

Research Article

Received: date:06.04.2025

Accepted: date:21.05.2025

Published: date:30.06.2025

Machine Learning-Based Wind Energy Forecasting Using Weather Parameters: The Example of Yalova

Abdulkadir Atalan^{1, *}, İlker Çam², Lütfi Alper Gündoğdu³, Harun Kahyalık⁴ and Yasemin Ayaz Atalan⁵

¹Department of Industrial Engineering, Çanakkale Onsekiz Mart University, Çanakkale Türkiye; abdulkdiratalan@gmail.com

²Department of Energy Department; Çanakkale Onsekiz Mart University, Çanakkale Türkiye; ilkerarda15@gmail.com

³Department of Energy Department; Çanakkale Onsekiz Mart University, Çanakkale Türkiye; lutfialpergundogdu@gmail.com

⁴Department of Energy Department; Çanakkale Onsekiz Mart University, Çanakkale Türkiye; hrnkhyllk@gmail.com

⁵Department of Energy Department; Çanakkale Onsekiz Mart University, Çanakkale Türkiye; yayazatalan@gmail.com

Orcid: 0000-0003-0924-3685¹ Orcid: 0009-0007-5313-5776² Orcid: 0000-0002-2002-0952³ Orcid: 0009-0008-9636-8008⁴ Orcid: 0000-0001-7767-0342⁵

*Correspondence: abdulkdiratalan@gmail.com, yayazatalan@gmail.com

Abstract: In this study, various machine learning algorithms were evaluated for estimating wind energy production using hourly meteorological data of Yalova province in 2018. The input parameters were input parameters of weather parameters such as temperature, relative humidity, air pressure, wind direction, and wind speed. In the analysis performed on a total of 50530 data points, methods such as Gradient Boosting (GB), Random Forests (RF), k-nearest neighbor (kNN), and Stochastic gradient descent (GBD) were compared. Model performances were evaluated according to Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), MAPE, and R² criteria. According to the results, the best-performing algorithm was RF with an MSE value of 0.039, RMSE value of 0.197, MAE value of 0.081, MAPE value of 0.377, and R² score of 0.961. On the other hand, the SGD model showed the lowest performance with an MSE value of 0.175, RMSE value of 0.418, MAE value of 0.303, MAPE value of 0.581, and R² score of 0.822. These findings show that machine learning models, supported by selecting the correct weather parameters, can provide high accuracy in estimating wind energy production and contribute to energy management policies in this direction.

Keywords: Wind energy, weather parameters, machine learning, forecasting

1. Introduction

Today, energy holds much greater global significance than in the past. A significant portion of the current energy demand is met through fossil fuels [1]. The various environmental, economic, and health-related negative impacts of fossil fuels make it necessary to seek new and sustainable solutions in the energy sector [2]. In this regard, renewable energy sources, developed as alternatives to depletable energy resources, are expected to play a much more significant role in future energy production [3]. Renewable energy sources, such as wind, solar, geothermal, biomass, hydroelectric, and wave energy, stand out due to their sustainable and environmentally friendly nature [4]. Enabling countries to produce energy within their borders, reducing dependence on foreign sources, and ensuring long-term economic benefits make these energy sources strategically valuable. Furthermore, the inexhaustible nature of renewable energy sources and their ability to minimize carbon emissions and reduce environmental impacts strengthen their critical role in the global energy transition process.

Wind energy has gained particular prominence among renewable resources due to its clean and inexhaustible nature [5]. As a renewable and clean energy source, wind energy has emerged as an alternative for energy production [6]. Its environmental benefits, increased electricity production, and

reduced fossil-based energy production have made wind energy increasingly important [7]. The amount of energy produced in wind farms is related to weather parameters [8]. For efficient use of wind energy, changes in wind speed need to be observed, and production potential must be accurately predicted [9].

Electricity production from wind energy is highly variable because it directly depends on atmospheric conditions. Variations in wind speed, changes in direction, and other meteorological factors are the main factors determining the daily and seasonal fluctuations in energy production. Machine learning techniques are frequently used to model these variations, and hybrid models are developed to improve accuracy [10]. Atmospheric parameters such as temperature, pressure changes, humidity, and air density directly affect the efficiency of wind turbines. Changes in air density can cause the same wind speed to carry different amounts of energy. These variations necessitate reliable forecasting methods to ensure the stable operation of energy plants. In this context, machine learning algorithms are used to forecast the power produced by wind turbines, and the performance of these methods is evaluated using various statistical criteria [11]. Ensuring grid security, making energy production planning more predictable, and enhancing economic sustainability are critical aspects where the accuracy of wind energy forecasts plays a vital role.

