Day ahead forecasting for solar and wind electricity production using machine learning techniques

Andrej Angelovski and Aleksandra Dedinec
Student at the Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
andrej.angelovski@students.finki.ukim.mk, aleksandra.kanevche@finki.ukim.mk

***Abstract*—This paper explores the forecasting of renewable energy power output, specifically wind power in Bogdanci and solar farms in the Republic of N. Macedonia, with the end goal of achieving a day ahead forecast. Accurate forecasting of renewable energy is critical for reliable and efficient energy generation, making this study important for the energy industry and policymakers. The study uses historical data from MEPSO for the last 4 years (since 2020) and employs various statistical and machine learning techniques, including autocorrelation function (ACF), partial autocorrelation function (PACF), periodogram, linear regression, decision tree regressor, random forest regressor, support vector machine, XGBRegressor, Lasso, and Ridge, to predict, or rather forecast power output. Results indicate that accurate forecasting can be achieved using these methods, with potential implications for the adoption of renewable energy sources. The models were evaluated using mean squared error, mean absolute error, and r2 score.**

I. Introduction

Renewable energy has gained significant attention in recent years due to the increasing global demand for electricity and the negative environmental impacts of non-renewable energy sources. However, renewable energy sources such as wind and solar are intermittent in nature, and their power output can be affected by several factors, such as weather conditions and seasonal changes. Accurate forecasting of renewable energy power output is essential for reliable and efficient energy generation, making this study important for the energy industry and policymakers.

II. Materials and Methods

In this section, we are going to explain all the materials and methods employed for this paper.

A. Data Preprocessing

The data used in this study was obtained from MEPSO, and it consists of the planned and actual power output of wind and from Bogdanci wind power plant and all the solar power plants in the Republic of Macedonia. The data spans the last 4 years (since 2020) with hourly observations. The data was preprocessed by performing feature scaling by normalization with the usage of Standard Scaler. Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation.

B. Statistical Analysis

1) ACF (Auto-Correlation Function): The ACF (Auto-Correlation Function) measures the linear relationship between a time series and its lagged values.

PACF (Partial Auto-Correlation Function) The PACF (Partial Auto-Correlation Function) measures the correlation between a time series and a lagged version of itself after removing the effects of the intermediate lags.

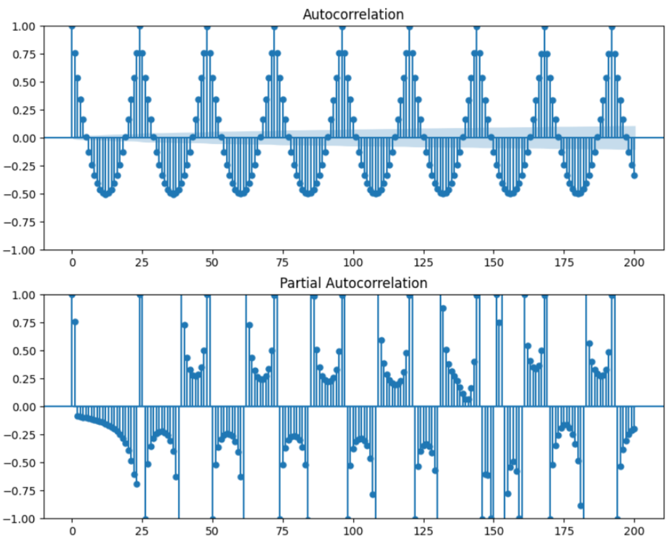

Fig. 1. ACF and PACF for the dataset.

From the statistical analysis we can conclude that the solar electricity generation is highly correlated to the same variable but 24 hours before, as well as 48 hours before (1)

2) Seasonal Decomposition: The seasonal decomposition is a method used in time series analysis to represent a time series as a sum (or, sometimes, a product) of three components - the linear trend, the periodic (seasonal) component, and random residuals.

From our dataset, yearly patterns can be noticed for the solar electricity generation, as shown in 2, which are not so visible for the wind electricity generation (3).

