Journal of Aeronautics and Space Technologies

17(2) (2024) 180-202

Journal Homepage: jast.hho.msu.edu.tr

RESEARCH ARTICLE

Forecast Analysis of Renewable Solar Energy Production Using Meteorological Data with Machine Learning Methods

Makine Öğrenmesi Yöntemleriyle Yenilenebilir Güneş Enerjisi Üretiminin Meteorolojik Veriler Kullanılarak Tahmin Analizi

Naciye Macit Sezikli[1*], Ümit Alkan[2], Metin Zontul[3], Zeynep Elabiad [4]

[1] *İstanbul Gelisim University, Vocational School, 34310 İstanbul, Türkiye,*
[2] *İstanbul Gelisim University, Faculty of Engineering and Architecture, Department of Computer Engineering, 34310, İstanbul, Türkiye,*
[3] *Sivas University of Science and Technology, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, 58000, Sivas, Türkiye,*
[4] *International Road Federation, 22314, Alexandria, Virginia, USA*

Received: *April 1, 2024* ***Revised:*** *July 3, 2024* ***Accepted:*** *July 11, 2024*

Abstract

Solar energy power plants play a significant role in meeting the demand for sustainable and clean energy. However, variable weather conditions, seasonal effects, and similar factors can result in the need for energy overproduction to be stored or lead to costs associated with energy deficiency. These situations can result in inefficiencies in solar energy production. The objective of this study is to predict energy production, increase efficiency, and develop more sustainable energy strategies by using machine learning methods with data obtained from meteorological and solar energy panels. This study aims to assess the results achieved by existing models and compare their successes. The Random Forest algorithm, which achieved the highest R^2 score, also obtained significantly lower values for MSE, RMSE, and MAE. This indicates that the Random Forest algorithm performs the best among the algorithms used in this study. This ranking of success is followed by Decision Trees and K-Nearest neighbors.

Keywords: Solar Energy, Solar Energy Plant, Solar Energy Systems (SES) Parameters, Meteorological Parameters, Machine Learning, Machine Learning Methods

Öz

Güneş enerjisi santralleri, temiz ve sürdürülebilir enerji taleplerini karşılama konusunda önemli bir rol oynar. Ancak değişken hava koşulları, mevsimsel etkiler ve benzeri faktörler, fazla üretilen enerjinin depolanması veya eksik enerjiden kaynaklanan maliyetlerle sonuçlanabilir. Bu durumlar, güneş enerjisi üretiminde verimsizliklere yol açabilir. Bu çalışmanın hedefi, meteoroloji ve güneş enerji panellerinden elde edilen verileri kullanarak makine öğrenmesi yöntemleriyle enerji üretimini tahmin etmek, verimliliği artırmak ve daha sürdürülebilir enerji stratejileri geliştirmektir. Bu çalışma, mevcut modellerin elde ettiği sonuçları değerlendirmeyi ve başarılarını karşılaştırmayı amaçlamaktadır. En yüksek R^2 puanını alan Random Forest algoritması, aynı zamanda MSE, RMSE ve MAE değerlerinde de önemli ölçüde düşük sonuçlar elde etmistir. Bu da Random Forest algoritmasının bu çalışmada kullanılan algoritmalar arasında en iyi performans gösterdiğini göstermektedir. Bu başarı sıralamasını Karar Ağaçları ve K-En Yakın Komşu izlemektedir.

Anahtar Kelimeler: Güneş Enerjisi, Güneş Enerjisi Santrali, Güneş Enerjisi Parametreleri, Meteorolojik Parametreler, Makine Öğrenmesi, Makine Öğrenmesi Yöntemleri

*Corresponding Author
E-mail: nmacit@gelisim.edu.tr

1. INTRODUCTION

Factors such as population growth, industrialization, and rising levels of prosperity are driving up energy demand. This increased demand is leading to the depletion of primary energy sources or an increase in our dependence on external resources [1]. Primary energy sources such as oil, coal, and natural gas serve as the leading resources to meet our energy needs. However, renewable energy sources remain underutilized and await further development [2].

Solar energy's environmentally friendly and unlimited nature is a significant factor influencing energy policies. Renewable energy sources, being inexhaustible, clean, and environmentally friendly, also have the potential to increase job opportunities. However, there are disadvantages, such as variability in energy sources depending on seasons and weather conditions and high investment and storage costs. More research is needed to make energy production efficient and effective. In the field of renewable energy production, studies are being conducted in mechatronics, computer science, and artificial intelligence to increase solar energy production and improve the conversion efficiency of mechanical systems [3]. The "Role of Artificial Intelligence in Energy Workshop and Panel " has emphasized the importance of using digital technologies in energy efficiency studies. Numerous studies have been conducted in different geographic areas worldwide and Turkiye using various statistical methods and approaches like fuzzy logic and artificial neural networks.

1.1. Related Work

Various artificial intelligence and machine learning-based studies that combine Solar Energy Systems (SES) with meteorological data have reached predictions that enhance efficiency in energy production. This study examines the challenges and recommendations by applying different algorithms across various regions and datasets. In a 2012 study, the Support Vector Machines model, combined with solar power plant and temperature data, achieved promising results in predicting power production[4]. The second phase of a study conducted in 2016 added meteorological data to SES data, applied a machine learning model, and observed a 75% reduction in error rates, improving predictions[5]. Research conducted in 2018 highlighted radiation, humidity, and time data as critical parameters in energy production forecasts. It was also recommended that the angle of solar panels be considered, as well as their maintenance and cleaning [6]. Another study suggested that using more data and different forecasting methods could further improve the accuracy of predictions [7]. One of the studies conducted in 2019 concluded that among various machine learning methods, artificial neural networks provided more reliable results in solar energy prediction [8];another suggested that incorporating other parameters such as temperature, pressure, and precipitation in future studies could enhance the learning processes of artificial intelligence methods[9]. A 2019 study focused on predicting disruptions in energy production due to solar power plants (SPP) being covered with optical obstructing materials like snow, sand, and mud. Early detection of disruptions through machine learning algorithms, using SES data along with humidity, precipitation levels, solar

radiation, and pressure data, would enable quicker maintenance activities [10]. A 2020 study observed that the Decision Trees model displayed low prediction performance and a tendency for overfitting [11]. In 2021, the Random Forest algorithm provided the best prediction. The study found that relative humidity, temperature, and radiation significantly affected solar energy production, while the amount of daily precipitation was a less significant factor. Including cloudiness data in the training set was suggested, and removing nighttime meteorological data from the dataset could slightly reduce the error rate [12]. Another study examining the relationship between meteorological data and energy production offered suggestions to minimize the impact of adverse weather conditions on energy production and supply [13]. Another study's correlation analysis revealed a strong relationship between the energy produced by SPP and humidity, temperature, and solar radiation [14]. In 2022, an application was made on six different machine learning models using SES data and measured meteorological data such as cloudiness, wind speed, wind direction, temperature, and pressure. According to the study results, the Gradient Boosting Regression Trees and Artificial Neural Network models outperformed all other methods [15].

1.2. Aim and Objective

The primary goal of this study is to improve efficiency in the energy sector and optimize energy production using machine learning methods with Solar Energy Systems (SES) and meteorological data. This involves making future-oriented energy production forecasts and evaluating the employed machine learning techniques through consistency and validity analyses, comparing the successes of existing models. This study combines data from solar energy systems and meteorological parameters to improve the prediction of solar energy production.

Various machine learning algorithms, including Random Forest, Decision Trees, Polynomial Regression, K-Nearest Neighbors, Ridge Regression, Lasso Regression, and Multiple Linear Regression, are applied and their performance is compared. Comprehensive data cleaning, data aggregation, normalization, and handling of missing values are conducted to improve model performance. Cross-validation techniques are used to assess model robustness and accuracy, determining the optimal strategy.

The study provides practical guidance for enhancing the efficiency of solar energy production and supporting sustainable energy practices. Different from existing literature, the study is unique in its use of region-specific SES data and 24 different meteorological parameters from the Bilecik, Bozüyük region of Turkiye, contributing to the originality and thoroughness of the research. This comprehensive approach to data integration, machine learning application, and detailed performance comparison creates a solid foundation for predicting solar energy production.

By making solar radiation variability predictable, the study enhances the accuracy and reliability of energy production forecasts, contributing significant advantages in cost-effectiveness, efficiency, and strategic planning within the energy sector. Precise prediction models help in better planning and management of the energy supply chain, reducing instabilities and ensuring continuous, stable energy production. This provides

energy companies with significant advantages in long-term strategic planning and investment decisions, leading to more efficient implementation and management of solar energy plants. This study also demonstrates the applicability of machine learning algorithms to energy production, contributing to interdisciplinary research and data analysis techniques for academia. It provides a foundation for other researchers by showing which methods are more effective under specific conditions, assisting in the development of comprehensive prediction models, and supporting model optimization.

