ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
SCHOOL OF ECONOMIC SCIENCES

ΤΜΗΜΑ ΔΙΕΘΝΩΝ & ΕΥΡΩΠΑΙΚΩΝ ΟΙΚΟΝΟΜΙΚΩΝ ΣΠΟΥΔΩΝ
DEPARTMENT OF INTERNATIONAL & EUROPEAN ECONOMIC STUDIES

Undergraduate Student: **Agiopoulos Kaptsikas, 4150037**
Supervisor Professor: **Dr. Phoebe Koundouri**

Evaluation of the Benefits from a Machine Learning Project for Wind Energy Prediction

Athens, September 2020

Abstract

The already apparent, and exponentially deteriorating, consequences of Climate Change, dictate that the decarbonatization of the world economy be realized the sooner possible. To that end, several steps towards the integration of clean energy, i.e. renewable sources, have been realized. Also, these steps include the harnessing of the available wind resources, by means of wind turbines, which are directly linked to the energy grid, providing green energy to the whole economy, and thus, contributing to the alleviation of the negative effects that Greenhouse Effect imposes to the society, economy, and the environment. However, wind energy has some drawbacks that mostly emanate from its volatile nature, which does not foster the ultimate assimilation of it to the energy market. Positively, due to recent technological advances in the field of information technology and computer science, have emerged lucrative avenues that could solve this persistent problem. This paper will try to give answers to two main questions. First, can machine learning be used in the wind energy sector, in order to successfully predict future hourly wind energy production of wind farms? Second, are there any benefits for the economy, and the environment, in case such a technology be exploited for the aforementioned purpose? If so, how could those be evaluated in monetary terms? The results of this study are rather encouraging. As the findings of the first part of the exercise indicate, machine learning can be used for short-term prediction of hourly wind energy production. Lastly, regarding the benefits of the new technology for the environment and the economy, these exist, and their monetary equivalents are quite noticeable.

Keywords: Wind Energy Generation Forecasting, Machine Learning, Monetarized Benefits, Evaluation, Nord Pool, Wind Energy Producers, Environment, Climate Change

Acknowledgments

This dissertation was conducted, during my fifth, and final, academic year as an undergraduate student of the Department of International and European Economic Studies (IEES) of Athens University of Economics and Business (AUEB). This inquiry is nothing more than a product of hard work, and faith to the Department and University, whose commitment to excellence, has been guiding students towards success, over the years. Also, I want to thank all of my professors, who have helped, taught, and shaped me, and my personality and background, as a student. Specially, I want to acknowledge the contribution of my supervisor, professor Dr. Phoebe Koundouri, for her useful help and instructions. Lastly, I also want to wholeheartedly thank my fellow student, Constantine Bardis, for being a loyal friend. This undergraduate dissertation is wholly dedicated to my one and only beloved mother country, Hellas.

List of Contents

1 Introduction

In the wake of the numerous devastating consequences of global warming, which refers to the long-term raising of the planet's temperature levels, due to human activities that first started during the industrial period, humanity has to find a promising solution. What causes temperature levels to rise, is the global economy's excessive reliance on fossil fuels, whose combustion is emitting great volumes of greenhouse gasses, which in turn get elevated to the atmosphere's higher levels and absorb the sun's radiation, thus increasing the Earth's temperature levels. Interestingly, many centuries have already passed since humans had first resorted to the harnessing of renewable sources of energy (RES), with the first happening in Europe around 200 BC. Contrastingly, the first commercial wind turbine was sold way later in USA at 1927, when the first official steps of the wind energy industry also took place. It is historically evident, that humanity has always been returning to nature for solving its most challenging problems. However, as RES are naturally volatile, meaning they cannot be accurately predicted, another remedy has to be found, so that they are exploited at full extend.

According to the related literature, machine learning can serve as a lucrative workaround to the problem of wind energy production's lack of predictability. There are many studies, indicating the aptitude of machine learning methods in very short-term to short-term predicting wind power production. Generally, very short-term predictions, are used for operating on the energy grid in real time (Chang, 2014) and having better technical control over the wind farms (Colak et al., 2012), while short-term forecasts serve as a tool for scheduling future load dispatches, thus ensuring the economic viability of wind energy. For example, Chaudhary et al. (2020), by a comparison of two machine learning algorithms, specifically of a Random Forest (RF) and a Support Vector Machine (SVM), have proven the aptitude of both of them to successfully deploy accurate short-term wind power forecasts, with the former being better from the latter one. Furthermore, Zendehboudi et al. (2018), have also shown the supremacy of machine learning algorithms, especially of a hybrid SVM, as a "an effective and precise modelling approach", for short-term wind power forecasting, in comparison with other conventional models. Moreover, Treiber et al. (2016), showed that when tuning its hyperparameters, the SVM algorithm can yield up to a 24% reduction in prediction error, in comparison with the persistence model, when used to create very short-term predictions of wind power production.

