1. Inajara RUTYNA, 2. Paweł PIOTROWSKI

Warsaw University of Technology, Electrical Power Engineering Institute
ORCID: 1. 0000-0001-8086-6231; 2. 0000-0002-1582-0912

doi:10.15199/48.2024.01.31

Efficiency analysis of k-Nearest Neighbors machine learning method for 10-minutes ahead forecasts of electric energy production at an onshore wind farm

***Abstract**. This paper presents tests of the effectiveness of the K-Nearest Neighbors (KNN) machine learning technique for short-term forecasting of energy production at an onshore wind farm with a horizon of 10 minutes. The tests were performed for several variants of input variables to KNN models (only backward variables of the forecasted time series and the use of additional exogenous input variables - meteorological data). For each of the variants, the selection of an appropriate number of k was performed using the cross-validation method, separately for each of the distance measures tested. Analyses were performed of the found k values depending on the variant of the input variables and the distance measure. Conclusions and observations of the performed tests were formulated.*

***Streszczenie.** W artykule przedstawiono testy skuteczności techniki uczenia maszynowego k najbliższych sąsiadów (K-Nearest Neighbors - KNN) do krótkoterminowego prognozowania produkcji energii na farmie wiatrowej lądowej z horyzontem 10 minut. Badania wykonano dla kilku wariantów zmiennych wejściowych do modeli KNN (tylko zmienne cofnięte prognozowanego szeregu czasowego oraz zastosowanie dodatkowych zmiennych wejściowych egzogenicznych – dane meteorologiczne). Dla każdego z wariantów wykonano dobór właściwej liczby k metodą walidacji krzyżowej, osobno dla każdej z testowanych miar odległosci. Wykonano analizy znalezionych wartości k w zależności od wariantu zmiennych wejściowych oraz miary odległości. Sformułowano wnioski i spostrzeżenia z wykonanych badań. (**Analiza efektywności metody uczenia maszynowego k-Nearest Neighbors dla prognoz produkcji energii elektrycznej z 10-minutowym wyprzedzeniem w lądowej farmie wiatrowej**)*

Keywords: wind farm, forecasting, machine learning, k-nearest neighbors.
Słowa kluczowe: farma wiatrowa, prognozowanie, uczenie maszynowe, k-najbliższych sąsiadów.

Introduction

Accurate short-term forecasts of energy production from onshore wind farms are very important for the proper operation of the power system in the processes of control, optimization and storage of electricity [1, 2]. The development of methodologies for forecasting energy production in RES is the subject of a great deal of research. Methods propose both single models [3, 4, 5, 6, 7, 8, 9, 10] as well as hybrid and ensemble methods [11, 12, 13, 14, 15]. Among the most commonly used methods are various machine learning techniques [13]. The analysis presented in this paper is based on the use of KNN models for short-term forecasts of energy production at a wind farm. This machine learning technique uses a different approach to input variables - all inputs to the model are treated as equally important, unlike other ML models where individual input variables have different importance - after training the model, the importance of individual input variables can be assessed. This treatment of inputs as equal can be both an advantage and a disadvantage, depending on the context and the nature of the data. Examples of the advantage: simplicity and robustness to outliers (since all inputs are treated equally, outliers have a relatively smaller impact on the predictions).

The legitimacy of using the KNN technique to forecast energy demand has been demonstrated, among other things, by the studies presented in [16, 17]. Meanwhile, the application of the KNN technique to wind farm energy production forecasts is described in [11, 12].

Statistical analysis of data

The data used to make the forecasts for the period of 2 full years (2021-2022) was taken from SOTAVENTO GALICIA, SA, a wind farm consisting of 24 wind turbines with a total power rating of 17 MW [18]. Three time series with 10-minute periods (measured data) were available: wind speed, wind direction and energy production in an extensive onshore wind farm. For statistical analyzes and forecasting models, in the first stage, the last six values from each of the three time series were selected as potential input variables for KNN models. The three time series were normalized to the ranges 0-1 (min-max normalization). Incorrect samples and those with numerous missing data were eliminated from the input-output data sets. In the case of single missing data gaps, data were supplemented based on neighboring values. Data outliers were replaced by the dataset using a mean plus standard deviation approach. After removing the outliers and before data normalization, a baseline statistics for the forecasted time series can be computed to gain insights into the data distribution. Basic statistics of analysed time series are presented in Tab.1.