2. METHODOLOGY AND DATA PROCESSING

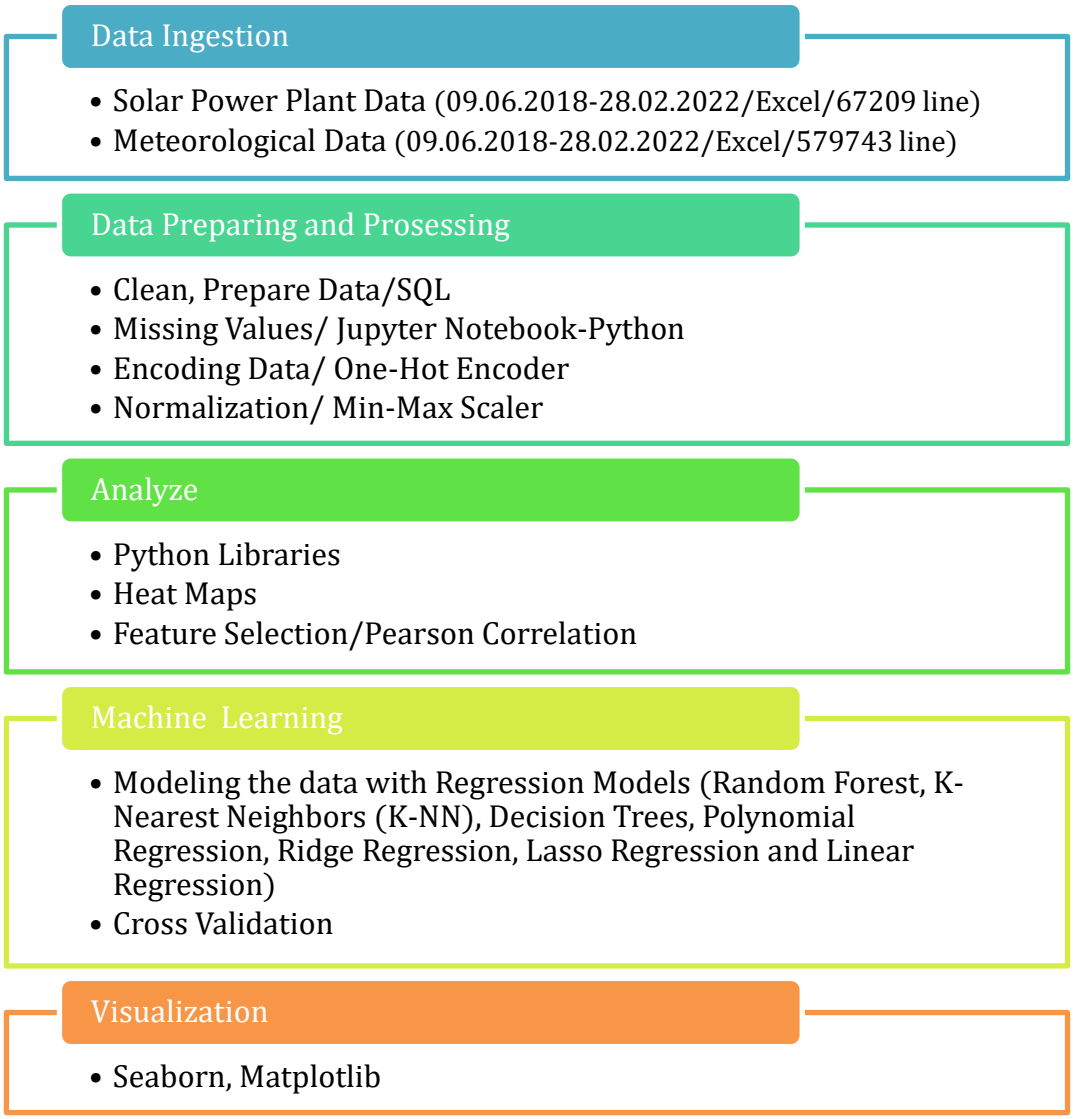


Figure 1. Methodology.

The raw data were merged and organized with the help of SQL, while all other processes were completed using the Python programming language within the Jupyter Notebook interface of the Anaconda program. SQL was essential for efficiently handling and

transforming extensive datasets. Specifically, SQL was used to merge multiple large Excel files into single tables, create aggregations for the parameter tables, and clean irregular timestamps and null values by standardizing and imputing data based on seasonal averages. This step reduced processing time, sustained data integrity, and minimized data movement between systems. It is emphasized that 80% of the data analysis process is allocated to the data cleaning and preprocessing phase [16]. This study consists of five main stages: Data Entry, Data Preprocessing (missing data, transforming data, normalizing data), Analysis (feature selection, cross-validation, heat maps), Machine Learning Applications, and Visualization (Figure 1).

2.1. Dataset

In this study, data recorded in hourly periods from June 9, 2018, to February 28, 2022, from three solar power plants (CEM SPP, CEMKA SPP, and Osman Aydın Şenol SPP) have been utilized. These data include parameters such as Inverter Energy Generation Cumulative (MW), Inverter Energy Generation (MW), Radiation Energy (W/m^2), and Cell Temperature (°C). Due to inconsistencies in the dataset's time parameters, it was decided to organize the data on a daily basis rather than using hourly intervals.

Hourly meteorological data obtained from the Bilecik Central Meteorological Station, provided by the Meteorological General Directorate, were also used within the scope of the study. Table 1 presents the data types, row counts, data ranges, and missing data information for Soil Temperature (°C), Actual Pressure (hPa), and the other 22 meteorological parameters, as well as Solar Energy System parameters. These data have been organized by considering the start and end times of solar energy production. As a result, a dataset consisting of 3,872 rows and 32 columns has been prepared (Table 1).

2.2. Missing Values

Incomplete, incorrect, or noisy data can negatively impact the performance of machine learning models, making preliminary exploratory analysis crucial before beginning data processing [17]. It is necessary to examine various details related to the dataset, such as the number of observations, types of variables, and amounts of missing data. Table 1 includes a 'Missing Data' column that summarizes the missing data for each feature.

The "Snow Depth" attribute contains the highest number of null values, totaling 3323. Since the Bilecik, Bozüyük region typically does not experience heavy snowfall, only the days with snowfall have been recorded, leaving records for days without snowfall missing. Thus, although more than 70% of this column contains missing data, it may not be appropriate to delete it completely. The missing data in this column have been assigned a snow depth value of 0, ensuring the dataset's consistency by assigning values to missing data. General assignment methods based on mean, median, and mode values have been considered to fill in missing data in other columns. However, since solar energy production varies with sunlight, it can exhibit dynamic changes between seasons and months. Therefore, a more specific approach has been adopted to fill in missing data. For all columns, missing data have been filled in, considering the mean, median, and mode values of the data in the same months as the missing data.

Table 1. Summary Table of Variables, Data Types, Ranges, and Missing Values.

VARIABLE NAME	PARAMETER NAME	DATA TYPE	RANGE MIN/MAX	ROW	MISSING DATA
SITE	Solar Energy Production Site	OBJECT	Cem/Cemka/OAS	67279	0
DATE (DAY, MONTH, YEAR)	Date	DATE TIME	9 June 2018/ 28 Feb. 2022	67279	0
START_TIME	Energy Production Start Time	INT	5/18	67279	0
END_TIME	Energy Production End Time	INT	10/22	67279	0
IRRADIANCE	Irradiance Energy	FLOAT	0.0/742.244	67279	0
CELL_TEMPERATURE	Cell Temperature	FLOAT	-4.78/45.88	67279	0
SOIL_TEMPERATURE_5CM	5 cm Soil Temperature	FLOAT	0.393/33.58	30737	24
ACTUAL_PRESSURE	Actual Pressure	FLOAT	932.25/970.22	31206	0
VAPOR_PRESSURE	Vapor Pressure	FLOAT	2.80/22.382	30431	33
EVAPOTRANSPIRATION	Evaporation Evapotranspiration	FLOAT	0.0/0.637	18699	1543
CLOUD_BASE_HEIGHT	Cloud Base Height	FLOAT	0.0/2500.0	15374	504
CLOUD_COVER	Cloud Cover	FLOAT	0.0/8.0	21357	0
INSOLATION_INTENSITY	Insolation Intensity	FLOAT	0.0/55.276	13559	1480
SUNBATHING_DURATION	Sunbathing Duration	FLOAT	0.0/0.988	11810	1795
WET_GLOBE_TEMPERATURE	Wet Globe Temperature	FLOAT	-7.176/22.96	21348	0
GLOBAL_SOLAR_RADIATION	Global Solar Radiation	FLOAT	0.0/640.0	13559	1480
MAX_RELATIVE_HUMIDITY	Maksimum Relative Humidity	FLOAT	30.75/98.062	27325	496
MAX_TEMPERATURE	Maksimum Temperature	FLOAT	-5.6/32.36	30971	48
SNOW_DEPTH	Snow Depth	FLOAT	0.0/35.0	2574	3323
MIN_RELATIVE_HUMIDITY	Minimum Relative Humidity	FLOAT	24.94/97.125	27324	496
MIN_TEMPERATURE	Minimum Temperature	FLOAT	-6.66/30.143	30971	48
RELATIVE_HUMIDITY	Relative Humidity	FLOAT	26.615/97.562	31019	0
TEMPERATURE	Temperature	FLOAT	-6.16/31.1	31263	0
TOTAL_GLOBAL_SOLAR_RADIATION	Total Global Solar Radiation	FLOAT	0.0/38400.0	13559	480
TOTAL_PRECIPITATION	Total Precipitation	FLOAT	0.0/2.712	31064	33
ABOVE_GROUND_MIN_TEMPERATURE	Above Ground Minimum Temperature	FLOAT	-4.135/40.057	29589	192
WIND_DIRECTION	Wind Direction	FLOAT	0.0/360.0	30005	4
WIND_SPEED	Wind Speed	FLOAT	0.0/4.53	30005	4
MAX_WIND_DIRECTION	Maximum Wind Direction	FLOAT	3.0/360.0	27997	39
MAX_WIND_SPEED	Maximum Wind Speed	FLOAT	1.125/13.72	27997	39
CUMULATIVE_ENERGY	Cumulative Energy Production by Inverter	FLOAT	0.0004/153.95	67279	0
ENERGY	Energy Production	FLOAT	0.0/16.658	67279	0