Wind energy generation occurs when wind turbines convert kinetic energy into electrical energy [12]. The movement of the wind strikes the blades of the turbine, causing them to rotate. This rotational movement is converted into mechanical energy through the rotor and transmitted to the turbine shaft [13]. The mechanical energy transferred from the turbine shaft to the electric generator is converted into electrical energy and is ready for use [14]. The electricity is increased to the appropriate voltage level with the help of transformers and transmitted to the electricity grid after adjusting the frequency [15]. Thus, wind energy is vital in meeting sustainable energy needs, reaching various applications from households to industrial facilities [16].

Wind energy fluctuates depending on meteorological variables [8]. Traditional statistical forecasting methods often fail to model the complex relationships between these variables accurately and may not provide sufficiently accurate predictions [17]. These limitations become more apparent in complex systems like energy markets with dynamic and multidimensional datasets. Traditional approaches typically rely on linear assumptions and fixed parameters, which cannot fully reflect the nonlinear interactions and time-dependent changes between variables [18].

Traditional forecasting methods include physical modeling techniques, statistical analysis methods, and time series models; however, their limitations in accurately modeling the complex variables in the atmosphere have increased the need for new and more advanced approaches [19], [20]. Today, artificial intelligence (AI) and machine learning techniques are increasingly used for more precise evaluation of meteorological data [21]. Deep learning algorithms, big data analytics, and hybrid forecasting models improve the predictability of wind energy production by modeling atmospheric variables more accurately [22]. In this regard, the detailed analysis of weather parameters and their integration into forecasting models significantly improves operational efficiency in the renewable energy sector. Machine learning algorithms, a subset of AI, can process large energy datasets and make more accurate predictions [23]. Therefore, data-driven models in wind power forecasting help make energy production processes more efficient and reliable [24].

Machine learning algorithms enable predicting future events by analyzing large datasets [25]. Used across many sectors, these algorithms are particularly significant for more efficient management of renewable resources in the energy sector [26]. In wind power forecasting, machine learning methods assist in identifying the relationships between variables affecting energy production by processing large datasets [27]. The most commonly used methods for wind power forecasting include Random Forest (RF), k -nearest Neighbors (k NN), gradient-boosted trees (GBT), and Artificial Neural Networks (ANN) [28].

Studies on machine learning methods in wind power forecasting aim to determine the most suitable methods by comparing the forecasting performances of different algorithms [29]. In wind power forecasting, the use of Genetic Algorithm (GA) has achieved a prediction accuracy of 98% by analyzing

SCADA data and NASA meteorological data collected [30]. In a similar study, Support Vector Regression (SVR), Gaussian Process Regression (GPR), and Decision Trees (DT) algorithms were used for short-term wind power forecasting, with the DT algorithm providing a prediction accuracy of $R^2=0.92$ [31]. In a previous study, Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) models were compared, and the highest accuracy for wind power forecasting was achieved with the LSTM model [32]. In another study, the effect of the placement of wind turbines within the plant on energy production was analyzed, and plant data were evaluated using Random Forest (RF) and Gradient Boosting Descent (GBD) models, and it was found that optimal turbine placement could increase energy production [33].

This study consists of four main sections. The first section provides a literature review on the subject and methodology of the study. The second section includes information about the methods used. The third section discusses the analysis and forecasting results obtained from the proposed methods. Finally, the last section presents the general structure of the study and its contributions to the literature.

2. Materials and Methods

This study aims to estimate wind energy production using hourly meteorological data from 2018 in Yalova province. The data set used includes weather parameters. These meteorological variables were included in the model as independent variables (input), while the amount of wind energy production was used as the dependent variable (output). In the pre-processing process of the data, missing and outlier values were detected and corrected with appropriate statistical techniques. Then, all variables were normalized and prepared for the modeling phase. This preparation process was implemented to increase the accuracy of the models and strengthen the learning processes. The variables considered in the study are provided in **Table 1**.

Table 1. The features of the dependent and independent variables

Variable	Types	Unit	Data Type	Data Source
Energy Output	Dependent	kW	Continuous	kaggle
Wind Speed	Independent	m/sec	Continuous	Open-Meteo
Wind Direction	Independent	(°)	Continuous	Open-Meteo
Temperature	Independent	°C	Continuous	Open-Meteo
Humidity	Independent	RH	Continuous	Open-Meteo
Dew Point 2m	Independent	°C	Continuous	Open-Meteo
Pressure MSL	Independent	hPa	Continuous	Open-Meteo
Cloud Cover	Independent	%	Continuous	Open-Meteo
Wind Gusts	Independent	m/s	Continuous	Open-Meteo
Precipitation	Independent	mm	Continuous	Open-Meteo

The primary data source used in this study is the Wind Turbine SCADA Dataset, obtained from the Kaggle platform [34]. This dataset includes operational data on wind turbines. In addition, Open API) where other datasets are provided, is a weather data source known for its reliability and widespread use [35]. However, there may be deviations in the data supplied by Open-Meteo API. This is because API includes model predictions and meteorological measurements; sometimes, deviations from real-time values may be observed. The following code was used to retrieve data from the Open-Meteo API Archive for this study:

```
import requests
import pandas as pd
import numpy as np
```

```
def open_meteo_to_full_climate_data_excel(lat, lon,
start_date, end_date, output_file):
    """
```