Similarly, machine learning can also be used in establishing medium-term to long-term wind energy predictions, yet the corresponding scientific literature is relatively lesser. Also, the medium-term and long-term predictions of wind energy have a wide range of applications. For example, they are mostly fostering managerial issues, such as maintenance of wind farms, and decision making, relating to energy reserves and wind turbines' energy commitments (Chang, 2014). For instance, Barbosa et al. (2017), created an ensemble, or a hybrid, of learning models, which were an ARIMA and two Neural Networks (NN), and showed that wind speed, and therefore wind energy generation, could be fairly, with the lowest prediction errors, predicted in every forecasting horizon, ranging from ultra-short to long-term. In addition, Barbounis et al. (2007), used three Recurrent Neural Networks (RNN) for a 72-hour ahead wind power prediction, and found that the RNN structure can outperform the persistence model by 50%, while the same number can be subject to improvements if more data are passed into the constructed models. Furthermore, Catalao et al. (2009), created a "three-layered feedforward ANN trained by the Levenberg-Marquardt algorithm", which could long-term forecast with relatively lower error than the persistent and ARIMA model, the future wind energy generation. Lastly, another application of machine learning for generating medium-term to long-term forecasts of wind speed and power production are presented by Cadenas and Rivera (2010), who used a hybrid model, consisting of an Artificial Neural Network (ANN) and a traditional ARIMA model, to forecast wind speed from three different regions, up and to 2 days ahead into the future. Conclusively, it is evident, from the respective scientific literature of the field, that deep learning algorithms, namely neural networks etc, have been more extensively used for medium and long-term forecasts, than other machine learning models, such as Random Forests and Support Vector Machines.

As the current study will encompass only machine learning models, and will not examine the suitability of deep learning for predicting future wind power generation, it would be interesting to see how short-term wind energy forecasts can benefit the energy market's functioning. More specifically, it would be helpful to evaluate those benefits, if any, in monetary terms, so that they become practically useful in decision making, both from wind energy producers, and the policy makers, who intend to maximize the social prosperity, though the mitigation of the consequences imposed by climate change. However, there are not any studies, which have broached the issue of evaluating a machine learning project, such as the one of this study, for environmental purposes. Contrastingly, some of the studies, which relate to topics of the energy markets microstructure, not only have they not incorporated such forecasts, but they have deemed them rather infeasible. More specifically, Mazzi and Pinson (2017) did so by assuming that wind resources "have a stochastic nature and can be predicted with a limited accuracy", while Bitar et al. (2014) that are "modelled as a random process". These, are some of the literate indications, which imply that there is not a greatly amassed volume of articles that study the underlying purpose of this exercise. Technically, however, Mazzi and Pinson (2017) in their paper, constructed a profit's maximization problem of the wind energy producers, which is expressed as a function of the electricity prices of the Nord Pool's market. On top of that, there exist many studies, which utilize machine learning methodologies, in order to predict these future electricity prices (see Beigaite et al., 2018; Kristiansen, 2014; Chaabane, 2014). Nevertheless, not quite many studies exist, which contemplate issues of wind power predictions in the framework of the energy market of Nord Pool.

This paper will broach, both the issue of predicting future wind energy generation with the aid of machine learning technologies, and finding potential benefits of it for the economy and the environment, which will also be converted into monetary terms, for decision making purposes. The study is divided into two concrete parts. On the one hand, the first part is allocated to the implementation of machine learning forecasts of future hourly wind energy production. Sub-partially, the first part adheres a specific sequence of content. First, the methodology of the following conducted study takes place, during which the proposed machine learning models of the comparative analysis are described, and then, the algorithmic procedure of their construction is outlined. Later, the part, concerning the collection and the pre-processing of the data is cited. Afterwards, follows a short part for the description of the evaluation metrics, or in other words, how is the predictive accuracy of the proposed models going to be measured and compared. Lastly, the empirical results of the first part are presented. First, a preliminary Exploratory Data Analysis (EDA) is performed, and right after that, are demonstrated the final results of the first part. On the other hand, the second part of this exercise, is dedicated in the evaluation of the exceeding predictive power of machine learning models, like Support Vector Regression (SVR) and Random Forest Regression (RFR), against other traditional methods, such as Linear Regression (LR) and Autoregressive Integrated Moving-Average (ARIMA). First of all, a theoretical setup is configured, so that every needed information is introduced. For instance, the framework of the Nord Pool's underlying energy market is described, and all the linked benefits, stemming from exploitation of machine learning by wind energy producers and the energy grid, are pinpointed. Then, the evaluation process is summarized, and lastly, the technical procedure and the final results are discussed.

2 Methodology

2.1 Description of Predictive Models

In this study, both statistical and machine learning models were selected. However, deep learning models, namely neural networks and their variants, have been purposely omitted from the analysis, since they serve as one of the most promising avenues for predicting wind power production, and thus, require a whole separate study, in order for their true aptitude to be unveiled. The goal of this part of the exercise is to compare the predictive accuracy of traditional with that of more advanced models. To that end, a baseline model has also been set up, which produces a naïve (myopic) forecast, and helps in determining the increment in terms of predictive accuracy that a more complex model adds to the forecasting task. Below, follows a short description of the selected predictive models.

2.1.1 Persistence Model

The persistence model is the most common reference model in wind power prediction and other tasks involving time series data. Also, it is one of the simplest models, as it only performs a myopic prediction. That is, the prediction of the next observation is always equal to the previous one. As for the study's short forecasting horizon, which has been stipulated to one hour, the persistence model is effective and difficult to outperform. Mathematically, it is outlined as:

$$\hat{y}_t = y_{t-1}, \forall\, t \in Z \qquad (1)$$

2.1.2 Linear Regression

The linear regression is one of the most basic statistical models. In particular, it establishes a linear relationship by fitting a straight line between a target and a predictive variable(s), given a historical set of observations, which describes their past behaviour. The best linear relationship, namely its intercept and coefficients, is found by minimizing the squared distance (ordinary least squares) of the line from the past observations. The estimated line is:

$$\hat{y}_t = \hat{a}_{OLS} + \sum_{i=1}^{n} \hat{b}_i^{OLS} \times x_{it}, \forall\, t, n \in Z \qquad (2)$$