2.3. Encoding Data

Machine learning models generally work more effectively with numerical data. Therefore, converting categorical data into a numerical format is essential to help the model understand these data. Popular methods used for this transformation include ordinal encoding, label encoding, and one-hot encoding.

One-hot encoding treats different levels of categorical variables independently, making it suitable for unordered (nominal) data. It is also useful for representing categorical data without any ranking or priority condition. As a result, converting categorical data into numerical data is a critical step in machine learning projects. In this study, the only column in the dataset containing categorical data is the "LOCATION" column, which includes the names of the SPP and contains 3 different data: "CEM", "CEMKA", and "OAS". The One-Hot Encoding method has been applied to process this column. As a result of this process, new columns have been created for the "CEM", "CEMKA", and "OAS" categories in the "LOCATION" column, with the value of the relevant column marked as 1 for the corresponding observations, while it is marked as 0 for all other observations.

2.4. Normalization

In this study, the min-max normalization method has been applied to normalize the dataset values between 0 and 1. This process ensures that the values of the features in the columns are on the same scale without altering their actual values, thereby aiding machine learning models in producing more consistent results and balancing the impact of different features in the model. It has been observed that the R^2 scores obtained before normalization have increased after the normalization process. Equation 1 is used for Min-Max Normalization [18].

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

2.5. Machine Learning Algorithms Used in the Study

Machine learning algorithms are classified into three main categories: Supervised, unsupervised, and reinforcement learning [19]. Supervised Learning involves creating functions that map inputs to outputs, where training data include both inputs and their corresponding outputs. This category is primarily divided into two subcategories: classification and regression. Classification refers to the process of assigning input data to specific categories, whereas regression involves predicting a continuous output variable based on input data.

In this study, supervised learning methods were used due to the presence of labeled features in the dataset, the ability to fully digitize the data, and the energy-related nature of the target output. These methods include classification and regression algorithms such as Random Forest, Decision Trees, Polynomial Regression, K-Nearest Neighbors (K-NN), Ridge Regression, Lasso Regression, and Linear Regression.

Linear Regression aims to model the linear relationship between a dependent variable (outcome) and independent variables (inputs), using this model to predict future values

[20]. It is commonly used to understand how one variable is affected by others or to make predictions about a dependent variable. However, if there is no linear relationship between independent variables or if the relationship is too complex, other regression methods or machine learning algorithms may be more appropriate. The linear regression equation is given in Equation 2 [21].

$$y = a_0 + a_1 x + a_2 x_2 + \cdots + a_n x_n \tag{2}$$

Lasso Regression employs the multicollinearity problem in traditional linear regression using L1 regularization [22]. L1 regularization tends to shrink the regression model's coefficients towards zero, completely eliminating some while reducing others. This helps reduce the model's tendency to overfit, making it more stable. It also offers the ability to perform feature selection and can reduce the complexity of models in high-dimensional datasets. Lasso regression is expressed as shown in Equation 3 [20].

$$SSE_{L_1} = \sum_{i=1}^{n} (y_{i-} \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \tag{3}$$

Ridge Regression uses L2 regularization to solve the multicollinearity problem encountered in linear regression [23]. L2 regularization reduces the estimate variances without completely nullifying them, making the model more stable and reducing the tendency to overfit. In Ridge Regression, increasing the alpha parameter, which measures the intensity of regularization, can lead to a simpler and smoother model structure. Conversely, decreasing alpha can activate more features, increasing the model's complexity. Ridge regression is expressed as shown in Equation 5 [20].

$$SSE_{L_2} = \sum_{i=1}^{n} (y_{i-} \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} {\beta_j}^2 \tag{4}$$

K-NN bases its predictions on the class or value of observed neighbors of data points, operating on the premise that similar data points are likely to belong to the same class or have similar values [24]. Parameter K, which affects the model's performance, determines how many neighbors are considered when making predictions. Small K values can lead to overfitting, while large K values can offer lower complexity at the expense of lower detail levels [25]. The calculation of Euclidean distance is used to measure the distance between two given points and is expressed by the formula in Equation 5 [26].

$$d(x, y) = \sqrt{\sum_{i=1}^{n} \sqrt{(y_i - x_i)^2}} \tag{5}$$

Polynomial Regression is an extended version of Linear Regression that can represent more complex relationships [27]. It is particularly useful in scenarios where the relationship between data is not linear. However, it risks overfitting, as a very high polynomial degree can fit the data perfectly but reduce the ability to generalize to new data. The polynomial regression equation is usually expressed in the format of Equation 3 [28].

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \quad (6)$$

The Decision tree algorithm works by splitting the existing dataset into smaller groups of datasets through various decision rules [29]. Decision trees represent decisions and outcomes visually and logically in a hierarchical tree structure. Each node in the tree structure represents a feature, each branch represents a decision, and each leaf represents an outcome [30]. However, they can be sensitive to overfitting issues. Hence, the depth of the trees and other parameters should be carefully adjusted.

Random Forest can be effectively used for regression problems. As an ensemble model created by combining multiple decision trees, it allows for the production of more consistent and robust models. The model benefits from the aggregation of multiple decision trees, enabling a more accurate classification of the dataset [31]. Generally, using more trees can yield better results but may require more computational power. To determine which attribute is more decisive in predictions and to ensure the homogeneous distribution of nodes during the splitting and selection process, the Gini index, a criterion of the CART algorithm, is calculated [29]. As the Gini index decreases, the homogeneity of the class increases, and a branch is considered successful if the Gini index of the child node is lower than that of the parent node [32]. The calculation of the Gini index for a node is provided in Equation 7 [33].

$$GINI(n) = 1 - \sum_{j=1}^{2} \left(p_j\right)^2 \quad (7)$$

2.6. Performance Evaluation Metrics

In this study, machine learning algorithms were evaluated using performance metrics such as R-squared (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). R^2 measures the degree of variance of the independent variable on the dependent variable explained in regression analyses. This metric is used to assess how well the regression model fits and its predictive ability.

The R^2 value typically ranges between 0 and 1. A value of 0 indicates that the independent variables do not explain any variance of the dependent variable, while a value of 1 indicates that the independent variables fully explain the variance of the dependent variable. The R-square formula is given in Equation 8 [34].

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - y_{ort})^2} \quad (8)$$

MAE measures how much deviation there is from the actual values in the predictions of a regression model. It is calculated by taking the absolute difference between each prediction and the actual value and then averaging these absolute errors [35]. The MAE formula is given in Equation 9 [36].