Retrieves climate data from the Open-Meteo API for the specified location (Yalova) and date range,
performs interpolation at 10-minute intervals,
and saves the data to an Excel file.
This function fetches the following climate parameters:

<pre> - temperature_2m, relative_humidity_2m, dew_point_2m, apparent_temperature, pressure_msl, cloud_cover, wind_speed_10m, wind_direction_10m, wind_gusts_10m, precipitation. """ base_url = "https://archive-api.open- meteo.com/v1/archive" params = { "latitude": lat, "longitude": lon, "start_date": start_date, "end_date": end_date, "temperature_unit": "celsius", "wind_speed_unit": "kmh", "timezone": "Europe/Istanbul", "hourly": ", ".join(["temperature_2m", # Temperature at 2m height "relative_humidity_2m", # Relative humidity at 2m height "dew_point_2m", # Dew point at 2m height "apparent_temperature", # Apparent temperature "pressure_msl", # Mean sea level pressure "cloud_cover", # Cloud cover "wind_speed_10m", # Wind speed at 10m height "wind_direction_10m", # Wind direction at 10m height "wind_gusts_10m", # Wind gusts at 10m height "precipitation" # Precipitation amount]), } try: # Fetch data from the Open-Meteo API response = requests.get(base_url, params=params) response.raise_for_status() </pre>	<pre> data = response.json() if "hourly" in data: hourly_df = pd.DataFrame(data["hourly"]) hourly_df["time"] = pd.to_datetime(hourly_df["time"]) # Interpolate the data to 10-minute intervals hourly_df.set_index("time", inplace=True) # Resample hourly data to 10-minute intervals with linear interpolation ten_minute_df = hourly_df.resample('10min').interpolate(method='li near') # Save the data to an Excel file ten_minute_df.to_excel(output_file, index=True) print(f"Data has been successfully written to {output_file}.") else: print("Hourly data not found.") except requests.exceptions.RequestException as e: print(f"Error during API request: {e}") except Exception as e: print(f"An unexpected error occurred: {e}") # Yalova coordinates yalova_lat = 40.6559 yalova_lon = 29.2715 # Define the date range for 2018 start_date = "2018-01-01" end_date = "2018-12-31" output_file = "yalova_2018_full_climate_data_10min.xlsx" # Call the function to fetch and save the data open_meteo_to_full_climate_data_excel(yalova_lat, yalova_lon, start_date, end_date, output_file) </pre>
---	---

In this study, ML algorithms, Random Forest (RF), Gradient Boosting (GB), *k*-nearest neighbor (kNN), and Stochastic Gradient Descent (SGD), were used to obtain the amount of energy produced from estimated wind energy. These algorithms were run in the Orange 3.34 Data Mining program to get

performance and prediction data. The workflow diagram of the methodology of the study is given in **Figure 1**.

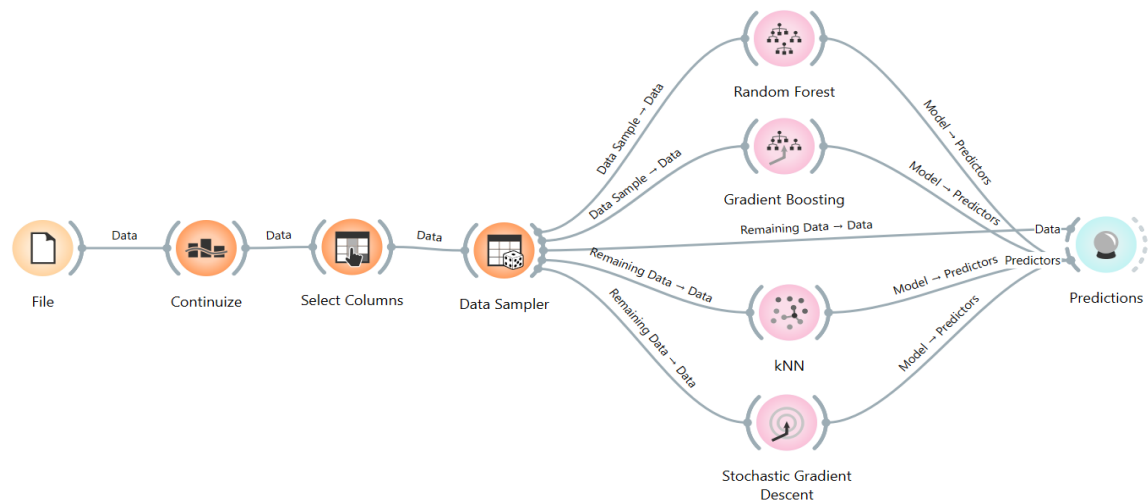


Figure 1. The flowchart of the method

Random Forest (RF) is an ensemble learning algorithm combining multiple decision trees. Each decision tree is trained based on a different subset of the dataset, and the results obtained are combined to increase the prediction accuracy [36]. The basic operating principle of the RF algorithm is based on training each decision tree independently by selecting random samples from the dataset and determining random subsets of features [37]. In this process, each decision tree produces its prediction. At the same time, the final result is determined by two different methods: majority voting in classification problems, and the final decision is formed by averaging the predictions in regression problems [38].