2.1.3 Elastic Net Regression

Although linear regression is not a complex model, there is a probability to overfit the data, should the number of features greatly increase. Hence, to address this issue, it would be useful to incorporate a form of penalization, which will eliminate any of the model's excessive, non-beneficial complexity. The elastic net regression utilizes a form of penalization that combines both lasso (L1) and ridge (L2) regressions. What makes elastic net different from linear regression, is the alteration of the cost function, which now integrates two more penalization terms. The new cost function is formulated as follows:

$$\min_{\hat{a}_{ENR}, \hat{b}_{ENR}} \sum_{i=1}^{n} \left(y_i - \hat{a}_{ENR} - \sum_{j=1}^{q} \hat{b}_j^{ENR} \times x_{ij} \right)^2 + (\lambda_1 \times \sum_{j=1}^{q} |\hat{b}_j^{ENR}|) + (\lambda_2 \times \sum_{j=1}^{q} \hat{b}_j^{ENR^2}) \qquad (3)$$

In the cost function, above, the second and the third term, describe the L1 and L2 penalization forms, respectively. The L1 term adds the magnitude of the coefficient, as a penalty, to the cost function, while the L2 term does the same, by adding the square of the coefficient. Hence, ridge regression shrinks the coefficients, thereby alleviating issues, regarding overfitting and multicollinearity, but lasso regression can even eliminate some of the coefficients, thus, not only helping in addressing overfitting, but also serving, as a feature selection method. Conclusively, the estimated regression line is:

$$\hat{y}_t = \hat{a}_{ENR} + \sum_{i=1}^{n} \hat{b}_i^{ENR} \times x_{it}\,, \forall\, t, n \in Z \qquad (4)$$

2.1.4 Polynomial Regression

In effect, polynomial regression is a variation of the linear regression model, whose set of features is expanded, not only including the predictor variables, but also their higher powers, i.e. quadratic, cubic etc. Although the model has non-linear features, it is only assumed a “quasi” non-linear algorithm, because the coefficients of the predictor variables are still linearly calculated. That is, the algorithm cannot clearly explain the non-linear relationship, which might exist between the target and predictor variables. However, the function of the generated predictions is part of the non-linear space. The estimated polynomial regressor is:

$$\hat{y}_t = \hat{a} + \sum_{i,k}^{n,K} \hat{b}_i \times x_{it}^{k}, \forall\, n, K \in Z \qquad (5)$$

2.1.5 Autoregressive Integrated Moving Average

In contrast with the previously described models, ARIMA is explicitly designed to handle timeseries data and respect their peculiar characteristics, which mostly emanate from the consequences of the residuals’ autocorrelation. More specifically, it is a conjunction of two parts, which are the autoregressive (AR) and moving average (MA). The former consists of a linear regression of the target variable onto its lagged observations, and the latter, of the target variable onto the previous error terms. Furthermore, the integration part regards prerequisites of the algorithm, concerning the nature of the target variable, which has to be stationary, namely have an invariant statistical distribution over time. In order for a non-stationary timeseries to be converted into stationary, (non-seasonal) differences have to be applied. The minimum order of differencing that successfully converts the series into a stationary process is the integration order. Rigorously, the estimation line is:

$$\text{ARIMA}(p, d, q): \quad \Delta(\widehat{y_t, y_{t-d}}) = \hat{a} + \sum_{i=1}^{p} \gamma_i \times y_{t-i} + \sum_{i=1}^{q} \delta_i \times \varepsilon_{t-i}, \forall\, t, p, q \in Z \qquad (6)$$

2.1.6 k-Nearest Neighbours Regression

The k-NN model is a machine learning method. Particularly, through the construction of a multi-dimensional space, whose number of dimensions equals the number of predictor variables, it predicts new cases, by averaging their nearest observations. The location of those is pinpointed, by calculating their multi-dimensional euclidean distances from every other past observation, and then the nearest cases are selected. Mathematically, predictions are formed as:

$$\hat{y} = \frac{\sum_{i=1}^{k} y}{k}, \forall\, k \in Z \qquad (7)$$

2.1.7 Random Forest Regression

A random forest is an ensemble of multiple decision trees (CART). In effect, the CART (Classification and Regression Tree) algorithm is the way of how a random forest really function. Essentially, a decision tree aims to minimize the current state of disorder (entropy) in the data, by splitting it in such a way, so that the greatest reduction of the error metric is realized. Also, a decision tree consists of nodes, at which the data is split, using the predictor variables of the regression. The amount of reduction incurred by splitting the data, utilizing a certain predictive feature, is called "information gain", as it describes the information that the corresponding feature provides, concerning its ability to effectively explain the nature of the data. Moreover, every decision tree, which comprises the random forest ensemble, is trained upon a randomly generated, through a process called "bootstrap aggregation", subset of the initial data. Jointly, the concepts of "entropy" and "information gain", form the ID3 algorithm, which is quite common for building decision trees. Comparatively, random forests are more efficient than individual decision trees, since they are able to reduce variance, thus eliminating the decision trees' extended risk of overfitting, and serving, as a significantly more reliable predicting option. Lastly, when all splitting nodes have been optimally created, a final, uniform prediction for every node is calculated. More specifically:

$$\hat{y}_{\mathrm{node}} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{N_{node}}} \mathrm{y_i}}{\mathrm{N_{node}}}, \forall\ \mathrm{N_{node}} \in \mathrm{Z} \qquad (8)$$