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (9)$$

MSE is a metric where the square of the difference between each prediction and the actual value is calculated, and the average of these squares is taken [37]. The formula for

MSE is given in Equation 10 [38].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (10)$$

RMSE is calculated by taking the square root of the average of these squared errors. This provides a value expressed in the original units of the errors, helping to assess the model's predictive ability in a more understandable way [37]. The formula for RMSE is given in Equation 11 [38].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (11)$$

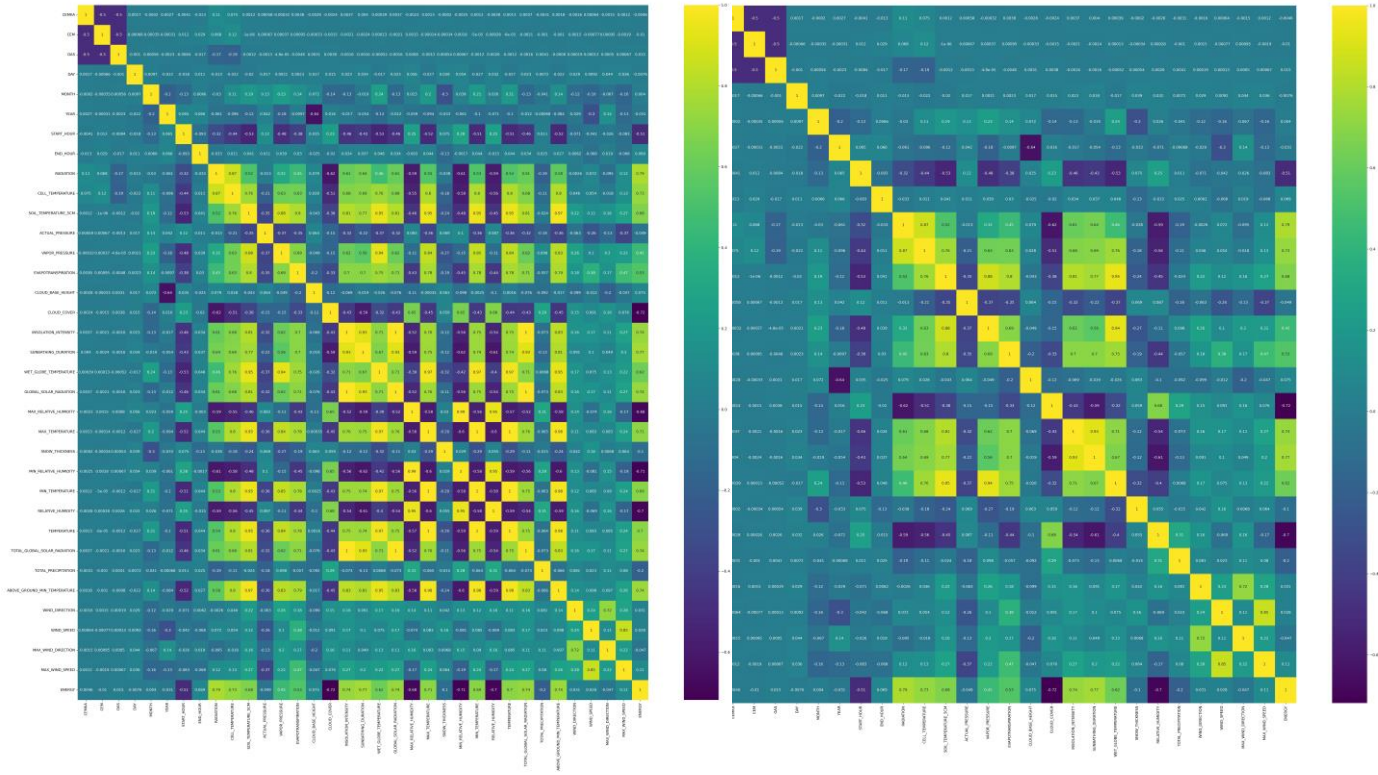
The closer the MSE, RMSE, and MAE values are to zero, the closer the model's predictions are to the actual values, and lower values reflect better model performance.

3. ANALYSIS

3.1. Feature Selection

In machine learning applications, selecting and creating the right features can enhance the model's performance and prevent overfitting. Therefore, using all features may not always be the best approach. The Pearson Correlation Method is commonly used to evaluate the relationship between numerical features. The correlation coefficient ranges between -1 and 1, indicating the nature of the relationship. A value of 1 represents a positive correlation, a value of -1 indicates a negative correlation, and 0 signifies no relationship. Heat maps help us understand the correlation relationships between data visually.

In this study, the Pearson correlation method was used to examine the relationships between variables, and features with high linear relationships were removed from the dataset. The removed parameters include Temperature, Minimum Temperature, Maximum Temperature, Above Ground Minimum Temperature, Maximum Relative Humidity, Minimum Relative Humidity, Global Solar Radiation, and Total Global Solar Radiation. Figure 2(b) indicates that the removal of these features reduced the distribution of yellow-colored and navy-colored areas (high relationship) in the heat map. This observation indicates that features that could lead to overfitting have been successfully removed from the dataset.



a) b)

Figure 2: Comparative Correlation Heatmaps Before a) and After b) Intervention.

3.2. Cross Validation

In the final data preparation stage, the primary objective is to complete the data preparation necessary for machine learning modeling successfully. This stage is crucial for assessing the accuracy and performance of the model and forms the foundation of machine learning projects. This process typically relies on the approach of splitting the data into 70% for training and 30% for testing. However, a disadvantage of this approach is that random splitting of the data can lead to deviations and errors in the model's training and testing phases. Cross-validation is used to solve this problem. It divides the data into equal parts according to a specified number of folds and uses each part for both training and testing. Cross-validation is a technique used to determine whether the model is sensitive to the data and if it faces an overfitting problem [39].

In this study phase, while dividing the data series into training and testing sets, the aim was to determine the most effective number of folds using cross-validation. During the cross-validation process, layers 3, 5, 7, and 10 were tested for each model. The R^2 scores obtained for each of these layers were analyzed through graphs, and the layer yielding the highest R^2 score was applied to the cross-validation method. Figure 3 indicates that, among the cross-validation experiments conducted for the Random Forest Algorithm, the layer achieving the highest efficiency was identified as layer 10.

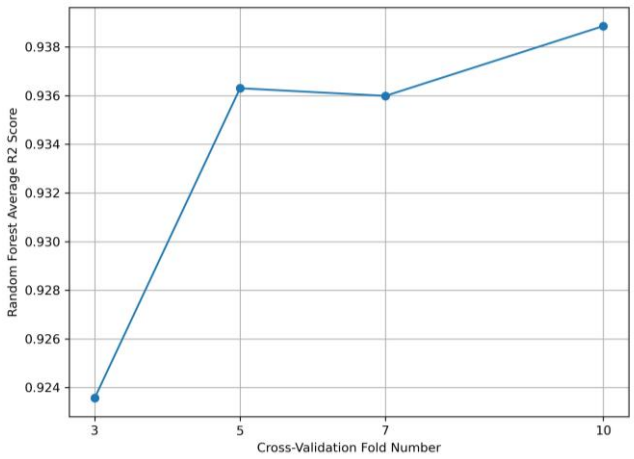


Figure 3. Comparison of layer numbers in cross-validation of the Random Forest algorithma.

For other models, the most efficient folds were as follows: 10 for Decision Trees, 5 for Polynomial Regression, 3 for K-Nearest Neighbors, 10 for Ridge Regression, 10 for Lasso Regression, and 5 for Linear Regression.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The application of the Random Forest Algorithm with 10-fold cross-validation has presented performance metrics, as shown in Table 2. According to these results, the most successful fold achieved an R^2 score of 0.977839, indicating a very close fit to the data, with an MSE value of 000470 and an MAE value of 0.010246. The model training time was 2.676702 seconds, and testing took 0.011588 seconds. The proximity of the R^2 score to 1 and the error metrics being near zero indicate that this model operates effectively, with predicted values very close to the actual values. Therefore, the Random Forest algorithm has been determined as the best-performing algorithm on our dataset. The close alignment and minimal deviation between the actual (green) and predicted (orange) values in Figure 4 effectively demonstrate the model's success.

Table 2. Cross-validation scores of a 10-fold Random Forest.