The Gradient Boosting (GB) algorithm is an ensemble learning method that focuses on obtaining more accurate predictions by minimizing error rates in the learning process [39]. The GB model aims to create a strong prediction model by sequentially training weak learners (usually decision trees) and performing optimizations to reduce the error rates of previous models at each stage [40]. Each new model improves the learning process by taking into account the errors made by the previous model more and increases the prediction accuracy.

k-nearest neighbor (k-NN) is an unsupervised learning algorithm that estimates a data point's class or value by looking at its nearest neighbors [41]. This algorithm is generally used in classification and regression problems. To determine the class or value of a data point, k-NN examines the labels or values of the nearest neighbors in the specified k number and usually estimates according to the class or average of these neighbors [42]. The user determines the k value, and it is generally necessary to select the most appropriate value using cross-validation. Although k-NN is widely preferred with its simple structure and understandability, the computational cost can be high in large data sets because distance calculations must be made over the entire data set for each estimate [43].

Stochastic gradient descent (SGD) is a widely used optimization algorithm in machine learning [44]. This method updates parameters using only the gradient of one data point at each iteration, to minimize the model's error function (usually the loss function) [45]. While the traditional gradient descent algorithm updates parameters by computing the gradient over the entire training data, SGD calculates only one randomly selected data point at a time. This provides a faster and lighter optimization process but can sometimes experience more fluctuations due to drift and noise. SGD is often preferred in large datasets and deep learning models because each iteration is faster and puts less pressure on the local memory [46]. However, it is important to choose the right learning rate and carefully monitor the parameter updates, as incorrect settings can negatively affect the success of the model.

To compare the performances of the ML algorithms used in the study, the root mean squared error (RMSE), the mean squared error (MSE), the mean absolute error (MAE), and the mean absolute

percentage error (MAPE) criteria, which express the error coefficients, as well as the R^2 values were calculated [47, 48]. The following formulas were used to calculate the data belonging to the performance criteria.

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

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

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

The formulas presented are common error metrics used in regression analysis to evaluate the performance of predictive models. In these equations, y_i represents the actual (observed) values, $\hat{y}_i$ represents the predicted values, $\bar{y}$ denotes the mean of the actual values, and n is the total number of observations.

3. Results

In this study, wind energy data obtained by considering the effects of environmental factors belonging to Yalova province were estimated with ML algorithms. Descriptive statistics of dependent and independent variables constituting the data set of the study are given in **Table 1**.

Table 1. Descriptive statistics of variables

Variable	N	Mean	StDev	Variance	Min	Max	Skew	Kurt
Wind Speed (m/s)	50530	75,580	42,272,0	178,689,0	0.00000	252,060	0.62	0.06
Wind Direction (°)	50530	123.69	93.4400	8731.730	0.00000	360.000	0.70	-0.75
Energy Produced (MW)	50530	13,077	13,125,0	17,225,00	-0.00247	36,1870	0.60	-1.16
Temperature 2m (°C)	50530	16,551	7,143,00	51,016,00	-0.20000	33,6000	0.01	-0.91
Relative Humidity 2m (%)	50530	75,736	13,699,0	187,657,0	23,0000	100,000	-0.69	-0.13
Dew Point 2m (°C)	50530	11,921	6,204,00	38,484,00	-5,80000	23,6000	-0.32	-0.78
Pressure MSL (hPa)	50530	1014.4	6.40000	41.00000	990.300	1033.80	0.20	-0.02
Cloud Cover (%)	50530	51,094	39,042,0	1,524,277	0.00000	100,000	-0.01	-1.64
Wind Gusts 10m (m/s)	50530	22,851	12,055,0	145,316,0	1,80000	79,6000	0.71	0.03
Precipitation (mm)	50530	0.08851	0.33285	0.110790	0.00000	710,000	7.47	84.84

Descriptive statistics for meteorological variables presented in **Table 1** reveal the environmental factors affecting wind energy production in Yalova province in detail. Wind speed and direction variables constitute the essential inputs of wind energy production (MW). The average wind speed is 75.58 m/s and varies between a minimum of zero and a maximum of 252.06 m/s. The high standard deviation (42.27) and variance (178.689) reveal that wind speed varies greatly. The average wind direction is 123.69° and is observed to differ slightly from the normal distribution with a kurtosis value of -0.75. The MW variable reflects the fluctuation in energy production and varies between -0.002 and 36.18 MW. In

addition, meteorological variables such as temperature (Temperature 2m), relative humidity (Relative Humidity 2m), dew point (Dew Point 2m), and atmospheric pressure (Pressure MSL) can also be related to wind energy production. While the average temperature is 16.55°C, the dew point is calculated as 11.92°C. The pressure variable shows a relatively low variability, averaging around 1014.4 hPa. The cloud cover variable varies between 0 and 100 and has a wide distribution due to the high standard deviation (39.04). Wind gusts (10m) have an average of 22.85 m/s and can reach a maximum value of 79.6 m/s. The precipitation amount variable varies between 0 and 710 mm, while the high kurtosis value (84.84) shows that the data distribution contains sharp and extreme values. These data provide an essential basis for analyzing how wind energy production is affected by atmospheric conditions.