Table 1. Basic statistics of analysed time series.

Time Series	Mean	Std	Min	Max
Wind speed (m/s)	6.25	3.57	0.00	29.48
Wind direction (º)	176.69	91.29	0.00	379.00
Energy production (kWh)	2756.24	3263.24	0.00	22552.20

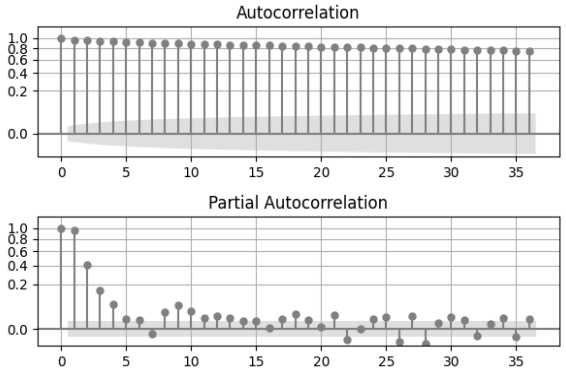

Fig.1. ACF and PACF results for energy production time series

To better understand the correlation in the energy production time series, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) are computed, considering up to 36 lags (6 hours) (see Fig. 1).

The ACF and PACF functions are able to provide significant insights about the correlation between observations at different time lags, allowing to find patterns or dependencies within the data. As results show, that even though there is a slow decay in ACF, it can be due to the fact that there is a low variation in the energy production for the given next periods, due to the higher decay on PACF correlation after 5 lags. Which means that one hour, or 6 steps, is more than enough to consider as input data for the 10-minutes ahead forecast.

In order to evaluate the impact of each input data on energy production (e(t)), a correlation matrix was generated (see Fig. 2). The correlation matrix provides a general view of the relationships between each input variable and the energy production output. Each input data is evaluated up to 6 steps back, giving the following input features: wind speed (ws(t-1), ws(t-2),…,ws(t-6)), wind direction (wd(t-1), wd(t-2),…,wd(t-6)) and energy production (e(t-1), e(t-2),…,e(t-6).

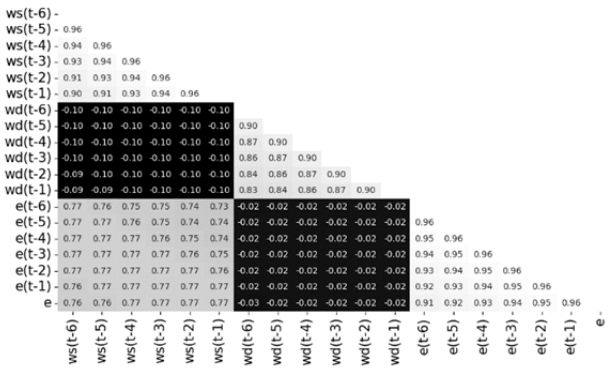

Fig.2. Correlation matrix for analysed input data and output data

Analysing the correlation matrix, it is evident that there is a clear definition of influence between the expected output e(t) and each input data. The energy production features e(t-1), e(t-2),…,e(t-6) present a strong linear correlation, varying from 0.96 to 0.91, which guarantees a direct relationship in the energy production e(t). In terms of importance, in second place is the wind speed features ws(t-1), ws(t-2),…,ws(t-6), with a correlation between 0.77 to 0.76, indicating that it is still a feature of great importance, even though less influential. On the other hand, the wind direction features wd(t-1), wd(t-2),…,wd(t-6) show a negative effect in the energy production, with correlation values ranging from -0.02 to -0.03, indicating an inverse relationship.

