforming variables.

The following stage of the process is devoted to exploratory data analysis (EDA), where the processed data is analyzed to find out the hidden patterns, correlations, and distributions. Such comprehensive analysis paves the way for identifying fundamental parameters that determine solar energy framework, in turn providing indispensable inputs for the devised model. Machine learning models, for example, linear regression, decision trees, random forests, support vector machines, and gradient-boosted machines are trained using the labeled data. Each model is judged on the basis of performance indicators such as RMSE, $R^2$, and MAE to find out which model is the most accurate and reliable for deployment. The realized model after careful development is then embedded into a Tableau dashboard, which is capable of providing real-time predictive insights and actionable recommendations to the consumers for their better energy usage.

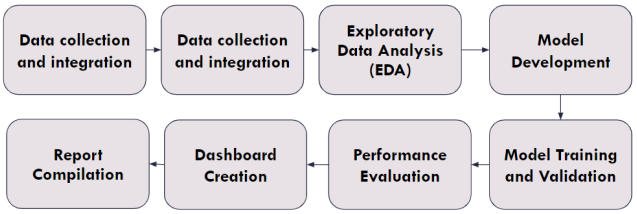


Fig. 2. Data flow process

## VIII. DATASET DESCRIPTION AND COLLECTION METHODS

The data from a solar panel in a single household is able to give detailed real-time insights into the production of solar power for home electricity usage and also electric vehicles(EV) charging records. The solar installation which is in Manteca, CA is registered with SolarPATH for hourly energy production in between August 2023 to March 2024. This fine-grained data includes solar production (kWh), home energy usage (kWh), and net grid energy exchange (kWh). Additionally, the system keeps a cumulative value of solar production, home usage, and net grid energy for longitudinal analysis and performance assessment decision-making. Tracking of weather data along with the solar data helps to establish the weather conditions that highly impact the efficiency of solar panels. The weather data for Manteca is collected by the weather data service (Visual Crossing) and it covers the same period as the solar data collection. The crucial parameters on the list of the weather data are temperature, humidity, wind speed, platform cover as well as irradiation, which are all factors widely used for correlating weather conditions with solar energy production levels.

Both the solar generation and weather datasets are timestamped, allowing for precise synchronization. There results in an all-inclusive dataset after every record of solar generation and consumption is synchronized with the corresponding weather conditions history of the hour. This dataset, which is the result of the integration of multiple data sources, is the foundation for the following analysis, allowing for a multidimensional view of how environmental factors influence solar power production.

## IX. DATASET DESCRIPTION AND COLLECTION METHODS

After the integration of the dataset, it is checked for its completeness and accuracy. Initial steps include checking for missing values and outliers. Missing values will definitely skew the analysis and results, while outliers might cause the distortion of the results. This will maintain data quality, and consequently, the dataset will be more reliable in successive steps and analysis.

To boost the ability of the dataset to predict the solar output for the purposes of optimal EV charging, new features are designed. Factors such as the time of the day, the day of the year, and the historical solar production averages are considered. For categorical variables, appropriate encoding techniques are applied, transforming them into numerical formats that are accepted by the machine learning algorithms.

Next, different data normalization and standardization techniques are employed, given the variety of data types and scales. This ensures that each variable has equal contribution in the analysis, none is more important than the others. Tools like Pandas and Scikit-learn are used in completing and handling these transformations.

The cleaning and preprocessing stages utilize Python for scripting, with libraries such as Pandas for data manipulation and Scikit-learn for applying machine learning techniques like encoding and scaling. These tools are essential for handling large datasets and preparing them for complex analytical tasks efficiently.

The cleaning and preprocessing stages use Python for scripting, data processing libraries such as Pandas for data manipulation and Scikit-learn for applying machine learning techniques e.g. encoding and scaling the data. Such tools are the ones that are the key to handling large datasets and efficiently preparing them for complex analytical tasks.

In summary, this detailed approach of data collection, integration, cleaning, and preprocessing, provides the data with the structuring that is necessary for further analysis. Moreover, it ensures the correctness and reliability of the conclusions obtained from the data. The procedures stated above serve as a platform for the employment of effective analytic models that are responsible for forecasting the solar energy production of households, as well as guaranteeing energy efficiency.

## X. EXPLORATORY DATA ANALYSIS

All visualizations were generated exclusively through Python code. The following charts and their respective descriptions are provided

**EDA 1:** This histogram visualizes the distribution of the target variable “solar production”, measured in kilowatt-hours (kWh). The x-axis represents the range of solar production values and the y-axis shows the frequency or count of observations within each range bin.

By looking at above plot, the shape of this histogram can provide insights into the central tendency and spread of solar production data from the household’s solar panel system. A normal bell-shaped distribution indicates that most observations fall around the mean value, with fewer instances of very low or very high solar production levels.

However, such distribution might be a bit skewed to one side, which means that the observations will probably be gathered towards the lower or higher side of the production range. The interruptions may be a result of seasonal changes, precise weather patterns, or the orientation and solar power system installations efficiency.

**Key Findings**

- The temperature is normally distributed, with most values between 60 and 80 degrees Fahrenheit.
- Here, we can see that more data is at 0 for Solar Production (kWh). This is because, during the night, there

is no solar production, and that's why it has a high bar of 0.