TRAIN TIME(sec)	TEST TIME(sec)	MAE	MSE	R^2
2.693017	0.011881	0.012725	0.000801	0.966911
2.699743	0.012547	0.015406	0.001240	0.947507
2.676702	0.011588	0.010246	0.000470	0.977839
2.686624	0.011264	0.014322	0.000875	0.962936
2.683924	0.011708	0.015851	0.001018	0.955305
2.712291	0.011445	0.012858	0.001198	0.942472
2.680358	0.011424	0.013175	0.001222	0.931541
2.645197	0.011388	0.017708	0.001327	0.942999
2.755475	0.009947	0.030055	0.002397	0.883474
2.709051	0.011982	0.029534	0.002617	0.882043

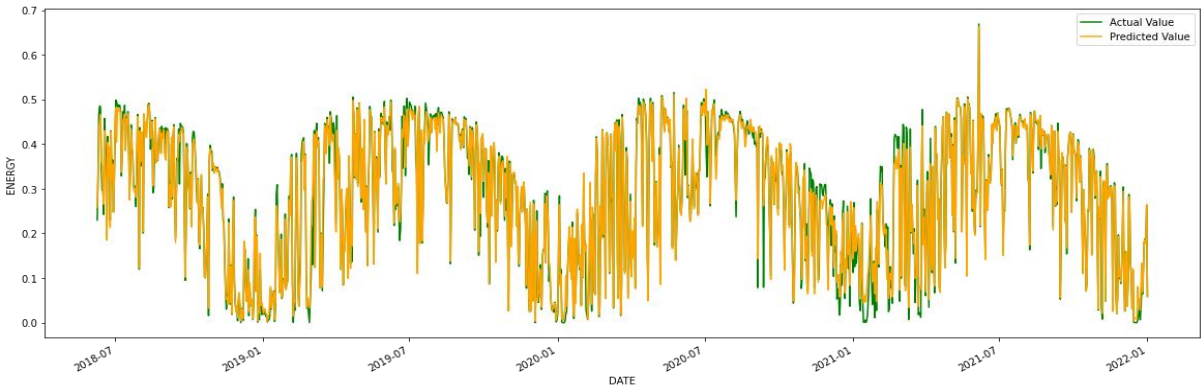


Figure 4. Actual vs. predicted values for Random Forest Algorithm.

When the Decision Trees algorithm was applied with 10-fold cross-validation, the performance metrics presented in Table 3 indicate that the most successful fold reached the second-highest R^2 score of 0.958071, with an MSE value of 0.001015 and an MAE value of 0.015920. The model's training time was 0.055333 seconds, and testing took 0.001745 seconds. This efficiency demonstrates that the model works effectively, with predicted values being very close to the actual values. Hence, the Decision Trees algorithm has been identified as the second-best performing algorithm on our dataset. While the modeling time is shorter compared to the Random Forest Algorithm, the R^2 score is lower. Figure 5 presents the actual (green) vs. predicted (orange) values for the Decision Trees algorithm.

Table 3. Cross-validation scores of Decision Trees with 10 folds.

TRAIN TIME(sec)	TEST TIME(sec)	MAE	MSE	R^2
0.055333	0.001745	0.015920	0.001015	0.958071
0.045795	0.001341	0.019678	0.002291	0.903002
0.044398	0.001374	0.014378	0.001387	0.934640
0.044352	0.001116	0.019781	0.002050	0.913171
0.042404	0.001022	0.018410	0.001508	0.933808
0.043641	0.001000	0.015193	0.002067	0.900792
0.042462	0.001001	0.016553	0.001467	0.917832
0.042120	0.000962	0.019907	0.001674	0.928063
0.043777	0.000954	0.038023	0.003757	0.817405
0.042618	0.000941	0.044692	0.006951	0.686678

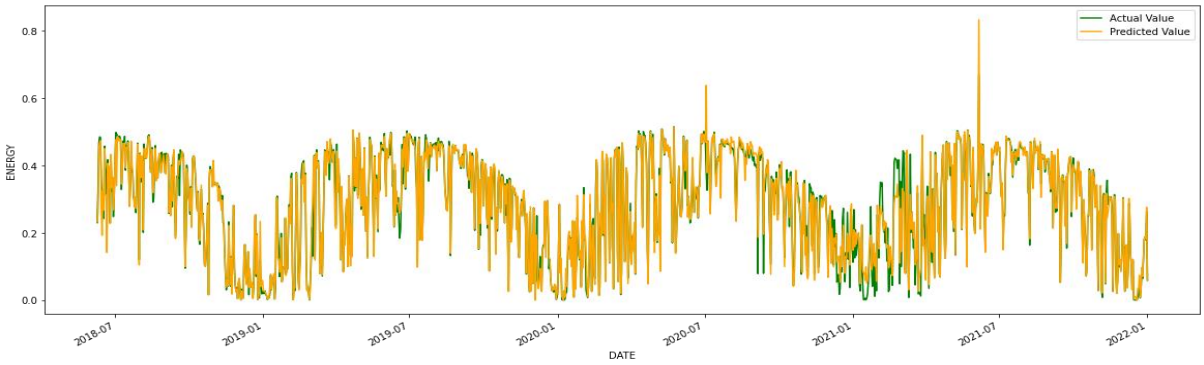


Figure 5. Actual vs. predicted values for Decision Trees Algorithm.

The polynomial algorithm, applied with 5-fold cross-validation, shows its performance metrics below in Table 4. According to these results, the most successful fold achieved the third-highest R^2 score of 0.934440, with an MSE value of 0.001499 and an MAE value of 0.022007, indicating that the model operates effectively. The model training took 0.043331 seconds, while the testing time was 0.000708 seconds. However, after applying polynomial regression to the dataset, the discrepancy between predicted values and actual values increased compared to other models, indicating a decrease in model performance (Figure 6).

Table 4. Cross-validation scores for Polynomial Regression with 5-fold.

TRAIN TIME(sec)	TEST TIME(sec)	MAE	MSE	R^2
0.246620	0.000776	0.023647	0.001667	0.930401
0.043331	0.000708	0.022007	0.001499	0.934440
0.109142	0.000834	0.026565	0.001683	0.922892
0.241825	0.001018	0.038778	0.003117	0.852737
0.193275	0.000864	0.057116	0.009124	0.575508

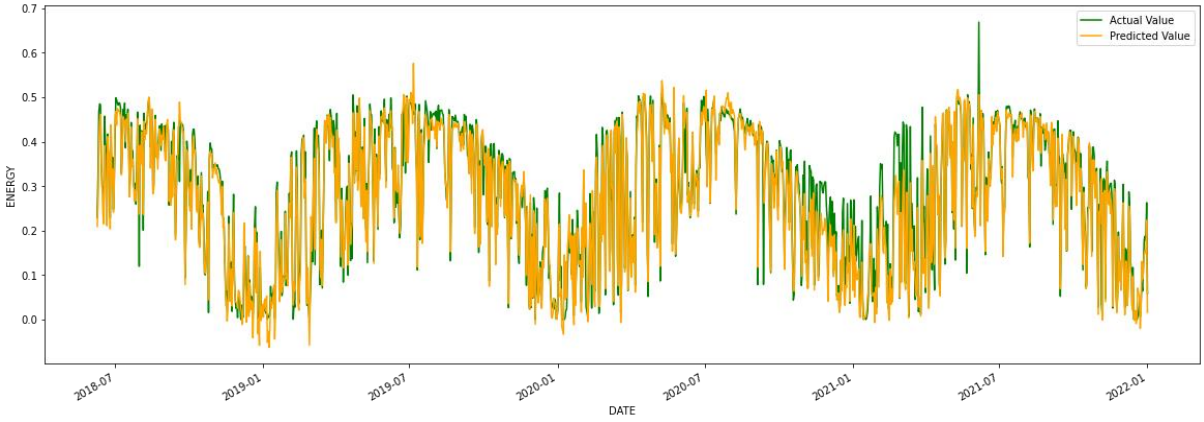


Figure 6. Actual vs. predicted values for Polynominal Algorithm.

The K-Nearest Neighbors (K-NN) algorithm, when applied with 3-fold cross-validation, has provided its performance metrics in Table 5. According to these outcomes, the most successful fold achieved a high R^2 score of 0.943126, indicating a strong fit to the data,

with an MSE value of 0.001308 and an MAE value of 0.023127. The model training took 0.004636 seconds, while the testing time was 0.057263 seconds. However, as indicated in the previous Figure 7, the discrepancy between predicted and actual values has increased compared to other models, signifying a decline in model performance.

Table 5. Cross-validation scores for K-Nearest Neighbors with 3-fold.

TRAIN TIME(sec)	TEST TIME(sec)	MAE	MSE	R^2
0.004636	0.057263	0.023127	0.001308	0.943126
0.001207	0.028936	0.047829	0.004600	0.798010
0.001203	0.028476	0.031314	0.002932	0.862946

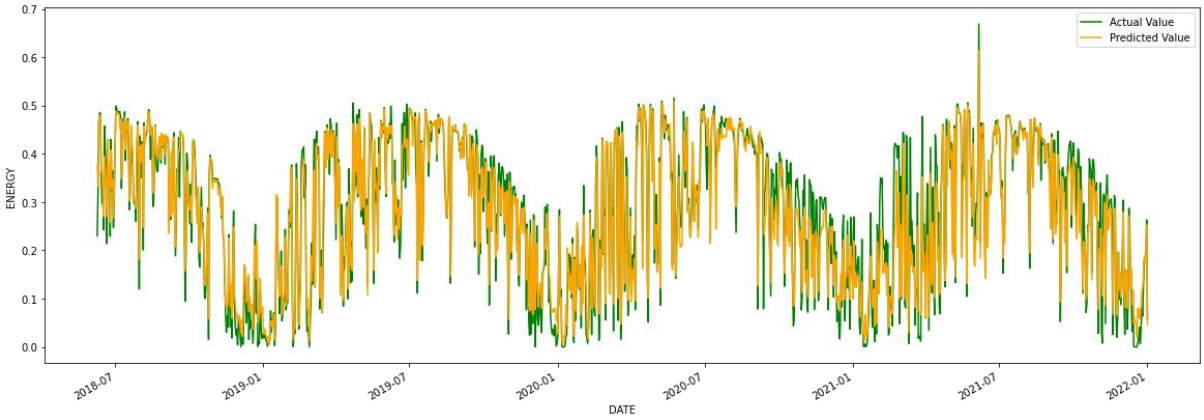


Figure 7. Actual vs. predicted values for K-Nearest Neighbors Algorithma.