2.1.8 Support Vector Regression

The support vector regression (SVR) is a variant of support vector machines (SVM), and it is used to solve regression, rather than classification problems, as a regular SVM does. Also, the same algorithm can handle, both linear, and non-linear data. In effect, the SVR is a linear regression method, because it intends to find a linear function that best explains a linear set of instances. However, if the data are non-linear, a kernel (see $K(x_i, x_j)$ below) function is utilized that projects them onto another, of higher dimensions, mathematical space, in which they are linearly transformed, yet without having the interrelationships between the samples eliminated. Now, since the data have been linearly transformed on the surrogate mathematical space, the best linear function (y), or hyperplane, is calculated. According to the SVR algorithm, the best linear hyperplane is the one that maximizes (maximum margin regressor) its distance from certain selected observations, which are called "support vectors" and are used to find the closest match between the samples and the hyperplane. Intuitively, when the distance among the "support vectors" and the hyperplane is maximized, the more accurate are the generated predictions by the model, due to the presence of random noise in the data. Finally, this maximized distance stipulates a range (ε), inside which any prediction errors occurred are neglected, while the ones being out of it, called "slacks" (ξ and ξ^*), are tracked and subjected to minimization. Hence, the conditional optimization problem is mathematically formed, as follows:

$$\min \frac{1}{2} \times \|\mathrm{w}\|^2 + \mathrm{C} \times \sum_{\mathrm{i=1}}^{\mathrm{n}} (\xi_{\mathrm{i}} - \xi_{\mathrm{i}}^*)$$

$$\mathrm{s.t.}$$

$$\mathrm{y_i - w \times x_i - b} \leq \varepsilon + \xi_{\mathrm{i}}$$

$$\mathrm{w \times x_i + b - y_i} \leq \varepsilon + \xi_{\mathrm{i}}^*$$

$$\rightarrow \hat{y}_{\mathrm{i}} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{w_i} \times \mathrm{K(x_i, x) + b}, \text{ where: } \mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = \mathrm{e}^{\left(-\|\mathrm{x_i - x_j}\|^2 / 2 \times \sigma^2\right)}, \forall\ \mathrm{n} \in \mathrm{Z} \qquad (9)$$

2.2 Evaluation Method

The objective of evaluation is to estimate the generalization error, which regards a model's accuracy to predict new cases based on its training, using only known data. The used evaluation methodology is a combinatory approach, since it jointly exploits the techniques of cross validation and grid search. These, are more specifically described below:

2.2.1 Grid Search

This method is used for the optimal determination (tuning) of a model's set of hyperparameters. There are two types of parameters to be tuned, during the construction of a model. These, are the regular parameters and the hyperparameters. What makes them different is that the latter have to be specified prior to the training phase, due to their inability to be estimated through the data. Essentially, hyperparameters are used to control the learning process, thereby being different from the regular parameters, which are estimated during the learning process itself. Moreover, most of the models possess tuneable hyperparameters, while there are others, which do not. In latter case, grid search can be skipped and its place occupy the training part. In effect, grid search is an exhaustive iteration over a matrix, which hosts every possible combination of the hyperparameters of a model. This process is iterative, since all combinations are used to train the model and the one yielding the best results, namely the lowest generalization error, is chosen. Furthermore, there are other tuning methods, which could be alternatively considered. However, it was not part of this study's purpose.

2.2.2 Cross-Validation

The cross-validation method is one of the most important concepts in machine learning, because, during which not only is performed the estimation of the generalization error, but also an improvement process of pinpointing it. Hence, this procedure yields the best possible generalization error of a model's predictive accuracy. According to it, the dataset is partitioned into a predefined number of equal parts, which are iteratively distributed to the training and testing sets. In every iteration, the maximum amount of data that can be occupied by the testing set is only one of the segmented folds, while the rest of them is allocated to the training set. The goal is for a number of training and testing sets' pairs to have been generated that match the number of equal parts, into which the original dataset had been divided. Then, the model is trained and evaluated on every of the aforementioned pairs, and the average of all the errors, corresponding to the different testing sets, is reported as the final generalization error.

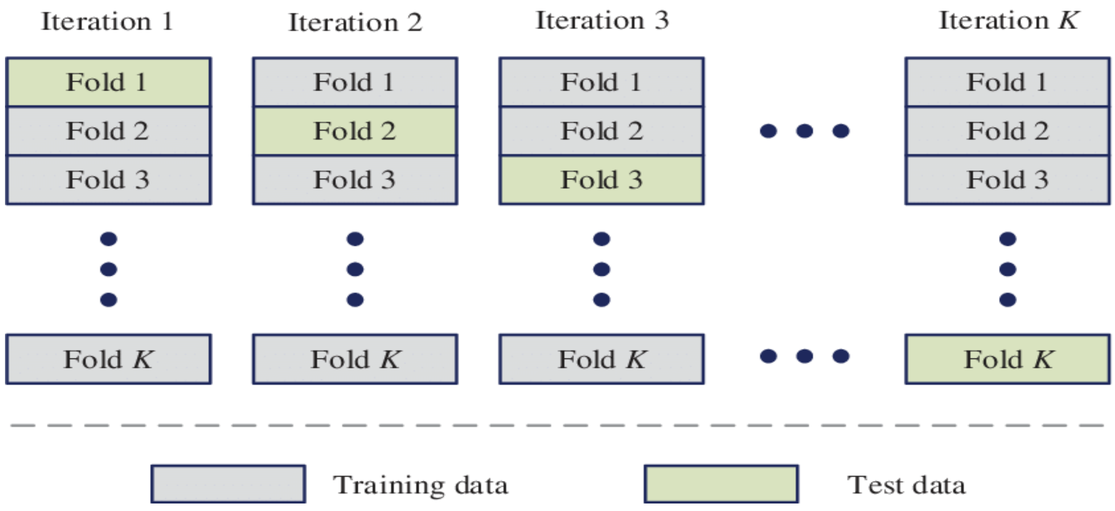


Image 1: Regular Cross-Validation, Source: Ren, Qiubing & Li, Mingchao & Han, Shuai (2019)