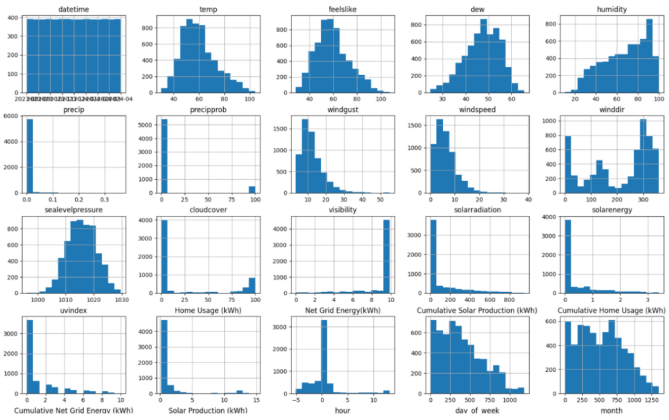


Fig. 3. Histogram for showing the spread of the data

**EDA 2:** The heatmap represents the correlations among the solar production and climate parameters, e.g. temperature, well and wind speed, cloud cover, and solar radiation. The color scale represents the strength and direction of the correlations, with darker shades of red indicating a strong positive correlation and darker shades of blue indicating a strong negative correlation. By examining this heatmap, we can identify which weather variables have the strongest relationships with solar production. For example, we might expect solar radiation to have a strong positive correlation, as higher levels of solar radiation often lead to increased solar energy generation. Conversely, cloud cover may have a negative correlation, as clouds can obstruct sunlight from reaching the solar panels.

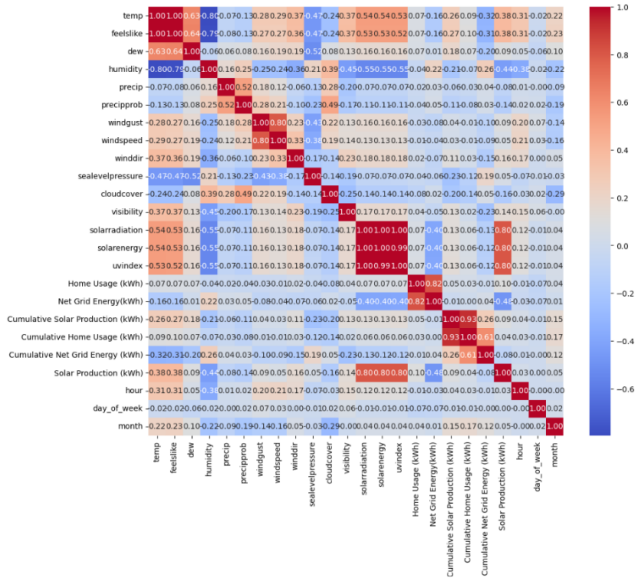


Fig. 4. Heatmap for Correlation Analysis for Numeric Variables

## Key Findings

- There is a strong positive correlation between solar radiation and solar energy. This indicates that the amount of solar energy produced is directly related to the amount of solar radiation.
- There is a strong negative correlation between wind speed and wind direction, which indicates that wind speed is inversely related to wind direction.
  Understanding these correlations helps in feature selection and prioritize the most influential variables for building accurate predictive models. Also, it provides insights into potential weather conditions that may impact solar production, informing energy usage strategies and electricity grid reliance.

**EDA 3:**The pair plot creates a matrix of scatter plots, visualizing the relationships between solar production and various weather features, as well as the relationships between the weather features themselves. Each scatter plot represents the pairwise correlation between two variables.

By analyzing these scatter plots, we can identify potential linear or non-linear relationships, as well as any outliers or clusters within the data. We might observe a strong positive linear relationship between solar production and solar radiation, or a negative non-linear relationship between solar production and cloud cover.

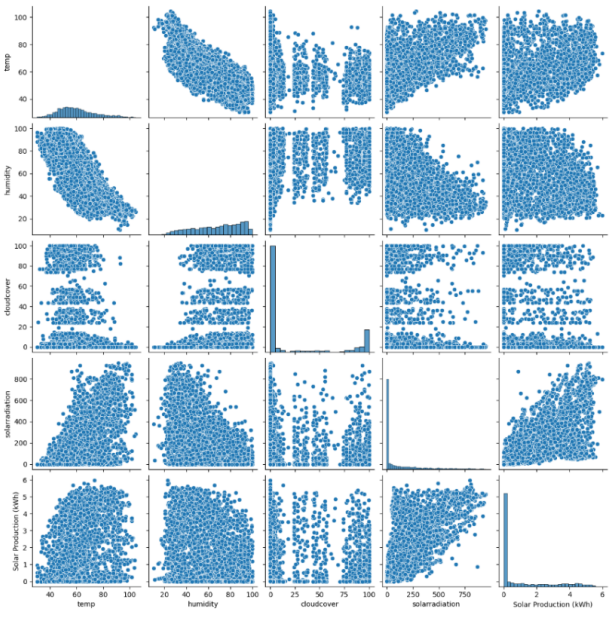


Fig. 5. Pair Plot for Variable Relationships

## Key Findings

- Pair plots are used to show the distribution of two variables (bivariate distribution).
- First, there is a strong positive relationship between temperature and solar radiation. This means that as the temperature increases, the amount of solar radiation also increases. This is likely because higher temperatures are

associated with clearer skies, which allow more sunlight to reach the Earth's surface.

- There is a weak negative relationship between cloud cover and solar radiation. This means that as cloud cover increases, the amount of solar radiation decreases. Clouds can block sunlight from reaching the Earth's surface.

  These insights inform feature engineering techniques, such as transformations to improve the predictive power of the machine learning models. Additionally, the pair plot can reveal patterns or structures that may guide the selection of appropriate model types (e.g., linear models, decision trees, or neural networks).

**EDA 4:** The boxplot is a valuable tool for visualizing the distribution of solar production and identifying potential outliers. The box represents the interquartile range (IQR), containing 50 percent of the data, while the horizontal line inside the box marks the median value. The whiskers extending from the box indicate the range of values that fall within 1.5 times the IQR, and any data points beyond the whiskers are considered outliers, represented as individual dots or markers.