3) SHAP: (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local expla-

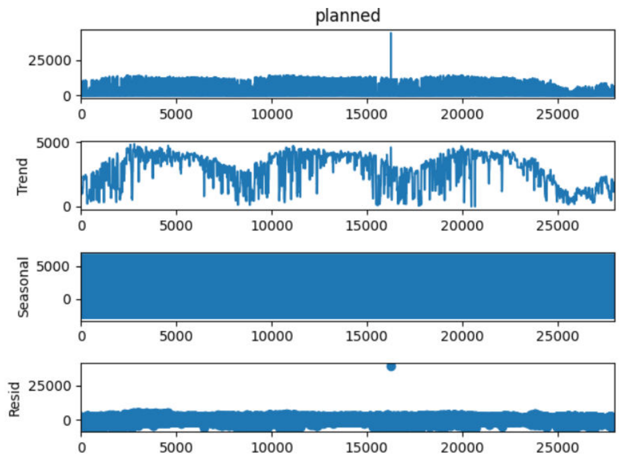

Fig. 2. Seasonal Decompose for solar power.

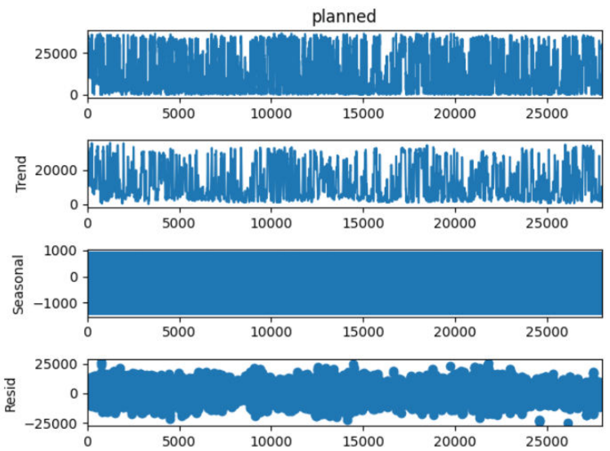

Fig. 3. Seasonal Decompose for wind power.

nations using the classic Shapley values from game theory and their related extensions. It assigns each feature an importance score for a particular prediction, based on the concept of Shapley values from cooperative game theory. In essence, SHAP allocates the 'credit' for a prediction among the features used in the model, providing a measure of each feature's contribution to the prediction. This allows for better understanding of which features are driving the model's predictions on both a global (overall) and local (per prediction) level. In the 4 and 5 the SHAP values for some of the features are given: hour, the average production of the previous 24 hours, hourly values for the previous day, and hourly values for the day two days prior.

4) Visualizations: Additionally, we used some common visualisations in order to have better view of the data. For the heat map we have added the features: the average production of the previous 24 hours, hourly values for the previous day, and hourly values for the day two days prior. The correlation shown in 6, confirms the high influence of the added features to the target value.

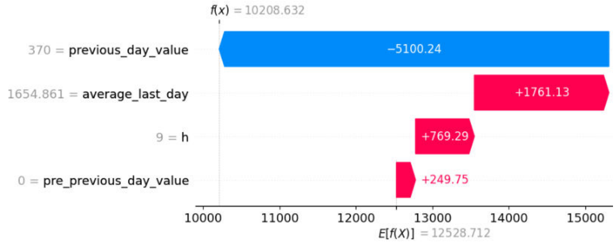

Fig. 4. SHAP for the wind dataset.

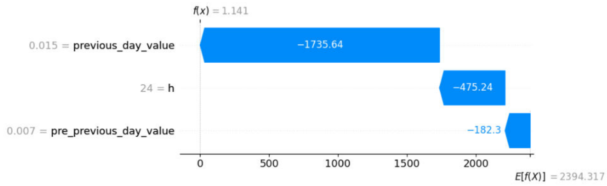

Fig. 5. SHAP for the solar dataset.

5) Input Parameters: Based on the statistical analysis, the conclusion is that input features are: day, month, year, hour, the average production of the previous 24 hours, hourly values for the previous day, and hourly values for the day two days prior.