However, the above version of cross-validation is not designed for tasks that incorporate time series data, since the random partitioning is infringing their fundamental characteristics. That is, the randomly segmented folds do not retain the coherence of the time component, whose presence yields the autocorrelation observed between the successive data points in a time series. Hence, in order to account for these temporal dependencies in the data, another version of cross-validation has to be employed. A suitable alternative version is called “nested cross-validation” and it is different, although inheriting the rationale of its previous variant, since it does respect the presence of the dataset’s time component. More specifically, an “expanding window” approach is implemented. Initially, the window covers a certain amount of the dataset, starting from the beginning, which is split into the first pair of training and testing sets. Continuingly, the window’s size iteratively increases by the amount of data that had been assigned to the testing set of the first iteration. Note that in every iteration the size of the evaluation set remains constant. Lastly, the model is constructed and evaluated, using every iteration’s pair of sets, and the average of all the out-of-sample (testing) errors is indicated as the generalization error.

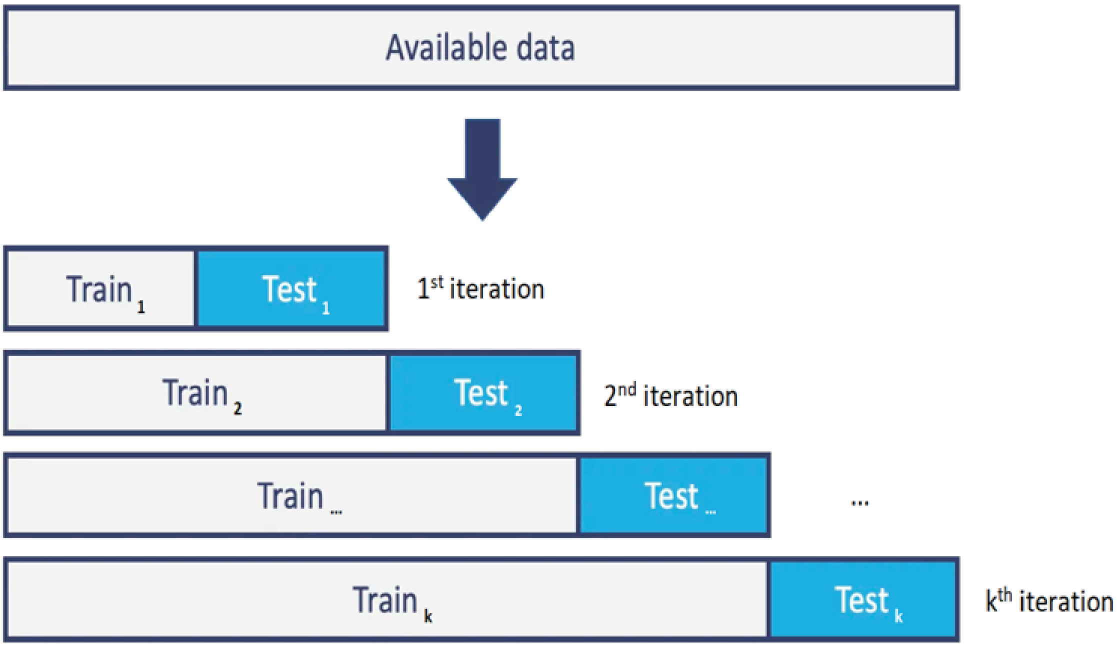
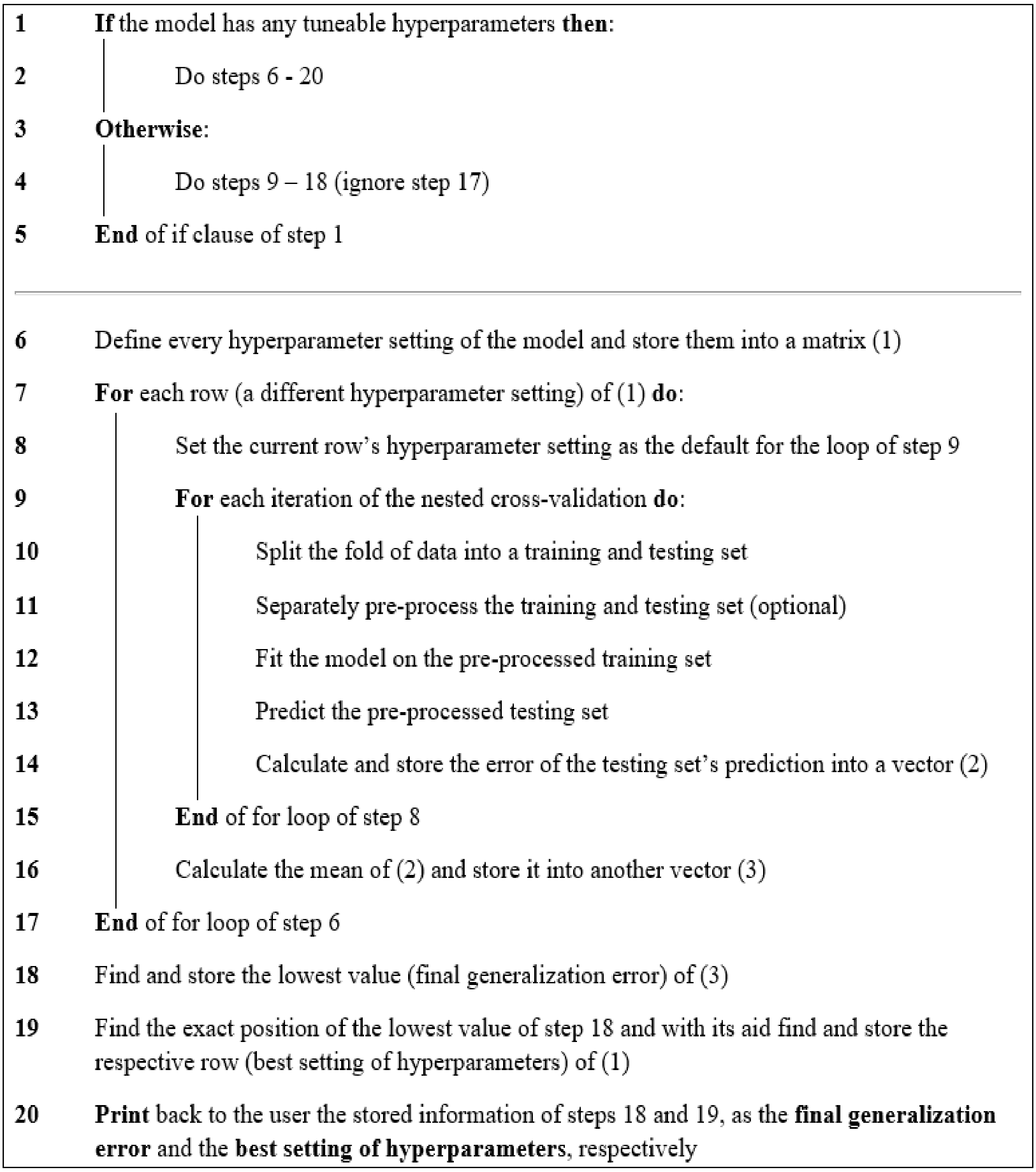