Outliers in solar production data could arise due to equipment issues, shade, or extreme weather events. Identifying and understanding these outliers is crucial, as they can significantly impact model performance and accuracy.

**Key Findings**

- In this case, the median solar production value is 2 kWh. The box extends from 0 to 4 kWh, indicating that most solar production values fall within this range. The whiskers extend to 6 kWh, indicating that there are a few outliers with high solar production values.

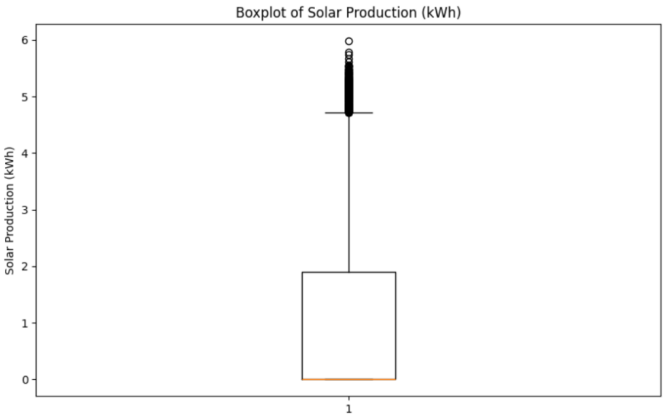


Fig. 6. Boxplot for Outlier Detection

**EDA 5:** This analysis explores the patterns and trends in solar production over different time periods, such as hours of the day, days of the week, or months of the year. Visualizations like line plots, bar charts, or heatmaps can be used to illustrate these temporal variations.

By analyzing these visualizations, we can identify peak periods of solar production, as well as periods of low or no production. For example, the situation could look different during the day when the sun is at its highest point and we might observe higher solar production during such midday hours. On the other hand, we might observe lower production during the winter months with shorter days and possibly more cloudy conditions.

Understanding these patterns and trends is crucial for optimizing energy usage strategies and scheduling activities that rely on solar power, such as electric vehicle charging or running household appliances. Additionally, this analysis can inform decisions related to energy storage solutions or reliance on the electricity grid during periods of low solar production.

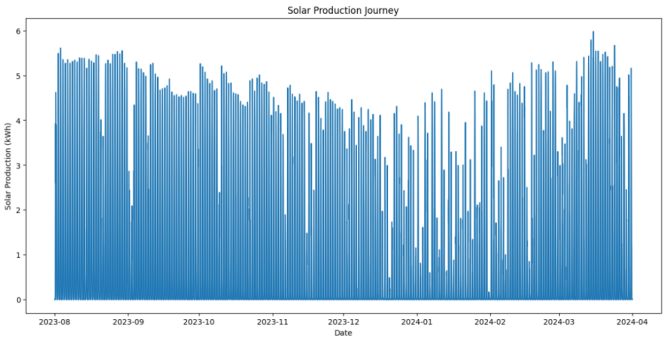


Fig. 7. Solar Production Patterns and Trends

**EDA 6:** Building upon the previous analysis, this visualization specifically focuses on identifying the optimal hours of the day for high solar production. A common approach is to create a bar chart or line plot showing the average solar production for each hour of the day.

By examining this chart, we can pinpoint the hours when solar production is consistently high, allowing us to target those periods for energy-intensive activities or charging electric vehicles. Conversely, we can also identify the hours with low or no solar production, during which it may be more efficient to draw power from the grid or alternative sources.

This analysis can be further enhanced by incorporating weather factors or seasonal variations, as the optimal hours for solar production may shift depending on these variables. For example, the peak production hours during summer months may differ from those in winter due to changes in daylight hours and cloud cover patterns.

By understanding the optimal hours for solar production, homeowners can make informed decisions about when to maximize their use of solar energy, potentially reducing their reliance on the electricity grid and lowering overall energy costs.

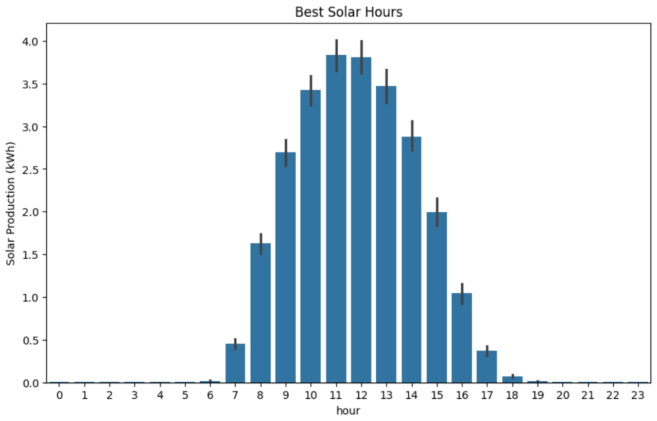


Fig. 8. Optimal Hours for Solar Production

## XI. DATA PREPARATION FOR MACHINE LEARNING MODELS

The correlation matrix shown in Figure 4 provides insights into the relationship between different variables in the dataset. By analyzing this correlation matrix, there is a high correlation between the temperature and feelslike features, which are close to 1. Also, there is a high correlation between solar radiation and solar energy, which is close to 1. The 'feelslike' feature is removed due to its high correlation with temperature, and the solar energy feature is removed due to its high correlation with solar radiation. Based on this analysis and domain knowledge, the features with high multicollinearity are removed. From the original processed dataset of 26 features, the final 16 features were selected with the target variable "Solar Production."