C. Machine Learning Techniques

In addition to the statistical methods discussed above, we also employed several machine learning models to forecast the power output of the wind and solar plants.

1) Decision Tree Regressor: The decision tree regressor is a non-parametric model that can be used for regression problems. It works by recursively partitioning the data into

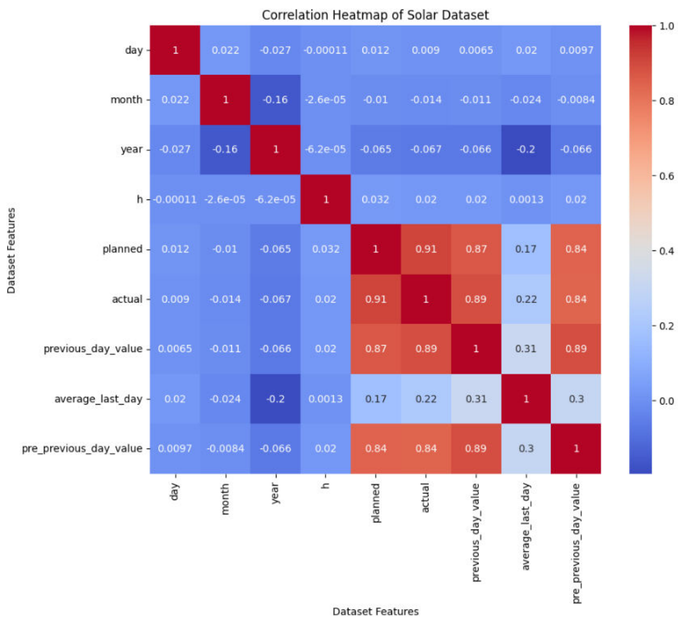

Fig. 6. Correlation of the parameters for the solar dataset.

subsets based on the values of the features, and then fitting a simple model to each subset.

2) Random Forest Regressor: The random forest regressor is an ensemble model that combines multiple decision tree regressors to improve performance and reduce overfitting. It works by building a large number of decision trees on random subsets of the data, and then averaging their predictions. In our implementation, we used a random forest regressor with 1000 trees.

3) Support Vector Machine: The support vector machine is a popular model for classification and regression tasks. It works by finding the hyperplane that maximally separates the data into different classes or predicts the output values for the regression problem. In our implementation, we used a support vector machine with a radial basis function kernel.

4) XGBRegressor: XGBRegressor is a popular gradient boosting machine learning model, that works by iteratively adding new models that attempt to correct the errors of the previous model, resulting in a strong ensemble model. In our implementation, we used the XGBRegressor with a maximum depth of 5, learning rate of 0.1 and 1000 estimators.

5) Lasso Regression: Lasso regression is a linear model that performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. In our implementation, we used the Lasso model with alpha=0.1.

6) Ridge Regression: Ridge regression is also a linear model that performs L2 regularization, which adds a penalty equal to the square of the magnitude of coefficients. In our implementation, we used the Ridge model with alpha=0.1.

7) Linear Regression: Linear regression is a statistical method used to study the relationship between two continuous variables, where one variable is considered as the dependent variable, and the other is the independent variable. It assumes that there is a linear relationship between the two variables, and tries to model this relationship using a linear equation.

D. Evaluation Metrics

Next, we evaluated the performance of the different models using the mean squared error (MSE), mean absolute error (MAE), and R-squared (R2) metrics. These metrics help to quantify the accuracy and goodness of fit of the models.

1) Mean Squared Error (MSE): MSE measures the average squared difference between the predicted and actual values, and is given by:

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

where n is the number of samples, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

2) Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted and actual values, and is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

3) R-squared (R^2): R-squared measures the proportion of variance in the dependent variable that is predictable from the independent variables.

III. Results and Future Work

Our analysis explored the data of wind and solar electricity generation. Day ahead forecast could be achieved using a combination of time series analysis and machine learning techniques. With additional fine-tuning and using additional input features such as wind speed, temperature, solar irradiance; implemented into a deep learning architecture our models can be used to generate reliable forecasts for renewable energy power output. The results show that best performances are achieved using XGBoost regressor, as shown in the Table 1 and Table 2. Additionally, the results obtained for the solar electricity generation are much more better than for wind electricity generation, which leads to the result that these models are suitable of the first profblem, but for wind electricity generation more deep architecture will be needed and probably additional input features.