Image 2: Nested Cross-Validation for Time Series Data,
Source: https://godatadriven.com/blog/its-time-to-trust-your-predictions/

2.2.3 Nested Cross-Validation

Now that the above methods have been described, it would be wise to analyse the combinatory algorithmic methodology, which was used to create and evaluate the predictive models of this exercise. To combine the approaches of cross-validation and grid search, an algorithmic procedure had to be configured, since there is a need to simultaneously find the model’s optimal generalization error and its respective best setting of hyperparameters. In order to accomplish that, a structure of two nested for loops has been implemented. The pseudocode of the algorithm is graphically shown in the image below. The outer loop (step 7) iterates over the rows of the matrix, containing all the available hyperparameters’ combinations, thus letting the inner loop (step 9) apply the “nested cross-validation” to every one of them. Substantially, this produces a matching of all the combinations of the hyperparameters against their corresponding best generalization errors. Then, a searching for the combination with the lowest generalization error is realized and its findings are what the program prints back to the user. As for the first five steps of the pseudocode, they help in amending the algorithm, in case where there are not any hyperparameters to be tuned for the model.



```
1   If the model has any tuneable hyperparameters then:
2       Do steps 6 - 20
3   Otherwise:
4       Do steps 9 – 18 (ignore step 17)
5   End of if clause of step 1

6   Define every hyperparameter setting of the model and store them into a matrix (1)
7   For each row (a different hyperparameter setting) of (1) do:
8       Set the current row's hyperparameter setting as the default for the loop of step 9
9       For each iteration of the nested cross-validation do:
10          Split the fold of data into a training and testing set
11          Separately pre-process the training and testing set (optional)
12          Fit the model on the pre-processed training set
13          Predict the pre-processed testing set
14          Calculate and store the error of the testing set's prediction into a vector (2)
15      End of for loop of step 8
16      Calculate the mean of (2) and store it into another vector (3)
17  End of for loop of step 6
18  Find and store the lowest value (final generalization error) of (3)
19  Find the exact position of the lowest value of step 18 and with its aid find and store the
    respective row (best setting of hyperparameters) of (1)
20  Print back to the user the stored information of steps 18 and 19, as the final generalization
    error and the best setting of hyperparameters, respectively
```

Image 3: Pseudocode of the Combinatory Algorithmic

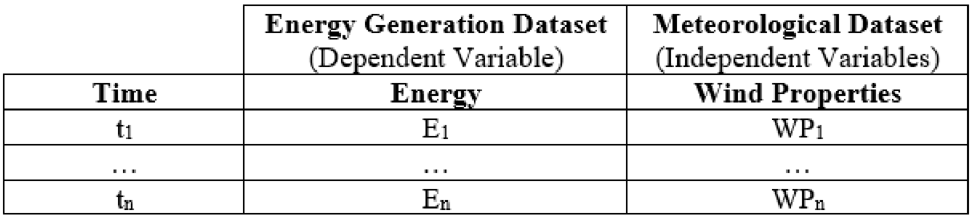
3 Empirical Results

3.1 Software

This study has been conducted using the R programming language (v4.0.2) and several of its embedded libraries, which are tailored for timeseries data. More specifically: the *dplyr* (v0.7.8) library was used for manipulating the dataset; *ggplot2* (v3.3.2) for data visualization; *stats* (v3.6.2) and *tseries* (v0.10-47) for deploying various statistical tests, and training linear models, such as linear and polynomial regression; *forecast* (v8.12) for addressing non-stationarity, implementing timeseries decomposition, and ARIMA modelling; *caret* (v6.0-86) as a baseline for writing the combinatory algorithmic methodology's code; *ranger* (v0.12.1) for training random forests; *kernlab* (v0.9-27) for training support vector machines; and lastly, *yardstick* (v0.0.6) for evaluating the predictive accuracy of the different machine learning models.