The selected features and target variables are defined for the machine learning models. The 'MinMaxScaler' is used to scale features to ensure all the features are within the same range. This normalization of the features helps to improve the performance of machine learning algorithms. The dataset is first split into 80% for training and validation and 20% for testing. In 80% of the dataset, 60% is used to train the machine learning models and 20% for validation. After the dataset split, the models are trained using the training data, and the validation dataset explains the model training and validates the predictions for overfitting. Then, the models are tested using the unseen test dataset to evaluate the model's accuracy and performance.

## XII. MODEL DEVELOPMENT

For our solar prediction project, the selection of the right machine learning models is the key to the accurate forecasting of solar energy production based on factors like solar radiation, temperature, and humidity. We have chosen five Machine Learning models, namely Linear Regression, Decision Tree, Random Forest, SVM (Support Vector Machines), and Gradient Boosting Machines to achieve this. The solar production target variable is the continuous variable, and the problem of predicting solar production based on weather and other environmental factors is a regression problem.

### A. Modeling with Linear Regression

Linear Regression is used as a base model for this project. It creates a linear connection between the input variables (e.g., temperature and humidity are the input variables) and the output variable (solar production). It enables us to understand the immediate effect of each predictor on the result, which is the main reason for knowing how different environmental conditions affect the production of solar energy. The linear regression aims to model the relationship between the dependent variable, solar production, and the other independent variables, which are the remaining selected features. This is done by fitting the linear regression line for the selected data. The relationship between the variables is explained using the below formula:

The linear regression equation is given by:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

The $y$ indicates the target variable, and the $x$ indicates the independent variables. The $\beta_0$ is the intercept, and $\beta_1$ to $\beta_n$ are the weights (coefficients) of the selected features. The linear regression assumes the relationship between the target and selected features is linear. The goal of this model is to find the coefficients that help to reduce the difference between actual and predicted target values. The cost function, mean squared error, in this model tells how well the model predicts to match the actual data. The cost function is reduced by using gradient descent in this model, and this process is repeated until all the coefficient values are found. The training and validation were completed using the split dataset. It uses the learned coefficients to make predictions when new data arrives.

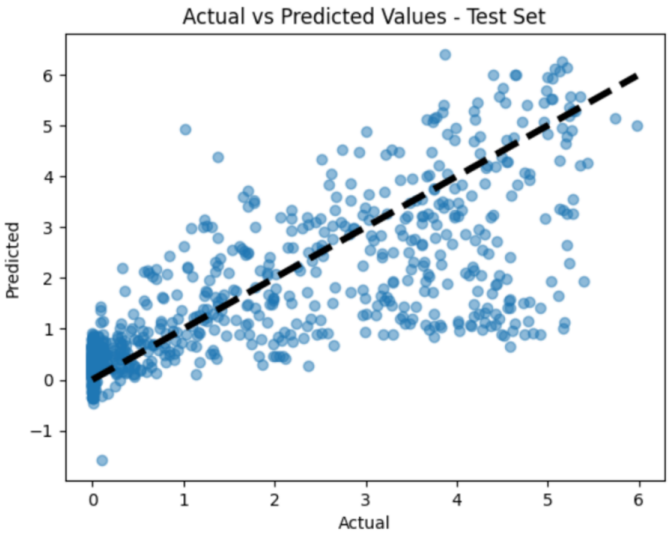


Fig. 9. Actual vs Predicted values by Linear Regression model

The scatter plot figure shows the relationship between predicted and actual values by the linear regression model. The X-axis represents the actual values, the y-axis represents the predicted values, and the black dashed line represents the linear equation. If the model's predictions are accurate, all the

points will lie on the linear regression line. The model performs better for small solar production values and it is shown by the points at the lower values that align on the linear line. However, the points away from the linear regression indicate that the model's predictions for those values are significantly off. The R-squared value 0.66 on the test dataset by linear regression model indicates that the 66% variance of solar production is explained by the model. Based on the MAE, the model predictions are 0.63 kWh away from the actual data, and based on the RMSE, the predicted values by this model are 0.96 kWh from the actual values. The model performs reasonably well in capturing trends but needs improvement in reducing prediction errors. Additional techniques like feature engineering might be employed to improve the accuracy of the model.

### *B. Modeling with Decision Tree*

This model is characterized by a tree-like graph of the decisions and their possible outcomes. It divides the dataset into branches so that the predictions can be made, which can process both linear and non-linear relationships. Decision trees are effective in dealing with the complicated, hierarchical decisions that are related to solar production, for example, the impact of weather conditions interactions. Moreover, they are also simple to understand and can take various data formats. The decision tree is a non-linear model used for regression tasks. The dataset is split, and this model forms a tree-like structure. The internal node represents the decision on the feature, the branch represents the outcome of that decision, and each leaf node is the predicted value. The figure explains the decision tree created for the solar production dataset, and the node has values that represent the features, mean squared error, which measures the variance achieved by the split, and samples explain the number of samples present in the training set. The value in the node is the predicted solar production value by that node.

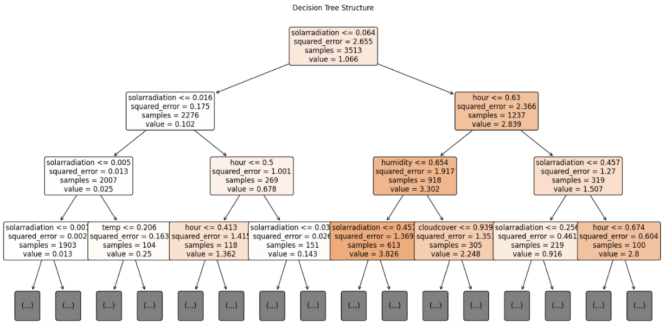


Fig. 10. Decision Tree

To make a prediction, the tree traverses from the root node to the leaf node. The predicted value is the mean value of the target variable in the leaf node. The model used learned splits and rules for taking decisions for the predictions of the new data that arrives. The R-squared value of 0.87 shows that the model can explain the 87% variance in the solar production data. The RMSE for this model is 0.61, indicating the predicted values are 0.61 kWH away from the actuarial solar production values. However, this model has a low mean squared error of 0.26.