Forecasting methods

KNN is a non-parametric method used for classification and regression where input data consists of the k closest training examples in the feature space. In KNN regression, the expected result is the value of the feature of the object, this value is the average of the distances of the k nearest neighbors [12]. The main hyperparameter for tuning is the number of nearest neighbors k, in our model k values from 1 to 100 were tested. The distance metric is the second hyperparameter, on which 3 different metrics were tested. The Manhattan distance, also known as the city block distance, taxi cab distance, or L1 norm, is calculated by summing the absolute differences between two points of their corresponding feature values. The Euclidean distance, also known as L2 norm, is calculated as the square root of the sum of the squared differences between two points of their corresponding feature values. It is important to emphasize that distance is sensitive to the scale of the features, therefore normalization of data is recommended. The cosine distance is recommended for high-dimensional data, is based on the angle between two points their feature vectors in the multi-dimensional space. The cosine distance has a different approach as instead of considering the magnitude of the feature values, it focuses on the direction or orientation of the vectors. This distinction allows the cosine distance to be robust to differences in feature magnitudes.

Some advantages of using KNN for forecasting include: simplicity (It does not require complex mathematical calculations or parameter tuning), robustness to outliers (it can be advantageous when dealing with noisy or unpredictable data) and ability to handle multi-dimensional data (allowing it to capture complex relationships between variables and potentially identify non-linear patterns in the data). There are also some disadvantages: computationally expensive (especially when dealing with large datasets or high-dimensional feature spaces), sensitivity to feature scaling (KNN relies on distance calculations, and the choice of distance metric can be affected by the scale of the features), proper choice of k (selecting an inappropriate value of k can lead to overfitting or underfitting of the data, affecting the accuracy of the forecasts), “curse of dimensionality” (KNN's performance can deteriorate as the dimensionality of the data increases, in high-dimensional feature spaces, the notion of distance becomes less reliable, as the points tend to be equidistant from each other) and irrelevant or noisy inputs (KNN treats all inputs equally, including potentially irrelevant or noisy data points, if the dataset contains irrelevant or noisy features, they can negatively affect the prediction accuracy).

The naive method was chosen as the base method for reference comparison, as it is a simple forecasting method that assumes that the future values will be the same as the most recent observed value.

The data sets (input-output) were divided into a training and validation set (75%) and a test set (25%). The division into the 2 sets was made sequentially - every fourth sample in the time series went to the test set. Analyses were conducted for three variants of input features: V(18 inputs) - all pre-proposed input data (6 regressed values of wind speed (ws(t-1), ws(t-2),…,ws(t-6)), wind direction (wd(t-1), wd(t-2),…,wd(t-6)) and electricity production (e(t-1), e(t-2),…,e(t-6)), V(6 inputs) - input data having Pearson's linear correlation value with output data above 0.8 (6 regressed values of electricity production), and V(12 inputs) - input data having Pearson's linear correlation value with output data above 0.7 (6 regressed values of wind speed and electricity production).

Evaluation criteria

To provide a comprehensive assessment of the KNN model’s performance and generalization capabilities, the 10-fold cross-validation approach is used together with evaluation criteria such as the Coefficient of Determination (R2), normalized Root Mean Squared Error (nRMSE), and normalized Mean Absolute Error (nMAE).

The cross-validation involves dividing the dataset into 10 equal-sized folds, so that the KNN model is trained using nine folds. The training process is made 10 times, and each time one of the folds is removed from training and used as a validation set. The evaluation metrics are then computed by aggregating the results from each fold, providing a robust

assessment of the model's performance. The averaging of the evaluation metrics through multiple folds allows to access the impact of variations in the training and validation data, guaranteeing more reliable results. The choice of appropriate *k* values was based on the lowest nMAE error.

The R2 score is a commonly used metric for regression tasks, providing an estimate of the model's ability to explain the variance in the target variable, with values closer to 1 indicating a better fit. The nRMSE and nMAE are normalized metrics. It allows comparison of prediction errors across different size wind farms.

$$R2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, for\ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad (1)$$

$$nRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{C_{norm}}\right)^2} \quad (2)$$

$$nMAE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{C_{norm}}\right| \quad (3)$$

where: C_{norm} is the normalization coefficient (the difference between the maximum and minimum values of the time series), $\hat{y}_i$ is predicted value, y_i is the actual value and *N* is the number of prediction points.

Results

To identify the best KNN model, a total of 900 models were trained. This involved using 3 different sets of input data, 3 different distance metrics, testing the *k* parameter from 1 to 100, and applying 10-fold cross-validation for enhanced reliability in selecting the appropriate *k* parameter.