3.2 Data Description

The forecasting of wind energy output, using weather's properties, requires two basic sets of data. These, consist of a dataset of the energy generation, measured over a certain period of time, and a dataset, containing the respective, spatially and time adjusted, meteorological data. Graphically:



	Energy Generation Dataset (Dependent Variable)	Meteorological Dataset (Independent Variables)
Time	Energy	Wind Properties
t_1	E_1	WP_1
…	…	…
t_n	E_n	WP_n

Table 1: Visual Inspection of the Dataset

Regarding the collection of the data, they were retrieved from different sources, because it was not possible to find a unified set, containing every needed information for the purposes of the study. First, the energy generation dataset was retrieved from the Ontario's Independent Electricity System Operator (IESO) official website, and second, the meteorological dataset was collected from the Government of Canada's official website.

This procedure required that both sets be calibrated in the same manner, in terms of spatial and time resolution. As for spatial resolution, the datasets concerned different geographical locations, due to the absence of weather stations in the location of the wind farm. Hence, the main problem was to optimally pinpoint that combination, among the available weather stations and the wind farm, with the least spatial distance. Inherently, this inevitably propagates some error to the final forecasts, yet not to the extent that the purpose of this exercise become unreachable. Regarding time resolution, both sets were successfully collected for the same period of a year, while their frequency is one measurement per hour. Thus, the sample's total size is 8760 observations.

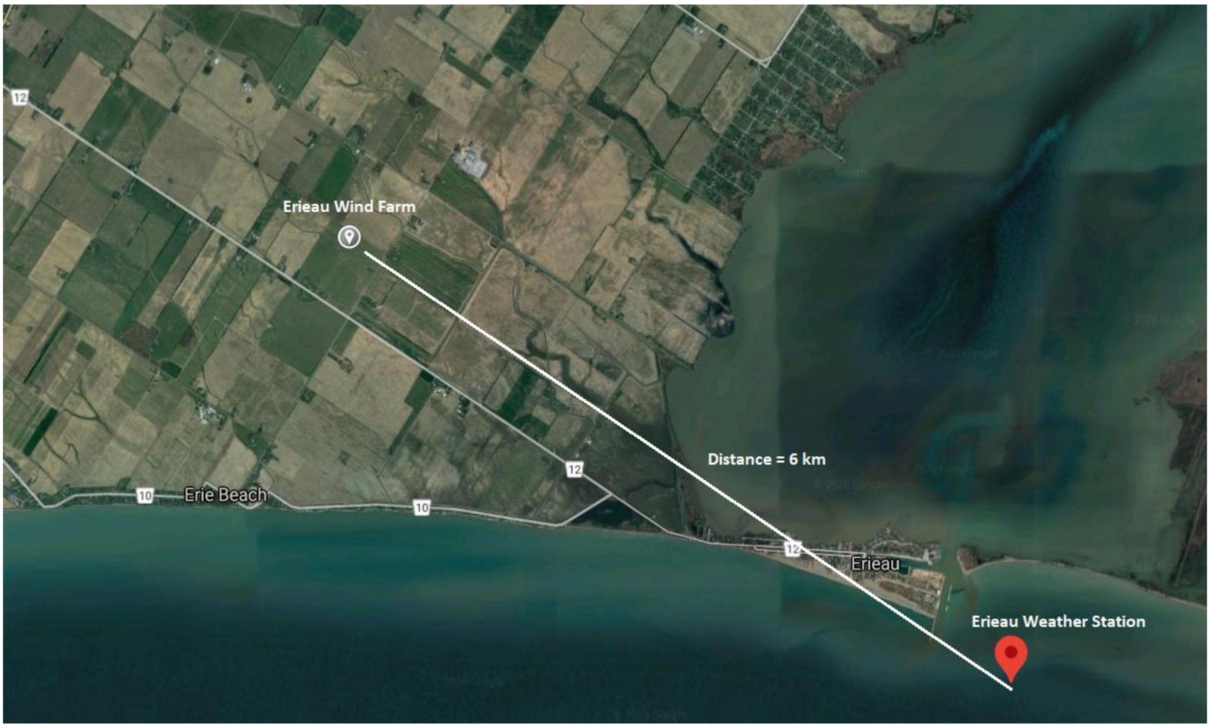


Image 4: The Location of the Wind Farm and the Closest Weather Station, Source: Google Maps

3.3 Data Preparation

The data preparation, or pre-processing, helps in converting a raw dataset into an understandable form by the machine. More specifically, some of the most fundamental parts of this technique, which are used in this study, are: *a*. data quality assessment, *b*. feature encoding, and *c*. splitting of the initial dataset into training and testing sets. Furthermore, other important, although not been used in this exercise, techniques are dimensionality reduction, feature aggregation and sampling.

3.3.1 Data Quality Assessment

In this step, a careful inspection of the initial dataset is performed, thus any particularities are spotted and treated accordingly. The most frequently observed blemishes in datasets are missing and inconsistent values, which refer to not and incorrectly assigned measurements, respectively. In this study, the latter were not the case, since the measurements were registered by accurate instruments that track down properties of the weather. However, there were some missing values in the dataset. There are two basic ways to cope with missing values, which are, either the elimination of the rows, containing the missing values, or the estimation of those values, using interpolation or filling methods. In this exercise, any encountered missing values were replaced by the mean of the column, in which they had been first found. This is a filling method and is the most commonly used among researchers.

3.3.2 Feature Encoding

This part of data preparation helps in the transformation of several types of variables, or features, so that they are ready for machine learning applications. The variables' type can be nominal, ordinal, or categorical. In this study, the variables were restricted only to the nominal type. Hence, the encoding should concentrate on feature scaling, which is how the nominal type of data is transformed, so that the range of values, among the variables, is equalized. This kind of data transformation is used, because certain models utilize Euclidian distances, and thus are dependent on the range of the features, which, if not equivalent, would deteriorate the training process. Therefore, every nominal independent variable has been standardized, using the following formula:

$$\tilde{x}_i = \frac{x_i - \bar{x}_{train}}{\sigma(x_{train})}, \forall\, i = \{train, test\} \qquad (10)$$

Note, that the training and testing sets' standardization is performed, using only the mean and standard deviation of the training set. This, accounts for the data leakage's consequences, if the precedure were performed, using statistical properties of the initial dataset.