TABLE I
Results of models for Wind Power

Model Used	Evaluation Method		
	MSE	***MAE***	R^2
Decision Tree Regressor	136042445.58	7436.92	0.2378
Random Forest Regressor	65608994.715	5739.1	0.63
Support Vector Machine	186259541.679	10289.59	-0.0434
XGBRegressor	46984009	4790.226	0.736
Linear Regression	77102133.57	6285.9	0.57
Lasso Regression	77366828.16	6272.93	0.566
Ridge Regression	77366838.3	6272.93	0.566

TABLE II
Results of models for Solar Power

Model Used	Evaluation Method		
	MSE	***MAE***	R^2
Decision Tree Regressor	3067838.58	787.63	0.773
Random Forest Regressor	1631601.77	585.3	0.879
Support Vector Machine	4249920.318	1078.28	0.685
XGBRegressor	1225085.225	524.539	0.909
Linear Regression	1790427.67	620	0.86
Lasso Regression	1875596.31	629.728	0.86
Ridge Regression	1875595.929	629.73	0.86

Future work involves further developing a deep learning framework for wind power forecasting, which leverages both Deep Deterministic Policy Gradient (DDPG) and Long Short-Term Memory (LSTM) for wind power output forecasting (as shown in 7. The LSTM trains on historical data, generating day-ahead predictions that feed into the DDPG agent. This agent formulates optimal controls for wind turbine blade and rotor adjustments, taking into account weather forecasts and day-ahead power predictions. The process optimizes the energy management system, schedules optimal wind power generation, and reduces energy costs through dynamic adjustments of wind turbine components.

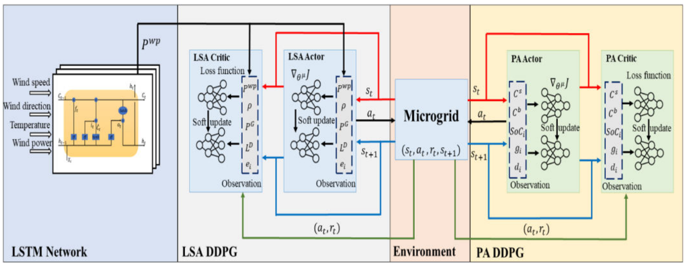

Fig. 7. Deep Learning Model - combining Deep Deterministic Policy Gradient (DDPG) and Long Short-Term Memory (LSTM) methods to optimize wind power forecasting for wind power.

IV. Acknowledgment

I would like to express my gratitude to the Macedonian Academy of Sciences and Arts for the assistance for this research project. Additionally, special gratidute towards MEPSO for providing their datasets for solar and wind electricity production.

References

[1] A. Ghasemi, A. Shojaeighadikolaei, and M. Hashemi, ”Combating Uncertainties in Wind and Distributed PV Energy Sources Using Integrated Reinforcement Learning and Time-Series Forecasting,” Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA. arXiv:2302.14094v1 [eess.SY] 27 Feb 2023

[2] A. R. Fuxjager, K. Kozak, M. Dorfer, P. M. Blies, and M. Wasserer, ”Reinforcement Learning Based Power Grid Day-Ahead Planning and AI-Assisted Control,” enliteAI, Vienna, A-1010, Feb. 15, 2023. arXiv:2302.07654v1 [cs.AI].

[3] F. Sabbatini and C. Grimani, ”Solar Wind Speed Estimate with Machine Learning Ensemble Models for LISA,” Department of Pure and Applied Sciences, University of Urbino Carlo Bo, Via S. Chiara, 27, Urbino, 61029, Italy and Section in Florence, INFN, Via B. Rossi, Sesto Fiorentino, 50019, Florence, Italy, arXiv:2302.06740v1 [astro-ph.HE], Feb. 13, 2023.