In Table 2, a summary of the results of training the KNN models is presented, including three prediction error measures, as well as the results for the reference model - the naive model. The best result for each error measure is highlighted in bold in Table 2. Evaluating the obtained results, the smallest nRMSE error for the KNN model is 20.97% lower than the nRMSE error of the reference method - the naive model. Additionally, the smallest nMAE error for the KNN model is 7.41% lower than the nMAE error of the naive model. However, the KNN model with the application of the Cosine distance metric performed significantly worse, when compared to the naive model.

Table 2. Summary of prediction results for different variations of KNN models and the naive method

Method	Distance metric	Input data variant	nMAE (10-fold cross-validation)	nRMSE (test)	nMAE (test)	R2 Score (test)
KNN (*k*=21)	Manhattan	V(6 inputs)	**0.0103**	**0.0325**	0.0104	**0.9497**
KNN (*k*=39)	Cosine		0.0694	0.0991	0.0692	0.5314
KNN (*k*=18)	Euclidean		0.0107	0.0328	**0.0103**	0.9486
KNN (*k*=19)	Manhattan	V(12 inputs)	0.0130	0.0335	0.0125	0.9464
KNN (*k*=16)	Cosine		0.0343	0.0596	0.0313	0.8306
KNN (*k*=15)	Euclidean		0.0139	0.0340	0.0131	0.9448
KNN (*k*=18)	Manhattan	V(18 inputs)	0.0211	0.0351	0.0153	0.9411
KNN (*k*=14)	Cosine		0.0372	0.0474	0.0243	0.8928
KNN (*k*=19)	Euclidean		0.0248	0.0369	0.0177	0.9352
Naive	-	e(t-1)	-	0.0411	0.0111	0.9195

The best KNN model is given by the parameter *k*=21, using 6 input features and employing the Manhattan distance metric (lowest nRMSE error and highest R2 score) (see Fig.3). However, when considering nMAE error as the evaluation criteria for model selection, the best KNN model is the one with a parameter *k*=18, using 6 input features, and employing the Euclidean distance metric. It is worth noting that the differences in results for both models are very small (the same quality class for both models). The Figure 4 shows the scatter plot of the actual energy production values and the values obtained from the forecast with the naive model.

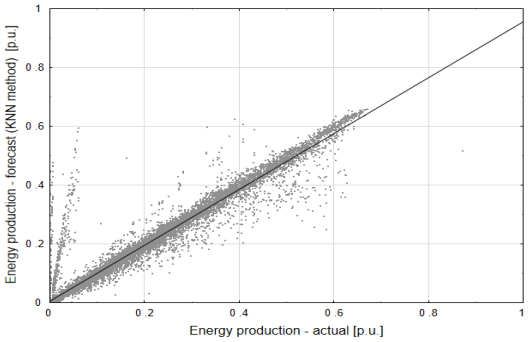

Fig.3. The scatter plot of the actual energy production values and the values obtained from the forecast with the best KNN model (k=21, Manhattan distance metric, V(6 inputs) (test data)

The selected value of the parameter *k* using cross-validation varies for 3 different distance metrics as well as for 3 different sets of input features. In the case where the Manhattan and Cosine distance metrics are considered, the selected value of *k* decreases as the number of input features increases (see Fig.5 and Fig.6). However, for the Manhattan distance metric, the changes in the value of *k* are negligible.

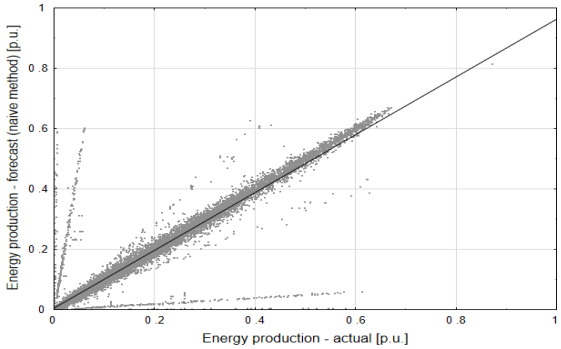

Fig.4. The scatter plot of the actual energy production values and the values obtained from the forecast with the naive model (test data)