Linear regression analysis was performed to measure the statistical effect and strength of the independent variables on the dependent variable. In this study, linear regression analysis was applied to measure the statistical impact and strength of the effect of various independent variables (Temperature, Humidity, Wind Speed, etc.) on a dependent variable. The ANOVA table (**Table 2**) of the linear regression analysis is presented to evaluate how much of the variance each independent variable explains on the dependent variable and whether this explanation is statistically significant. ANOVA is a method used to assess the mean differences between groups statistically. In **Table 2**, the degrees of freedom (DF), corrected sum of squares (Adj SS), corrected mean squares (Adj MS), F value, and P value of each independent variable are presented. While P values show the statistical significance of the effect of independent variables on the dependent variable, F values and Adj SS values show the effect size. **Figure 2** shows the power of influence of independent variables on the dependent variable. The variable with the most impact on the dependent variable has been determined as Wind Speed (m/s).

Table 2. ANOVA analysis

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Temperature at 2m (°C)	1	13.4000	13.4000	47.3900	0.001
Relative Humidity at 2m (%)	1	8.30000	8.30000	29.3000	0.000
Dew Point at 2m (°C)	1	12.6000	12.6000	44.5800	0.001
Mean Sea Level Pressure (hPa)	1	62.6000	62.6000	221.310	0.001
Cloud Cover (%)	1	1.80000	1.80000	6.46000	0.011
Wind Gusts at 10m (m/s)	1	94.6000	94.6000	334.400	0.002
Precipitation (mm)	1	13.2000	13.2000	46.6500	0.003
Wind Speed (m/s)	1	72157.9	72157.9	255115	0.002
Wind Direction (°)	1	4.90000	4.90000	17.4500	0.001

ANOVA analysis statistically reveals the effect of various meteorological variables on a specific dependent variable. All analyzed variables were statistically significant, as all p-values were less than 0.05. In particular, variables such as temperature (p=0.001), relative humidity (p=0.000), dew point (p=0.001), sea level pressure (p=0.001), and wind direction (p=0.001) all have high F-values and strongly reveal their effects with very low p-values. For example, sea level pressure stands out with its adjusted mean square value of 62.6 and F-value of 221.31. At the same time, wind direction has a relatively lower effect (F=17.45) but still has a statistically significant effect. This situation reveals that many components related to weather conditions directly and significantly affect the dependent variable in the system.

As a result of the analysis, the wind speed variable stands out with a corrected mean square value of 72,157.9 and an extraordinarily high F-value of 255,115, making it the most effective factor on the dependent variable. This is followed by wind intensity with a corrected mean square value of 94.6 and an F-value of 334.4. In addition, factors such as precipitation amount (F=46.65, p=0.003) and cloudiness rate (F=6.46, p=0.011) also have significant effects despite lower F-values. This shows that the system is affected by thermodynamic and atmospheric variables and that complex interactions between meteorological parameters should be considered. In conclusion, this ANOVA table powerfully reveals the effect of meteorological variables at individual and comparative levels and provides important clues about which factors should be prioritized in modeling.