For the Ridge Regression algorithm applied with 10-fold cross-validation, the performance metrics presented in Table 6 indicate that the most successful fold reached the second-highest R^2 score of 0.929428, with an MSE value of 0.001498 and an MAE value of 0.027092. The model training took 0.004549 seconds, while the testing time was 0.001214 seconds. The predicted and actual values for Ridge Regression are shown in Figure 8.

Table 6. Cross-validation scores for Ridge Regression with 10-fold.

TRAIN TIME(sec)	TEST TIME(sec)	MAE	MSE	R^2
0.017350	0.001596	0.034323	0.002038	0.915765
0.004416	0.001207	0.032414	0.002561	0.891551
0.004549	0.001214	0.027092	0.001498	0.929428
0.005079	0.001236	0.035063	0.002390	0.898773
0.004131	0.001203	0.036170	0.002433	0.893215
0.004400	0.001530	0.033049	0.002144	0.897057
0.004039	0.002012	0.030421	0.002481	0.861025
0.004102	0.001081	0.036942	0.002902	0.875300
0.004174	0.000986	0.049912	0.004374	0.787406
0.004017	0.001562	0.049322	0.006135	0.723466

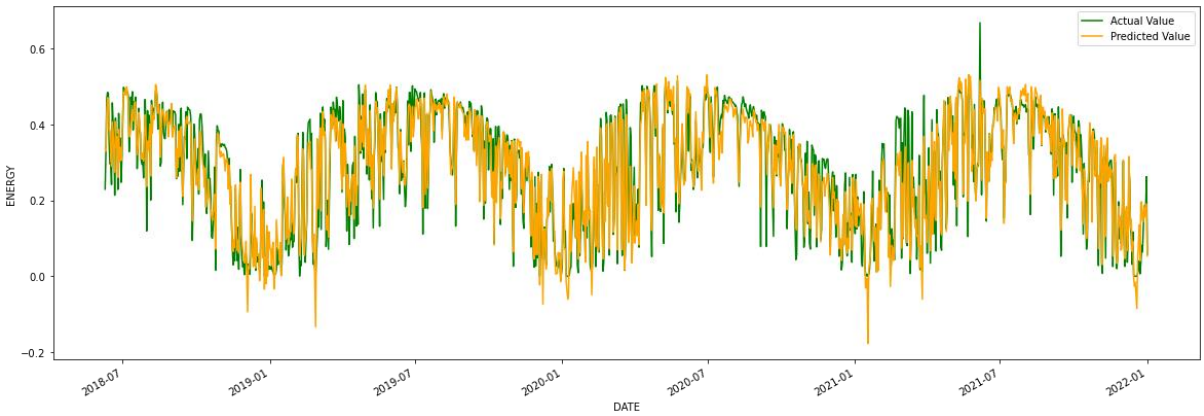


Figure 8. Actual vs. predicted values for Ridge Regression.

The Lasso Regression algorithm, when applied with 10-fold cross-validation, showed its performance metrics in Table 7. According to these results, the most successful fold achieved an R^2 score of 0.930667, an MSE value of 0.001471, and an MAE value of 0.026507. The model training took 0.073838 seconds, while the testing time was 0.001043 seconds. The predicted and actual values for the Lasso Regression are shown in Figure 9.

Table 7. Cross-validation scores for Lasso Regression with 10-fold.

TRAIN TIME(sec)	TEST TIME(sec)	MAE	MSE	R^2
0.088474	0.001396	0.034482	0.002055	0.915055
0.078327	0.001167	0.032605	0.002598	0.889981
0.073838	0.001043	0.026507	0.001471	0.930667
0.073755	0.001203	0.034960	0.002395	0.898565
0.072219	0.001068	0.036647	0.002502	0.890186
0.073340	0.001204	0.033124	0.002145	0.897031
0.072662	0.001109	0.030473	0.002506	0.859606
0.078145	0.001266	0.036719	0.002898	0.875467
0.070600	0.001138	0.050945	0.004554	0.778652
0.071725	0.001000	0.048838	0.006067	0.726527

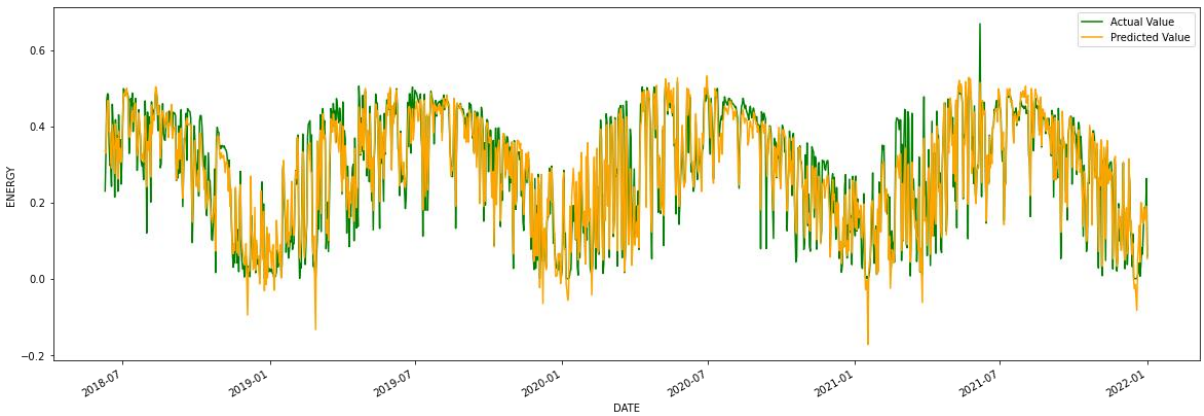


Figure 9. Actual vs. predicted values for Lasso Regression.

For the Linear Regression algorithm applied with 5-fold cross-validation, the performance metrics presented in Table 8 indicate that the most successful fold reached the second-highest R^2 score of 0.91252, with an MSE value of 0.002 and an MAE value of 0.03169. The model training took 0.00375 seconds, while the testing time was 0.00127 seconds. The predicted and actual values for Linear Regression are shown in Figure 10.

Table 8. Cross-validation scores for Linear Regression with 5-fold.

TRAIN TIME(sec)	TEST TIME(sec)	MAE	MSE	R^2
0.016861	0.001614	0.034770	0.002372	0.900943
0.003758	0.001279	0.031693	0.002001	0.912526
0.002433	0.001098	0.034438	0.002290	0.895052
0.002113	0.001093	0.035152	0.002846	0.865547
0.003341	0.001920	0.058762	0.007975	0.628970

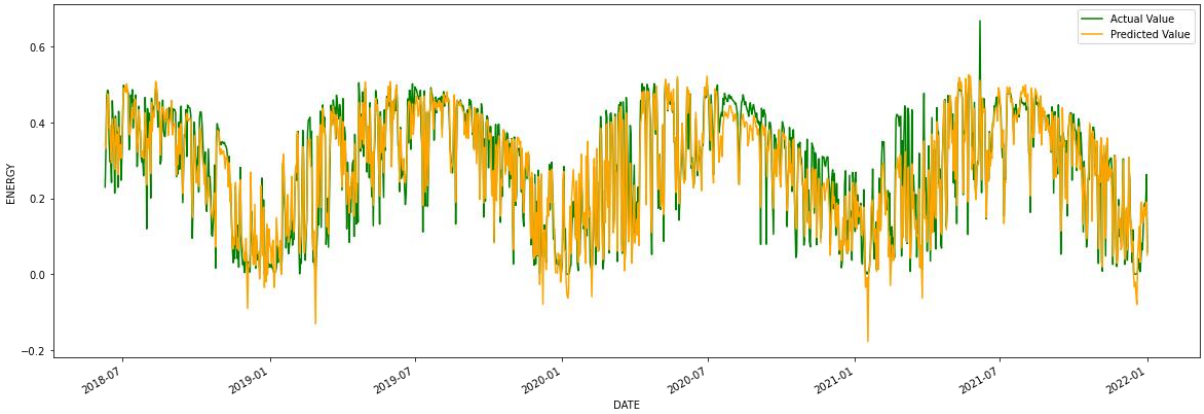


Figure 10. Actual vs. predicted values for Linear Regression.