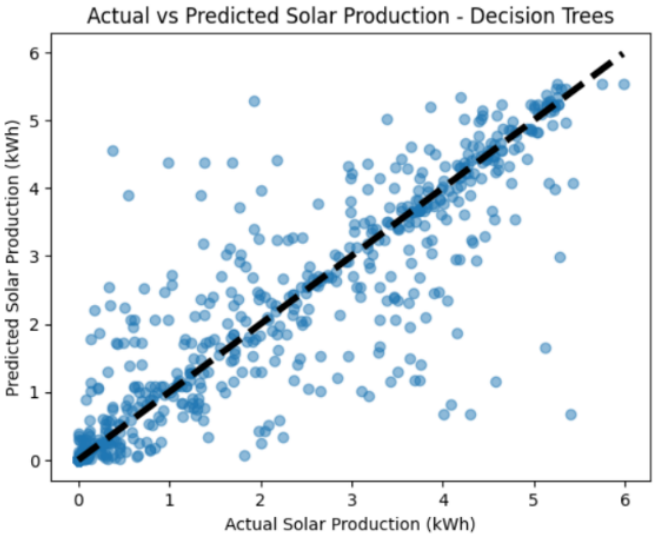


Fig. 11. Actual vs Predicted values by Decision Tree

The scatter plot for actual values vs predicted values by the decision tree model is shown in the figure. Most of the data points lie close to the diagonal line, indicating the model's predictions are accurate for these observations. It has fewer outliers compared to the linear regression model, which suggests outliers, and the model's predictions can be improved for these data points. These outliers suggest that complex model techniques can be used to improve the accuracy.

### *C. Modeling of Support Vector Machine*

The SVM is chiefly a classifier that aims to discover the hyperplane that best separates a dataset into classes. Besides being used for classification, SVM can also be applied to regression (SVR) to predict continuous values.SVM is considered a working tool in areas with high-dimensional spaces and with non-linear boundaries by the use of the kernel functions. Hence, it is the perfect tool for the forecasting of solar output when the relationship between the input features and the output is not linear. The support vector machines can be used for both classification and regression tasks. In this project, Support vector regression SVR is used to predict solar energy production. The SVR aims to find a function with a margin of tolerance specified that approximates the data.

This model's goal is to find the best-fit line that can stay within the margin of tolerance. This function that SVR tries to reduce is the objective function, which has two components: the regularization term and the loss function. The regularization term helps the model to remain the same, simple and helps to avoid overfitting. This is achieved by penalizing large weights. The expression $\frac{1}{2}\|w\|^2$ represents a part of the objective function in support vector regression. The loss

function is the epsilon-insensitive loss function, which allows a margin within which the errors are ignored, as mentioned in the formula below:

$$C \sum_{i=1}^{n} \max(0, |y_i - (w \cdot x_i + b)| - \epsilon)$$

Here, $y_i$ is the actual value, $(w \cdot x_i + b)$ is the predicted value, and $\epsilon$ is the epsilon margin. of tolerance. By combining both of these components, SVR tries to minimize the function. The R-squared value of 0.77 indicates that 77% variance is explained by the model, and this model has a high RMSE of 0.80 and MAE of 0.40. The scatter plot in the below figure shows the relationship between the actual and predicted solar production values by this SVM model. The model's predictions are accurate for smaller values. For the larger values, there is a noticeable spread around the diagonal line, which indicates model predictions need to be improved.

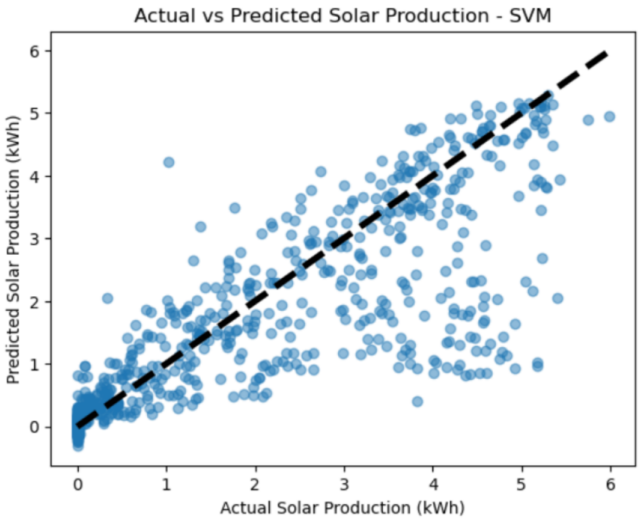


Fig. 12. Actual vs Predicted values by SVM

This model performs well for small and medium values, and improvement is needed for higher value prediction. Employing ensemble methods could improve the accuracy and performance of this solar production dataset.

### *D. Modeling using Gradient Boosting Machines*

A very strong boosting ensemble method that builds models in stages just as other boosting methods do and generalizes them by enabling the optimization of a given arbitrary differentiable loss function.GBM is able to deal with all types of predictive modeling, including the robustness to outliers in output space (solar production). It is the fundamental one that corrects the mistakes of the previous models and increases predictability, which is especially useful in forecasting tasks where precision is of the essence.GBM is particularly effective for regression tasks and can handle various types of predictive modeling. GBM model structure is built on an ensemble of trees, which is in sequential order. Each tree predicts the error from the previous tree and predictions are combined for the final model. This model aims to reduce the difference between the predicted and actual values and reduce overall error. The model is initialized with parameters of the maximum depth as 5, 250 n_estimators to specify the number of boosting stages and with a learning rate of 0.05. This model's R squared values indicate this explains 93% of the variance in the data, and this shows the model has a high ability to detect the underlying patterns in the data. This model has an RMSE of 0.42 and MAE of 0.21.