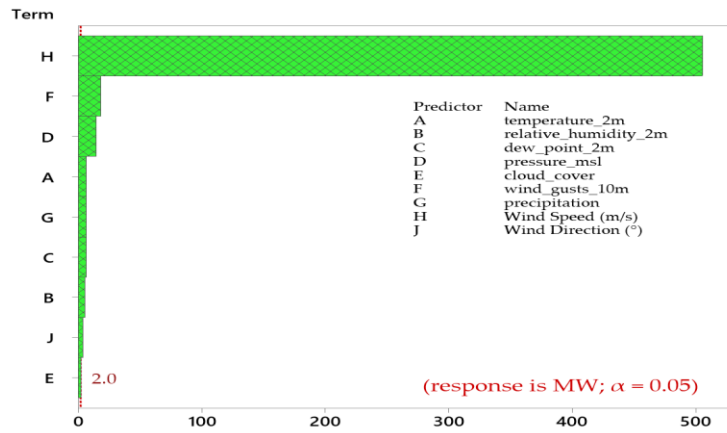


Figure 2. Pareto chart of standardized effects of independent variables

The Pareto chart presented in **Figure 2** shows the standardized effects of the independent variables on the dependent variable (MW). According to the chart, the Wind Speed (m/s) (H) variable has the most potent effect on the dependent variable, while the impact of the Cloud Cover (E) variable is relatively low. This contradicts the claim in the text that the "Cloud Cover" variable is the most effective. The chart also shows that the variables Precipitation (F), Sea Level Pressure (D), Temperature at 2m (A), and Dew Point at 2m (G) have significant effects on the dependent variable. The effects of the variables Relative Humidity (B) at 2m and Wind Direction ($^{\circ}$) (J) are at the lowest level. The expression $\alpha = 0.05$ in the chart indicates the level of statistical significance. This is a threshold value used to evaluate whether the length of the bars is statistically significant. If the bar length exceeds this threshold value, the effect of the relevant independent variable is statistically significant. This analysis shows that the variables with the most substantial impact on the dependent variable are Wind Speed and Rainfall. The performance measurement criteria of ML algorithms, RMSE, MSE, MAE, MAPE, and R^2 data, are given in **Table 3**.

Table 3. Data on performance measurement criteria of ML algorithms

Model	MSE	RMSE	MAE	MAPE	R^2
RF	0.039	0.197	0.081	0.377	0.961
GB	0.071	0.266	0.117	0.421	0.928
kNN	0.039	0.198	0.111	0.689	0.960
SGD	0.175	0.418	0.303	0.581	0.822

Table 3 presents the data for RMSE, MSE, MAE, MAPE, and R^2 metrics used to evaluate the performance of machine learning algorithms. According to the results in the table, the Random Forests (RF) algorithm shows the best performance with the lowest MSE (0.039), RMSE (0.197), and MAE (0.081) values. This indicates that the RF algorithm makes fewer prediction errors and produces results closer to the actual values. In addition, the R^2 value of RF is relatively high (0.961), indicating that the model can explain a large portion of the variance in the data. Although the kNN algorithm has similar MSE (0.039) and RMSE (0.198) values to RF, its MAE value is higher than RF (0.111). The Gradient Boosting (GB) algorithm has higher MSE (0.071), RMSE (0.266), and MAE (0.117) values than RF and K-NN, indicating that GB performs worse than the other two algorithms.

SGD algorithm has the highest MSE (0.175), RMSE (0.418), and MAE (0.303) values compared to the other algorithms, indicating that SGD performs the worst. In addition, SGD's R^2 value is the lowest (0.822) compared to the other algorithms, indicating that the model cannot explain the variance in the data. The results in **Table 3** reveal that RF and K-NN algorithms perform better than the other algorithms for this dataset. However, which algorithm is most suitable may vary depending on the specific requirements and priorities of the application. For example, when the MAPE value is taken into account, it is seen that RF (0.377) performs better than GB (0.421). This means that the percentage of errors for RF is lower than that of GB.

4. Conclusion

This study aims to increase predictability in renewable energy production by investigating the effectiveness of machine learning (ML) algorithms in wind energy forecasting in Yalova province. Comparisons between different ML models revealed that Random Forest (RF) and Gradient Boosting algorithms provide high accuracy rates. The effect of meteorological variables (e.g., temperature, humidity, pressure, and wind speed) used in the study on wind power production was analyzed in detail, and it was seen that wind speed was the most dominant parameter in forecast performance. It was shown that even in regions with moderate wind potential, such as Yalova, satisfactory forecast results can be achieved with appropriate modeling and data analysis. These findings emphasize the importance of environmental variables and data-driven approaches in energy production planning. As a result, this study constitutes an example that will contribute to renewable energy management at the local level and shows the applicability of similar methodologies in different geographical regions. The results prove that wind energy forecasts should be addressed from a multi-dimensional perspective, not only with technical data but also by considering environmental and statistical components. In future studies, it may be possible to increase the prediction accuracy by using longer data sets and integrating more advanced ML and deep learning methods. In addition, integrating these models in energy supply-demand management with real-time systems will play a strategic role in implementing sustainable energy policies.

Author Contributions: The conceptualization of the study was carried out by A.A. and Y.A.A. Data analysis and methodology development were conducted by L.A.G. and H.K. The original draft was prepared by A.A. and Y.A.A. Review, editing, and final revisions were performed by İ.Ç. and Y.A.A. All authors have approved the final version of the manuscript and share equal responsibility for the content.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