Summary and conclusions

The conducted research has demonstrated that the KNN method can be applied to forecast energy production in a wind farm with a 10-minute horizon, using lagged values of the predicted time series and meteorological data measurements as input data. The inherent characteristic of the KNN method, where all input features are treated equally, resulted in the best outcomes when employing 6 input features. These input features exhibit the highest

linear correlation with the output data. Undoubtedly, this poses a limitation of the KNN method. Typically, in machine learning methods, the more input features that have a statistically significant correlation with the output data, the higher the quality of the forecasts. There are variations of the KNN method that consider weights for the input features values, and the authors of the article plan to test such algorithms. Based on the obtained results, it can be concluded that the appropriate distance metric for this forecasting problem is Manhattan. The Mean Bias Error (MBE) for the best KNN model is very close to zero (0.000163). Therefore, this model does not exhibit a tendency to overestimate or underestimate forecasted values.

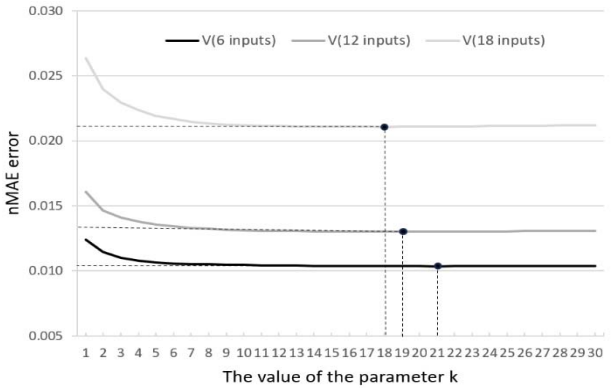

Fig.5. The relationship between the optimal value of the parameter *k* and the number of input features for the KNN model utilizing the Manhattan distance metric (10-fold cross-validation – nMAE error)

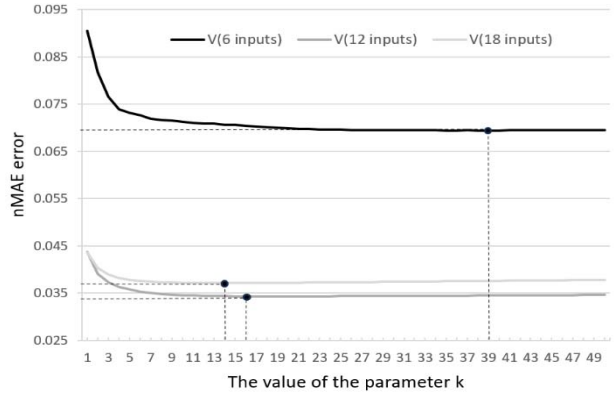

Fig.6. The relationship between the optimal value of the parameter *k* and the number of input features for the KNN model utilizing the Cosine distance metric (10-fold cross-validation – nMAE error)

As the number of input features increases (12 and then 18), the prediction errors also increase for the KNN method when utilizing Manhattan and Euclidean distance metrics. Conversely, the opposite phenomenon was observed for the Cosine distance metric - as the number of input features increased, the prediction errors decreased (Fig.6). However, they were still significantly worse compared to the prediction errors of the naive method.

In the case of dividing the results into three groups (6, 12, and 18 input features), the highest number of smallest prediction errors was achieved using the KNN method with the Manhattan distance metric. Therefore, the Manhattan distance metric can be recommended for this type of forecasting. A drawback of the KNN method is undoubtedly the long training time, especially when using cross-validation to determine the proper value of the *k* parameter. The conducted research also indicates that the KNN method is the most effective for forecasting when only the input data with a high correlation to the output data is used. The use of input data with low correlation to the output data leads to a decrease in the quality of the KNN predictions. Therefore, in cases where it is possible to use meteorological forecasts (such as wind speed) as input data for the prediction period (which have a high correlation with the input data), the KNN method would achieve a significant reduction in prediction errors.