The learning curve of Gradient Boosting learning curve shows the relationship between training samples and the model's performance in the figure. The model benefits from the training data more and cross-validation scores are improved with decreasing variance. This model outperforms all the above three models with this ensemble technique.

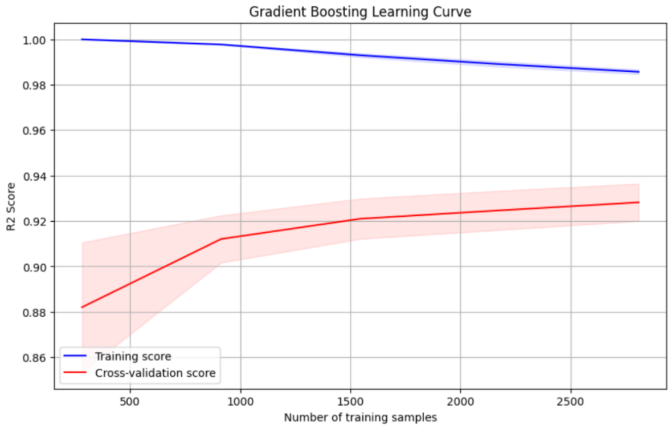


Fig. 13. Gradient Boosting learning curve

### *E. Modeling with Random Forest*

A method that is an ensemble and works by creating many decision trees while training. It enhances the model accuracy by averaging or 'bagging' many trees to decrease overfitting and variance of the predictions. Through the complexity of the environmental data and its influence on the solar output, Random Forest is able to capture this complexity more effectively than a single decision tree. It is resistant to overfitting and is perfect for large datasets with high dimensionality. For regression problems like this project, each tree makes a prediction and all the results are combined finally from all the trees in the model. For each node in the features, the features are chosen randomly. This combination of multiple trees smooths the prediction and reduces the variance. The overfitting is reduced in this model since it uses an average of multiple trees, which is explained in the learning curve figure for the random forest model. Analyzing how the model performs when data points increase and this cross-validation score keeping increasing denotes that the model performs well for unseen data points. The training curve and cross-validation score stabilized around 4000 data points, and a smaller gap between curves indicates no model overfitting.

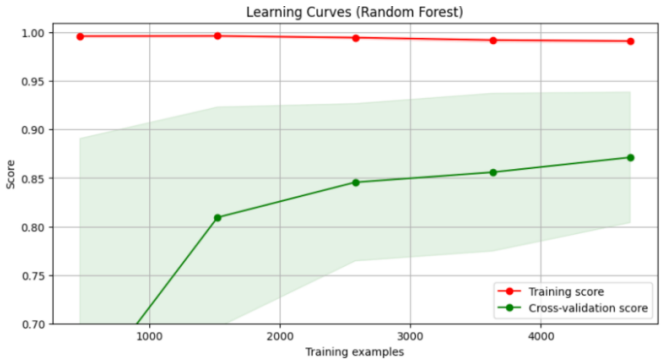


Fig. 14. Random Forest learning curve

The random forest model has an R squared value of 0.93 with RMSE 0.40 and MAE of 0.19 in the validation set indicating that this model outperforms all the other models with lowest error and high R-squared value. The scatter plot figure shown below indicates how well the data points align on the line or closer to the line indicating higher prediction rate with less outliers.

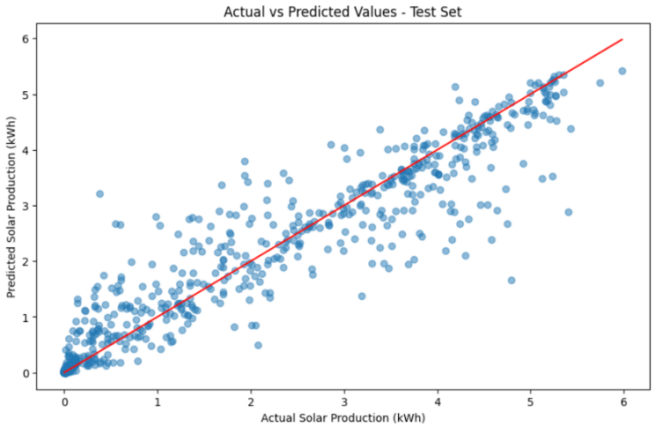


Fig. 15. Actual vs Predicted values by Random Forest

The K-fold cross-validation analysis was performed on the Random forest model to check how the model behaves for different data splits. The 5 folds are used to check the performance of the model and it is shown in the figure. The heatmap shows the different performance metrics across the folds, and fold 3 outperformed other folds with the highest R squared value of 0.94. The random forest model has an average Rsqaured of 0.93 with an MAE of 0.18. This analysis shows that this model is highly robust and effective for our solar prediction task.