Comparing the results of Ridge, Lasso and Linear Regression, it is observed that despite the R^2 scores being close to 1, these algorithms performed lower compared to other applications. Additionally, the increase in green areas in Figures 8, 9, and 10, which indicate the discrepancy between actual and predicted values, suggests that the performance of these models is lower compared to others. The comparison of R^2, MSE, MAE, RMSE, training times, and testing times for all algorithms applied to the dataset has been presented in Table 9.

Table 9. Performance comparison table of machine learning algorithmas.

	R^2	MAE	MSE	RMSE	TRAIN TIME(sec)	TEST TIME(sec)
RANDOM FOREST	0,977839	0,010246	0,00047	0,021679483	2,679702	0,011588
DECISION TREES	0,958071	0,01592	0,001015	0,031859065	0,055333	0,001745
K-NEAREST NEIGHBORS	0,943126	0,023127	0,001308	0,036166283	0,004636	0,057263
POLYNOMINAL	0,934440	0,022007	0,001499	0,038716921	0,043331	0,000708
LASSO REGRESSION	0,930667	0,026507	0,001471	0,038353618	0,073838	0,001043
RIDGE REGRESSION	0,915765	0,034323	0,002038	0,045144213	0,017350	0,001596
LINEAR REGRESSION	0,912526	0,031693	0,002001	0,044732538	0,003758	0,001279

5. CONCLUSION

The R^2 scores are considered an important performance metric for assessing the success of an algorithm. While the R^2 score is crucial, it alone may not suffice to determine a model's performance comprehensively. A model is considered to perform well if its R^2 score is close to 1 and its error metrics (MAE, MSE, and RMSE) are low. The graph of R^2 scores for machine learning models is provided in Figure 11.

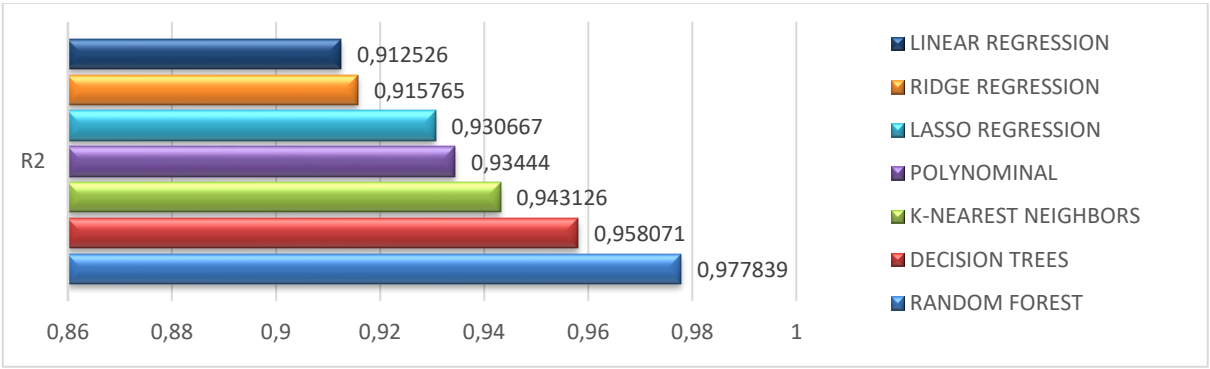


Figure 11. R^2 score chart for machine learning models.

As observed in Figure 11 and Figure 12, the Random Forest algorithm, which achieved the highest R^2 score, also obtained significantly low values for MSE(Figure 12b), RMSE(Figure 12c), and MAE(Figure 12a). This indicates that the Random Forest algorithm performs the best among the algorithms used in this study. This ranking of success is followed by Decision Trees and K-Nearest Neighbors.

Figure 12. MAE(a), MSE(b), RMSE(c) charts for machine learning models.

The outcomes and findings of this study present a series of recommendations aimed at promoting sustainable transformation within the energy sector and supporting the efficient utilization of renewable energy sources.

Collaborations with Industry: Collaborations with industry and partnerships with the private sector can be established to facilitate the rapid development and implementation of renewable energy projects.

Cooperation with the Meteorological Directorate: Ongoing cooperation with the Meteorological Directorate can ensure that energy production potential analyses are regularly updated using current climate data. This collaboration will help identify the most suitable regions for Solar Energy Plants and facilitate the conduction of feasibility analyses.

Energy Source Diversification: It is suggested that studies focus not only on solar energy but also on other renewable energy sources such as wind energy. This approach encourages the diversification of the energy supply and enhances energy security.

Optimization of Random Forest Algorithm: The Random Forest algorithm stands out for its modeling success and shorter training duration. However, limitations in tree depth could lead to overfitting issues. In future studies, the algorithm's parameters could be examined in more detail, and more effective solutions could be produced by exploring different methods.

Automation of Maintenance and Cleaning: Recording maintenance and cleaning times in solar energy plants and using this data can contribute to the development of automation capable of performing automatic maintenance and cleaning tasks. Additionally, implementing automations that vary depending on the position of the sun and predicting the effects of these variables on energy production can enhance efficiency.

These recommendations can enhance the sustainability of the energy sector and increase Turkiye's capacity to meet its energy demand through environmentally friendly and sustainable practices.

REFERENCES

[1] M. Yılmaz, *The energy potential of Türkiye and its importance of renewable energy sources in terms of electricity production,* in Ankara University Journal of Environmental Sciences, **4**(2), 33-54 (2012).

[2] M. C. Şenel, *Design Principles And Dynamic Behaviour Of Power Transmission Mechanisms In Wind Turbine*, Master's Thesis, Institute of Science, Ondokuz Mayıs University. Türkiye: Samsun, 2012. [Online]. Available: Açık Bilim, https://acikbilim.yok.gov.tr /handle /20.500.12812/675572 [Accessed: 11 Jan. 2024].

[3] B. Düzgün, TUBA-Role of Artificial Intelligence in Energy Workshop and Panel, Turkiye: Ankara, 2022, pp.30 [Online]. Available: Researchgate, https://www.researchgate.net/profile/Mehmet-Ezan/publication/368246136_TUBA-Enerjide_Yapay_Zekanin_Rolu_Raporulinks/63de053362d2a24f92f7c966/TUeBA-Enerjide-Yapay-Zekanin-Rolue-Raporu.pdf [Accessed: 11 Jan. 2024].

[4] J. Shi, W. J. Lee, Y. Liu, Y. Yang and P. Wang, *Forecasting power output of photovoltaic systems based on weather classification and support vector machines*, IEEE Transactions on Industry Applications, **48**(3), 1064-1069 (2012).

[5] B. Wolff, J. Kühnert, E. Lorenz, O. Kramer and D. Heinemann, *Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data*, Solar Energy, **135**, 197-208, October 2016.

[6] C. Yavuz, "Short-term solar energy production estimation in Konya", Master's Thesis, Institute of Science, Necmettin Erbakan University, Konya, Turkiye, 2018. [Online]. Available: Proquest, https://www.proquest.com/openview/d1b953c084074bab5723a84 02ba24e38/1?pq-origsite=gscholar&cbl=2026366&diss=y, [Accessed: 11 Jan. 2024].

[7] T. Akman, *Analysis of Electrical Load Forecasting Methods*. Turkiye, Ankara, 2018.

[8] S. Uğuz, O. Oral and N. Çağlayan, *Estimation of Energy to be Obtained from PV Power Plants Using Machine Learning Methods*, in International Journal of Engineering Research and Development, **11**(3), 769-779 (2019).

[9] A. O. Gök, "Developing A Short Term Generation Forecast System For Solar Power Plants", Master's Thesis, Institute of Science, Kahramanmaraş Sütçü İmam University, Kahramanmaraş, Turkiye, 2019.

[10] C. G. Haba, "Monitoring solar panels using machine-learning techniques", in *8th international conference on modern power systems (MPS), pp.1-6, May 2019* [Online]. Available: https://doi.org/10.1109/MPS.2019.8759651, [Accessed: 11 Jan. 2024].

[11] A. Khalyasmaa, S. Eroshenko and D. C. Tran, "Very-short term forecasting of photovoltaic plants generation based on meteorological data from open sources using machine learning", *In 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 564-568, October 2020* [Online]. Available: https://doi.org/10.1109/ICSTCEE49637.2020.9276765, [Accessed: 11 Jan. 2024].

[12] K. Mahmud, S. Azam, A. Karim, S. Zobaed, B. Shanmugam and D.Mathur, *Machine learning based PV power generation forecasting in alice springs*, IEEE Access, **9**, 46117-46128, (2021). [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3066494, [Accessed: 11 Jan. 2024].