- [1] N. Abas, A. Kalair, and N. Khan, "Review of fossil fuels and future energy technologies," *Futures*, vol. 69, pp. 31–49, 2015.
- [2] P. Wilkinson, K. R. Smith, M. Joffe, and A. Haines, "A global perspective on energy: health effects and injustices," *Lancet*, vol. 370, no. 9591, pp. 965–978, 2007.
- [3] O. Ellabban, H. Abu-Rub, and F. Blaabjerg, "Renewable energy resources: Current status, future prospects and their enabling technology," *Renew. Sustain. energy Rev.*, vol. 39, pp. 748–764, 2014.
- [4] A. Rahman, O. Farrok, and M. M. Haque, "Environmental impact of renewable energy source based electrical power plants: Solar, wind, hydroelectric, biomass, geothermal, tidal, ocean, and osmotic," *Renew. Sustain. energy Rev.*, vol. 161, p. 112279, 2022.
- [5] A. Atalan and Y. A. Atalan, "Nonlinear Optimization Models of Box-Behnken Experimental Design: Turbine Simulation for Wind Power Plant," *3rd International Conference on Engineering and Applied Natural Sciences*, 2023.
- [6] A. D. Şahin, "Progress and recent trends in wind energy," *Prog. energy Combust. Sci.*, vol. 30, no. 5, pp. 501–543, 2004.
- [7] E. Toklu, "Overview of potential and utilization of renewable energy sources in Turkey," *Renew. Energy*, vol. 50, pp. 456–463, 2013, doi: 10.1016/j.renene.2012.06.035.
- [8] F. Cassola and M. Burlando, "Wind speed and wind energy forecast through Kalman filtering of Numerical Weather Prediction model output," *Appl. Energy*, vol. 99, pp. 154–166, 2012.
- [9] I. Sanchez, "Short-term prediction of wind energy production," *Int. J. Forecast.*, vol. 22, no. 1, pp. 43–56, 2006.
- [10] S.-Q. Dotse, I. Larbi, A. M. Limantol, and L. C. De Silva, "A review of the application of hybrid machine learning models to improve rainfall prediction," *Model. Earth Syst. Environ.*, vol. 10, no. 1, pp. 19–44, 2024.
- [11] A. Alkesaiberi, F. Harrou, and Y. Sun, "Efficient wind power prediction using machine learning methods: A comparative study," *Energies*, vol. 15, no. 7, p. 2327, 2022.
- [12] D. Wang, X. Gao, K. Meng, J. Qiu, L. L. Lai, and S. Gao, "Utilisation of kinetic energy from wind turbine for grid connections: a review paper," *IET Renew. Power Gener.*, vol. 12, no. 6, pp. 615–624, 2018.
- [13] J. W. Zhou, W. Zhang, X. Jiang, and E. D. Zhai, "Investigation on dynamics of rotating wind turbine blade using transferred differential transformation method," *Renew. Energy*, vol. 188, pp. 96–113, 2022.
- [14] C.-T. Chen, R. A. Islam, and S. Priya, "Electric energy generator," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 53, no. 3, pp. 656–661, 2006.
- [15] E. Serban, M. Ordóñez, and C. Pondiche, "Voltage and frequency grid support strategies beyond standards," *IEEE Trans. power Electron.*, vol. 32, no. 1, pp. 298–309, 2016.
- [16] J. B. Welch and A. Venkateswaran, "The dual sustainability of wind energy," *Renew. Sustain. Energy Rev.*, vol. 13, no. 5, pp. 1121–1126, 2009.
- [17] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," *PLoS One*, vol. 13, no. 3, p. e0194889, 2018.