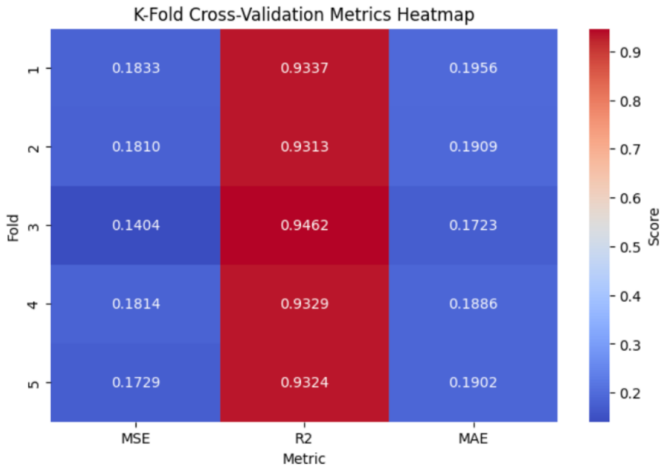


Fig. 16. K-Fold Cross Validation Metrics Heatmap

## XIII. Results

The performance of various machine learning models in terms of their explanatory power and prediction accuracy on a given dataset. Models set for evaluation include Linear Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and Decision Trees.

The linear regression model has the lowest $R^2$ value of 0.68, indicative of the fact that it is just able to explain about 68% of the variance in the data sets. On the other hand, it also has relatively higher error rates RMSE of 0.9 and MAE of 0.59 compared to the other models. In contrast, both Random Forest and Gradient Boosting models give the most accurate representation with $R^2$ values at 0.93, which means that they predict the variability within the dataset with 93% accuracy, thus implying a very strong predictive power. Among these, the Random Forest model has a slightly lower RMSE of 0.4 and MAE of 0.19, which insinuates that it usually forecasts more correctly and with lesser deviation from the exact values relative to Gradient Boosting where RMSE is 0.42 and MAE is 0.21.

Besides, the performance metrics for Support Vector Machine (SVM) and Decision Trees also show a high level of predictive ability. According to the SVM, an $R^2$ of 0.78, RMSE of 0.74, and MAE of 0.37, on the other hand, Decision Trees do better when the $R^2$ is 0.84, RMSE of 0.62, and MAE of 0.26. Thus, these results show that Decision Trees explain better and SVM makes faultier predictions.

In conclusion, the Random Forest model is the most reliable in terms of accuracy and consistency in error metrics, just after Gradient Boosting. This comparison is, therefore, necessary to find the model that will most appropriately solve the problem for a given data set between the power of explanation and predictive accuracy.

| Model | R-Squared (R^2) | Root mean squared error(RMSE) | Mean Absolute Error (MAE) |
|---|---|---|---|
| Linear Regression | 0.68 | 0.90 | 0.59 |
| Support Vector Machine | 0.78 | 0.74 | 0.37 |
| Decision Trees | 0.84 | 0.62 | 0.26 |
| Random Forest | 0.93 | 0.40 | 0.19 |
| Gradient Boosting | 0.93 | 0.42 | 0.21 |

Fig. 17. Performance Evaluation

- Solar Radiation: The bubble chart shows the solar radiation values, each bubble being a certain value. The bigger the bubble, the higher the solar radiation level is. The insights mean to find the periods that have more solar radiation, which are very important for the solar energy production to be at its peak.
- Temperature: The bubble chart displays the temperature values. Each of the bubbles' size is an indicator of the temperature magnitude. The higher the temperature readings are, the more solar production there is because more sunlight usually means higher temperature.
- Humidity: The bubble chart shows humidity levels, which are represented by the larger bubbles, thus, the higher the humidity, the bigger the bubbles are. Humidity can affect solar production, thus, it is important to keep an eye on these levels together with solar production
- Actual Solar Production By Month: This bar chart shows the solar production for different months. The chart reveals the seasonal patterns and trends of solar energy generation, thus, it displays the months that have higher or lower solar production.
- EV Charging Recommendation by Predicted Solar Production: The pie chart segregates days as either "Put on the Recommended list" or "Not Recommended" for EV charging according to the predicted solar production. This recommendation system assists in the determination of the EV charging schedules which will be in correspondence with the high solar energy availability
- Actual Vs Predicted Solar Production by Day: This line chart depicts the real solar production as compared to the forecasted values on a daily basis. It reveals the reliability of the prediction model and the major differences, which can be for the improvement of the model and the planning.
- Filters Panel: Filters enable the users to pick the appropriate dates, times, and other parameters to tighten the results in the dashboard. The characteristic of this feature is very important for the detailed examination and to concentrate on certain periods or conditions.
- Summary Metrics: The side panel shows the average temperature, the solar production and the humidity. These metrics of the summaries give a brief idea of the environmental situation and the solar production performance in the selected period.

Currently, the dashboard is designed for the historical data for demonstration purposes, but it is planned to be connected with the real-time data streams in the future. The automatic updating of the dashboard will make the recommendations more accurate and up to date. Future work on the project will involve adding real-time weather data, the improvement of the predictive model's accuracy and the adjustment of the dashboard's user interface to better support the real-time decisions in solar energy management.

## XIV. Dashboard Creation

We have developed a prototype of a near real-time dashboard which is going to be of use for the betterment of solar power production through the predictions and recommendations for electric vehicle (EV) charging using a predictive model based on a Random Forest algorithm. This dashboard, which is made using Tableau, exploits the historical data to show how the integration with the real-time weather data can possibly improve the prediction accuracy and the operational efficiency.