***Authors**: mgr Inajara Rutyna, Warsaw University of Technology, Electrical Power Engineering Institute, Koszykowa 75, 00-662 Warsaw, Poland, E-mail: inajara.rutyna.dokt@pw.edu.pl; dr hab. inż. Paweł Piotrowski (corresponding author), Warsaw University of Technology, Electrical Power Engineering Institute, Koszykowa 75, 00-662 Warsaw, Poland, E-mail: pawel.piotrowski@pw.edu.pl.*

REFERENCES

[1] Baczyński D., Parol M., Piotrowski P., Współczesne problemy prognozowania w elektroenergetyce. Zagadnienia wybrane, Pod redakcją naukową Mirosława Parola, *OWPW*, (2020)

[2] Popławski T., Problematyka prognoz generacji wiatrowej w KSE, *Przegląd Elektrotechniczny*, 90 (2014), nr 7, 119-122

[3] Piotrowski P., Rutyna I., Baczyński D., Kopyt M., Evaluation Metrics for Wind Power Forecasts: A Comprehensive Review and Statistical Analysis of Errors, *Energies* 15(24) (2022), 1-38

[4] Piotrowski P., Baczyński D., Kopyt M., Szafranek K., Helt P., Gulczyński T., Analysis of forecasted meteorological data (NWP) for efficient spatial forecasting of wind power generation, *Electric Power Systems Research*, 175 (2019), 1-9

[5] Piotrowski P., Baczyński D., Prognozowanie dobowej produkcji energii elektrycznej przez turbinę wiatrową z horyzontem 1 doby, *Przegląd Elektrotechniczny*, 90 (2014), nr 9, 113-117

[6] Piotrowski P., Analiza statystyczna danych mających wpływ na produkcję energii elektrycznej przez farmę wiatrową oraz przykładowe prognozy krótkoterminowe, *Przegląd Elektrotechniczny*, 88 (2012), nr 3a, 161-164

[7] Popławski T., Wężgowiec M., Implementacja informatyczna modelu trendu pełzającego do prognozowania mocy farm wiatrowych, *Przegląd Elektrotechniczny,* 93 (2017), nr 2, 246-249

[8] Popławski T.,Dąsal K., Łyp J., Szeląg P., Zastosowanie modeli ARMA do przewidywania mocy i energii pozyskiwanej z wiatru, *Polityka Energetyczna*, T.13 z.2, 2010, 385-400

[9] Gholamreza M., Farshid K., A new short-term wind speed forecasting method based on fine-tuned LSTM neural network and optimal input sets, *Energy Conversion and Management*, 2013 (2020), 1-15

[10] Kejun W., Xiaoxia Q., Hongda L., Jiakang S., Deep belief network based k-means cluster approach for short-term wind power forecasting, *Energy*, 165 (2018), 840-852

[11] Piotrowski P., Baczyński D., Kopyt M, Gulczyński T., Advanced Ensemble Methods Using Machine Learning and Deep Learning for One-Day-Ahead Forecasts of Electric Energy Production in Wind Farms, *Energies* 15(4) (2022), 1-30

[12] Piotrowski P., Kopyt M., Baczyński D., Robak S., Gulczyński T., Hybrid and Ensemble Methods of Two Days Ahead Forecasts of Electric Energy Production in a Small Wind Turbine, *Energies* 14(5) (2021), 1-25

[13] Kopyt M., Power Flow Forecasts: A Status Quo Review. Part 1: RES Generation Prediction, *Przegląd Elektrotechniczny*, 96 (2020), no. 11, 1-4

[14] Oveis A., Mohamed L., Mehdi B., Behrouz S., Miadreza S., Improved EMD-Based Complex Prediction Model for Wind Power Forecasting, *IEEE Transaction on Sustainable Energy*, 11 (2020), no. 4, 2790-2802

[15] Hao Y., Zuhong O., Shengquan H., Anbo M., A cascaded deep learning wind power prediction approach based on a two layer of mode decomposition, Energy, 189 (2019), 1-11

[16] Dudek G., Janicki M., Nearest Neighbour Model with Weather Inputs for Pattern-based Electricity Demand Forecasting, *Przegląd Elektrotechniczny*, 93 (2017), no. 3, 7-10

[17] Dudek G., Pełka P., Prognozowanie miesięcznego zapotrzebowania na energię elektryczną metodą k najbliższych sąsiadów, *Przegląd Elektrotechniczny,* 93 (2017), nr 4, 62-65

[18] https://www.sotaventogalicia.com/en/technical-area/real-time-data/historical/ (Accessed 2023-02-01)