- [18] P. Reichert and J. Mieleitner, "Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters," *Water Resour. Res.*, vol. 45, no. 10, 2009.
- [19] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, "Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects," *Atmos. Environ.*, vol. 60, pp. 656–676, 2012.
- [20] A. Atalan, "The ChatGPT application on quality management: a comprehensive review," *J. Manag. Anal.*, pp. 1–31, May 2025, doi: 10.1080/23270012.2025.2484225.
- [21] S. Dewitte, J. P. Cornelis, R. Müller, and A. Munteanu, "Artificial intelligence revolutionises weather forecast, climate monitoring and decadal prediction," *Remote Sens.*, vol. 13, no. 16, p. 3209, 2021.
- [22] J. Devaraj, R. Madurai Elavarasan, G. M. Shafiullah, T. Jamal, and I. Khan, "A holistic review on energy forecasting using big data and deep learning models," *Int. J. energy Res.*, vol. 45, no. 9, pp. 13489–13530, 2021.
- [23] Y. A. Atalan and A. Atalan, "Integration of the Machine Learning Algorithms and I-MR Statistical Process Control for Solar Energy," *Sustain.*, vol. 15, no. 18, Sep. 2023, doi: 10.3390/su151813782.
- [24] E. T. Renani, M. F. M. Elias, and N. A. Rahim, "Using data-driven approach for wind power prediction: A comparative study," *Energy Convers. Manag.*, vol. 118, pp. 193–203, 2016.
- [25] A. Atalan and C. Ç. Dönmez, "Dynamic Price Application to Prevent Financial Losses to Hospitals Based on Machine Learning Algorithms," *Healthcare*, vol. 12, no. 13, p. 1272, Jun. 2024, doi: 10.3390/healthcare12131272.
- [26] H. Szczepaniuk and E. K. Szczepaniuk, "Applications of artificial intelligence algorithms in the energy sector," *Energies*, vol. 16, no. 1, p. 347, 2022.
- [27] H. Demolli, A. S. Dokuz, A. Ecemis, and M. Gokcek, "Wind power forecasting based on daily wind speed data using machine learning algorithms," *Energy Convers. Manag.*, vol. 198, p. 111823, 2019, doi: <https://doi.org/10.1016/j.enconman.2019.111823>.
- [28] P. Piotrowski, D. Baczyński, M. Kopyt, and T. Gulczyński, "Advanced ensemble methods using machine learning and deep learning for one-day-ahead forecasts of electric energy production in wind farms," *Energies*, vol. 15, no. 4, p. 1252, 2022.
- [29] L. Wang, X. Zhou, X. Zhu, Z. Dong, and W. Guo, "Estimation of biomass in wheat using random forest regression algorithm and remote sensing data," *Crop J.*, vol. 4, no. 3, pp. 212–219, Jun. 2016, doi: 10.1016/j.cj.2016.01.008.
- [30] P. Moshtaghi, N. Hajialigol, and B. Rafiei, "A Comprehensive Review of Artificial Intelligence Applications in Wind Energy Power Generation," *Available SSRN 5061006*, 2024.
- [31] K. Yazıcı, "Makine öğrenmesi yöntemleri kullanılarak kısa dönem rüzgar gücü tahmini." Sakarya Üniversitesi, 2021.
- [32] Y. Qin *et al.*, "Hybrid forecasting model based on long short term memory network and deep learning neural network for wind signal," *Appl. Energy*, vol. 236, pp. 262–272, 2019.
- [33] U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, "A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments," *Energies*, vol. 14, no. 16, p. 5196, 2021.
- [34] B. Erisen, "Wind Turbine Scada Dataset 2018 Scada Data of a Wind Turbine in Turkey," *Kaggle*, 2018. <https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset>
- [35] Open-meteo, "Accurate Weather Forecasts for Any Location," 2025. <https://archive-api.open-meteo.com/v1/archive>
- [36] A. Atalan, "Forecasting drinking milk price based on economic, social, and environmental factors using machine learning algorithms," *Agribusiness*, vol. 39, no. 1, pp. 214–241, Jan. 2023, doi: 10.1002/agr.21773.
- [37] D. Thakur and S. Biswas, "Permutation importance based modified guided regularized random forest in human activity recognition with smartphone," *Eng. Appl. Artif. Intell.*, vol. 129, p. 107681, 2024, doi: <https://doi.org/10.1016/j.engappai.2023.107681>.
- [38] H. İnaç, Y. E. Ayören, A. Atalan, and C. Ç. Dönmez, "Estimation of Postal Service Delivery Time and Energy Cost with E-Scooter by Machine Learning Algorithms," *Appl. Sci.*, vol. 12, no. 23, p. 12266, Nov. 2022, doi: 10.3390/app122312266.
- [39] W. Khan, S. Walker, and W. Zeiler, "Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach," *Energy*, vol. 240, p. 122812, 2022, doi: <https://doi.org/10.1016/j.energy.2021.122812>.
- [40] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman, "Boosting and additive trees," *Elem. Stat. Learn. data mining, inference, Predict.*, pp. 337–387, 2009.
- [41] G. McKenzie and D. Romm, "Measuring urban regional similarity through mobility signatures," *Comput. Environ. Urban Syst.*, vol. 89, p. 101684, Sep. 2021, doi: 10.1016/j.compenvurbsys.2021.101684.
- [42] Z. Liu, Q. Pan, and J. Dezert, "A new belief-based K-nearest neighbor classification method," *Pattern Recognit.*, vol. 46, no. 3, pp. 834–844, 2013.
- [43] J. Huang *et al.*, "Cross-validation based K nearest neighbor imputation for software quality datasets: an empirical study," *J. Syst. Softw.*, vol. 132, pp. 226–252, 2017.
- [44] Y. Tian, Y. Zhang, and H. Zhang, "Recent advances in stochastic gradient descent in deep learning," *Mathematics*, vol. 11, no. 3, p. 682, 2023.
- [45] J. Yang and G. Yang, "Modified convolutional neural network based on dropout and the stochastic gradient descent optimizer," *Algorithms*, vol. 11, no. 3, p. 28, 2018.
- [46] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
- [47] A. Keskin, "Türkiye enerji piyasasında piyasa takas fiyatı tahmini: Makine öğrenimi yöntemlerinin karşılaştırılması," *Üçüncü Sektör Sosyal Ekonomi Dergisi*, vol. 60, no. 2, pp. 1707–1719, 2025, doi: 10.63556/tisej.2025.1509.
- [48] Y. Ayaz Atalan and A. Atalan, "Testing the Wind Energy Data Based on Environmental Factors Predicted by Machine Learning with Analysis of Variance," *Appl. Sci.*, vol. 15, no. 1, p. 241, Dec. 2024, doi: 10.3390/app15010241.