[13] N. Özdoğan, "The Impact Of Weather On Energy Supply - Examination Of Solar Power Plants In Kahramanmaraş And Adiyaman Region" Master's Thesis, Institute of Science, Kahramanmaraş Sütçü İmam University, Kahramanmaraş, Turkiye, 2021.

[14] C. Yıldız, H. Açıkgöz, *A kernel extreme learning machine-based neural network to forecast very short-term power output of an on-grid photovoltaic power plant* in Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, **43**(4), 395-412 (2020).

[15] S. Sivakumar, B.Neeraja, M. Jamuna Rani, H. Anandaram, S. Ramya, G. Padhan and S. Gurusamy, et al, *Machine learning approach on time series for PV-solar energy*, in Advances in Materials Science and Engineering, (2022). [Online]. Available: https://doi.org/10.1155/2022/6458377, [Accessed: 11 Jan. 2024].

[16] T. Dasu and T. Jahnson, *Exploratory Data Mining and Data Cleaning,* Wiley-IEEE, Published by John wiley and Sons, New Jersey, Canada, (2003). [Online]. Available:

Books Google, https://books.google.com.tr/books?hl=tr&lr=&id=HStUfJnCvT4C&oi=fnd&pg=PR5&dq=Exploratory+Data+Mining+and+Data+Cleaning.+Wiley-IEEE.+dasu&ots=duIW L9n7t&sig=wqoAeg3Q0B3WY_r4DpFPbqMBJhc&redir_esc=y#v=onepage&q=Exploratory%20Data%20Mining%20and%20Data%20Cleaning.%20Wiley-IEEE.%20dasu& f=false, [Accessed: 11 Jan. 2024].

[17] C. Ye, H. Wang, J. Li, H. Gao and S. Cheng, *Crowdsourcing-Enhanced missing values imputation based on Bayesian network*, in International Conference on Database Systems for Advanced Applications, Springer 67–81, (2016). [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-32025-0_5, [Accessed: 11 Jan. 2024].

[18] T.Ş. Yapraklı and H. Erdal, *Firm Failure Prediction: A Case Study Based On Machine Learning*, Journal of Information Technologies, **9**(1), 21-31 (2016).

[19] J. A. Castellanos-Garzón, E. Costa, J. M. Corchado, *An evolutionary framework for machine learning applied to medical data*, Knowledge-Based Systems, **185**, 1-8, 104982, December 2019 [Online]. Available: https://www.sciencedirect.com/science/article./abs/pii/S0950705119304046, [Accessed: 11 Jan. 2024].

[20] E. N. Nacar and B. Erdebilli, *Sales prediction using data analytics: A machine learning approach*, Journal of Industrial Engineering, **32**(2), 307-320 (2021).

[21] M. H. Kutner, C. J. Nachtsheim, C. J., J. Neter, and W. Li, "Applied linear statistical models", (2005).

[22] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society Series B: Statistical Methodology, **58**(1), 267-288 (2018).

[23] A. E. Hoerl and R. W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, **42**(1), 80-86, (2000).

[24] G. Dilki, Ö. D. Başar, *Comparison Study Of Distance Measures Using K- Nearest Neighbor Algorithm On Bankruptcy Prediction*, İstanbul Ticaret University İstanbul Commerce University Journal of Science, **19**(38), 224-233 (2020).

[25] V. Sinap, *Prediction of Counter-Strike: Global Offensive Round Results with Machine Learning Techniques*, in Journal of Intelligent Systems: Theory and Applications, **6**(2), 119-129 (2023).

[26] P. Cunningham and S. J. Delany, K-nearest neighbour classifiers-a tutorial, "K-nearest neighbour classifiers-a tutorial." *ACM computing surveys (CSUR)* 54.6 (2021): 1-25.

[27] S. Arslankaya, Ş. Toprak, *Using Machine Learning and Deep Learning Algorithms for Stock Price Prediction*, International Journal of Engineering Research and Development, **13**(1), 178-192 (2023).

[28] E. Ostertagová, "Modelling using polynomial regression." *Procedia Engineering* , 48 (2012): 500-506.

[29] F. Akalın, *Polyp Diagnostic With Heuristic Algorithm Techniques On Small Bowel Images*, Master's Thesis, Institute of Sciece, Sakarya University. Turkiye: Sakarya, 2020.

[30] Y. Gültepe, *A Comparative Assessment on Air Pollution Estimation by Machine Learning Algorithms*, European Journal of Science and Technology, **16**, 8-15 (2018).

[31] C. Hedge, S. Wallace, K. Gray, "Using trees, bagging, and random forests to predict rate of penetration during drilling" *in SPE Middle East Intelligent Oil and Gas Conference and Exhibition, SPE-176792-MS, September 2015.*
[32] Ö. Akar, O. Güngör, Classification of multispectral images using Random Forest algorithm, Journal of Geodesy and Geoinformation, **106**, 139-146, (2012). [Online]. Available: https://doi.org/10.9733/jgg.241212.1t, [Accessed: 11 Jan. 2024].
[33] C. T. Okolo, *Diabetes Prediction Using Machine Learning Algorithm*. University of Louisiana at Lafayette, 2022.
[34] F. Yetiz, M. Terzioğlu, and M. Kayakuş. "An Analysis of Turkish Deposit Banks' Customer Forecasting with Machine Learning Methods" *Sosyoekonomi Society* 29.50 (2021): 413-432.
[35] N. S. Özen, S. Saraç, M. Koyuncu, *Prediction of COVID-19 Cases in the United States of America with Machine Learning Algorithms*, European Journal of Science and Technology, **22**, 134-139 (2021).
[36] T. Chai and RR. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)." *Geoscientific model development discussions* 7.1 (2014): 1525-1534.
[37] M. Özşahin, K. B. Yilmaz, S. Akcan, Makine Öğrenmesi Algoritmalari İle Ev Kirasi Tahmini. [Online]. Available: https://www.researchgate.net/profile/Metin-Oezsahin/publication/366986757_MAKINE_OGRENMESI_ALGORITMALARI_ILE_EV_KIRASI_TAHMINI/links/63bd257fc3c99660ebe2de6a/MAKINE-OeGRENMESI-ALGORITMALARI-ILE-EV-KIRASI-TAHMINI.pdf, [Accessed: 11 Jan. 2024].
[38] Ş. Seyhan, "Prediction of Safety Factor with Machine Learning on Retaining Walls", Master's Thesis, Institute of Sciece, Karatay University., Turkiye, 2021.
[39] E. Sorhun, *Python ile Makine Öğrenmesi,* CA: Turkiye, İstanbul, 2023.

To Cite This Article, N. Macit Sezikli, Ü. Alkan, M. Zontul, Z. Elabiad, Forecast Analysis Of Renewable Solar Energy Production Using Meteorological Data With Machine Learning Methods, Journal of Aeronautics and Space Technologies **17**(2), 180-202 (2024).

VITAE

Naciye Macit Sezikli received her B.Sc. degree in Computer Systems from the Faculty of Technical Education, Suleyman Demirel University, Turkiye, in 2009. She received her M.Sc. degree in Mechatronics Engineering from the Faculty of Engineering and Architecture, Istanbul Gelisim University, Turkiye, in 2023. She is currently working at Istanbul Gelisim University as a Lecturer.

Ümit Alkan received his B.Sc. degree in Physics from the Faculty of Arts and Sciences, Department, Yıldız Technical University, Turkiye, in 2001. He received his M.Sc. degree in Physics from the Institute of Science, Yıldız Technical University, Turkiye, in 2004. He received his Ph.D. degree in Physics from the Institute of Science, Yıldız Technical University, Turkiye, in 2011. He is currently working at İstanbul Gelişim University as a Doctor Lecturer of the Faculty of Engineering Archıtecture.

Metin Zontul received his B.Sc. degree in Computer Engineering from the Faculty of Engineering, Middle East Technical University, Türkiye, in 1993. Later, he received his

M.Sc. degree in Computer-Aided Design, Production, and Programming from Erciyes University, Türkiye, in 1996. Finally, he received his Ph.D. degree in Numerical Methods from Cumhuriyet University, Türkiye, in 2004. He is currently working at Sivas University of Science and Technology as Dean of the Faculty of Engineering and Natural Sciences.

Zeynep Elabiad received her B.Sc. degree in Computer Systems from the Faculty of Technical Education at Suleyman Demirel University, Türkiye, in 2009. She earned her M.Sc. degree in Computer Engineering from the Faculty of Engineering at Istanbul Aydin University, Türkiye, in 2013. She obtained her second M.Sc. degree in Data Science from Indiana University Bloomington, USA, in 2023. She is currently working as an IT manager at the International Road Federation in the US.