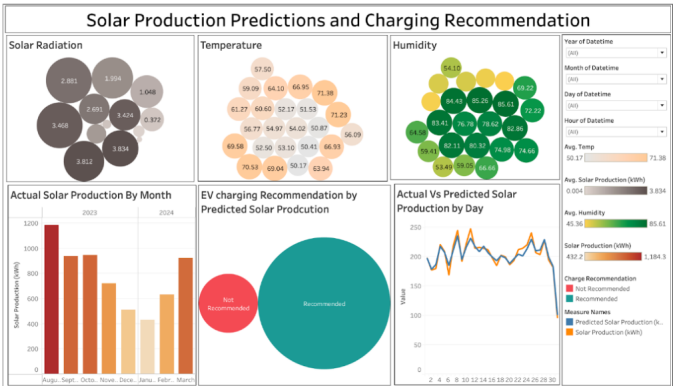


Fig. 18. Dashboard

The idea of creating this dashboard is driven by the necessity to efficiently control the solar energy creation and usage in the context of the growing dependence on renewable energy sources and electric vehicles. Management of solar production and EV charging schedules are the key elements to increasing solar production and matching energy availability with the charging schedules, which, in turn, will promote sustainability and reduce energy waste.

The technique adopted in our research includes all the steps of data collection, model creation, and the development of a dashboard. We take a huge dataset composed of about 536,435 images from Amazon, which are meticulously categorized with metadata like the measurements of the items, their weight, and the quantity of expected deliveries. The dataset is the foundation for the creation of our predictive model. The Random Forest model, which is known for being stable in

the cases of non-linearity and the diversity of the datasets, was the one that was used to train on this historical data to predict solar production values. The dashboard was designed utilizing Tableau to present the information and forecasts in the best possible way. It comprises various interactive components such as filters and charts that enable the users to get the most relevant information as shown in the figure.

## XV. CONCLUSION

This project has successfully demonstrated the potential of machine learning techniques in optimizing the utilization of solar energy for electric vehicle charging and residential energy consumption. By leveraging real-time data from solar panels, weather conditions, and historical patterns, the developed models can accurately predict solar energy production and provide homeowners with valuable insights into the best times to charge their electric vehicles or use household appliances.

The implementation of this project can have a significant impact on promoting sustainability and reducing reliance on traditional energy sources. By aligning energy consumption with solar production peaks, homeowners can minimize their dependence on the electricity grid, leading to substantial cost savings and a reduced carbon footprint. Moreover, the efficient use of renewable energy sources, such as solar power, contributes to mitigating the effects of climate change and promoting a greener, more environmentally friendly future.

The user-friendly Tableau dashboard created for the project enables homeowners to make smart energy use decisions, which results in a greater awareness and active participation in sustainable practices. Besides just the personal households, this approach also has the capacity to drive the collective shift towards a more advanced and sustainable energy system when scaled up.

In general, our project can be considered as an innovative and practical solution to the integration of renewable energy sources into one's daily life. Machine learning and data-driven insight enable it to be designed in a way that is more efficient, less expensive, and more sustainable for energy management.

## XVI. FUTURE SCOPE

While the project has achieved significant milestones, there are several avenues for further exploration and improvement:

- Integration with Smart Home Systems: Incorporating the developed machine learning models into existing smart home systems could enable seamless and automated control of energy consumption based on solar production forecasts. This integration could involve intelligent scheduling of appliances, lighting, and other household devices to optimize energy usage during peak solar production periods.
- Real-time Monitoring and Adaptation: Incorporating real-time monitoring and adaptive components aid in the system to improve the accuracy and responsiveness. Through continual data monitoring from solar panels and weather stations, the model could be updated dynamically to reflect the unchanging conditions which in turn run optimally the energy management decisions.
- Expansion to Larger Scale: While the current project focuses on a single household, scaling the solution to larger communities or neighborhoods could unlock additional benefits. By coordinating energy usage across multiple households with solar installations, the overall efficiency and grid impact could be further optimized, fostering a more sustainable energy ecosystem.
- Integration with Energy Storage Solutions: Incorporating energy storage systems, such as batteries or thermal storage, could further enhance the effectiveness of the "Smart Solar Scheduling" approach. By storing excess solar energy during peak production periods, households could leverage this stored energy during periods of low solar output, reducing reliance on the grid and increasing energy self-sufficiency.
- Exploration of Advanced Machine Learning Techniques: As machine learning algorithms continue to evolve, exploring and incorporating advanced techniques, such as deep learning or ensemble methods, could potentially improve the accuracy and robustness of solar production forecasting and energy usage optimization.

## XVII. LIST OF FIGURES



## REFERENCES


[1] Almughram O, Abdullah ben Slama S, Zafar BA. A Reinforcement Learning Approach for Integrating an Intelligent Home Energy Management System with a Vehicle-to-Home Unit. Applied Sciences. 2023; 13(9):5539. https://doi.org/10.3390/app13095539

[2] Aguilar-Dominguez, D., Ejeh, J., Dunbar, A. D. F., and Brown, S. F. (2021). Machine learning approach for electric vehicle availability forecast to provide vehicle-to-home services.Energy Reports, 7, 71–80. https://doi.org/10.1016/j.egyr.2021.02.053

[3] Forootan, M. M., Larki, I., Zahedi, R., and Ahmadi, A. (2022). Machine Learning and Deep Learning in Energy Systems: A Review. Sustainability, 14(8), 4832.https://doi.org/10.3390/su14084832

[4] Sharma, N., Sharma, P., Irwin, D., and Shenoy, P. (2011). Predicting solar generation from weather forecasts using machine learning. In IEEE SmartGridComm.https://doi.org/10.1109/SmartGridComm.2011.6102373

[5] Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., and Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. Renewable Energy,105, 569-582. https://doi.org/10.1016/j.renene.2016.12.095

[6] MathWorks. (n.d.). Understanding Support Vector Machine Regression. Retrieved from https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html

[7] Corke, T. (2012, February 16). Specialist Margin Prediction: Epsilon Insensitive Loss Functions. Matter of Stats. Retrieved from https://www.matterofstats.com/mafl-online/2012/2/16/specialist-margin-prediction-epsilon-insensitive-loss-functions.html