3.3.3 Splitting into Training and Testing Sets

This is one of the most fundamental parts of the pre-processing module. In machine learning, the spotlight of interest is inclined towards the practical implementation, which actually is the construction of predictive models, used in predictions of everyday life, such as the task of forecasting the wind energy production from a wind farm. Inductively, a need exists to assess the performance of these models and their ability to work effectively, when applied in real life situations. For that reason, the initial dataset is to be divided into two subparts, which are called training and testing sets. The former is used for training the model, so that it learns the insightful correlations within the data. The latter, which consists only of newly seen cases, is used to evaluate the effectiveness of the previous learning process, serving as a target of prediction for the trained model. Lastly, in this study, the avenue of splitting the initial dataset into the two aforementioned subparts is automatically addressed, using algorithmic techniques.

Lastly, every part of the data preparation is going to be separately implemented onto the training and testing sets, in order to prevent the consequences of data leakage from having negative effects on the performance of the models, had the original dataset initially been used. These consequences are related to falsesly overoptimistic generalization errors, due to the models' retrospective acquaintance with information of the testing set, during the training phase, which is normally meant to be independenlty predicted in the testing part.

3.4 Evaluation Metrics

The evaluation regards the way of calculating the prediction's error. This, refers to the deviation, between the actual value and the corresponding generated prediction by a predictive model. Moreover, there are two types of errors, called "in-sample" and "out-of-sample", which are calculated, during the training and testing phases, respectively, while their comparison serves, as an indication of the presence of an overfitted or underfitted model. Technically speaking, there are many error metrics, which are being used in regression problems. However, the most common are: the root mean squared error (RMSE), mean absolute error (MAE). In this study, both these metrics are going to be used, so that there is a clearer and more robust comparison, among the constructed predictive models. Distinctively, the above two metrics, which are described in equations (a) and (b), solely measure the prediction's accuracy of the models. Below, are given their formulas:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}, \forall\, n \in Z_+^* \qquad (11) \qquad\qquad \mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}, \forall\, n \in Z_+^* \qquad (12)$$

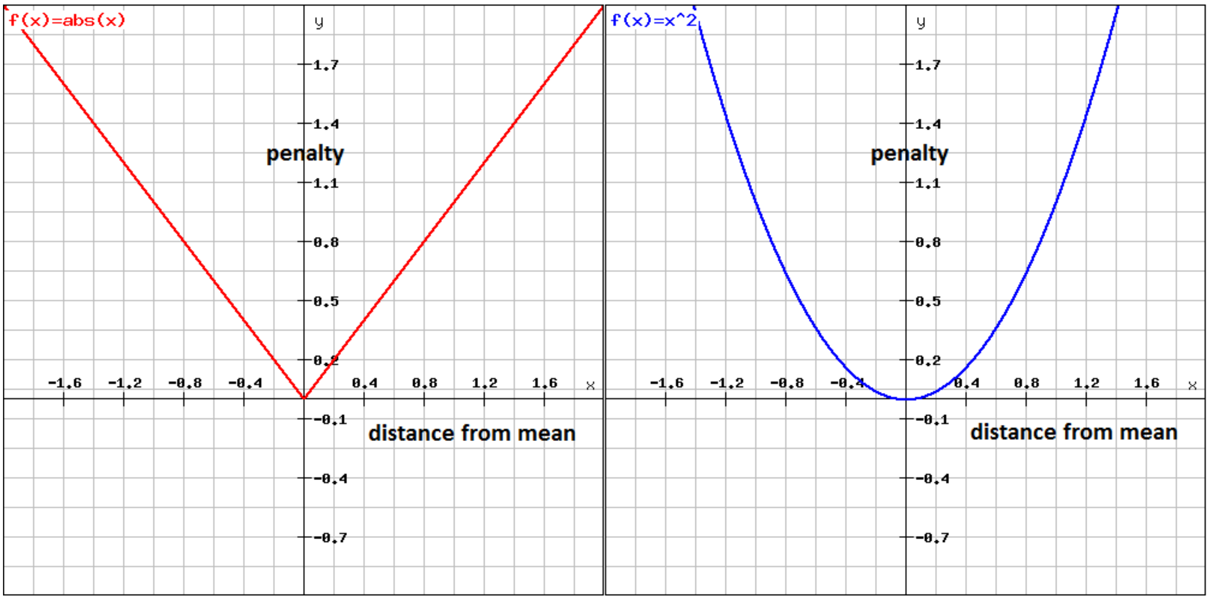


Figure 1: MAE (left) and RMSE (right) graphically illustrated, Source: https://i.stack.imgur.com/PtfUm.png

3.5 Explanatory Data Analysis (EDA)

Above, there is a graph of the target and predictor variables. The dependent variable does not appear to have a constant mean and variance, which preliminary suggests that it be a non-stationary process. Also, there are not any noticeable outliers, spikes or sudden shifts. Moreover, no strong seasonal patterns and cyclical movements are discernible through visual depiction. Furthermore, it is observed that a relationship exists between the predictors and target variables, while the most noticeable is the one between wind energy generation and wind speed, air pressure. However, although graphs' inspection helps in revealing hidden characteristics of the data, more rigorous statistical tests are to be employed, in order to technically assure their existence.

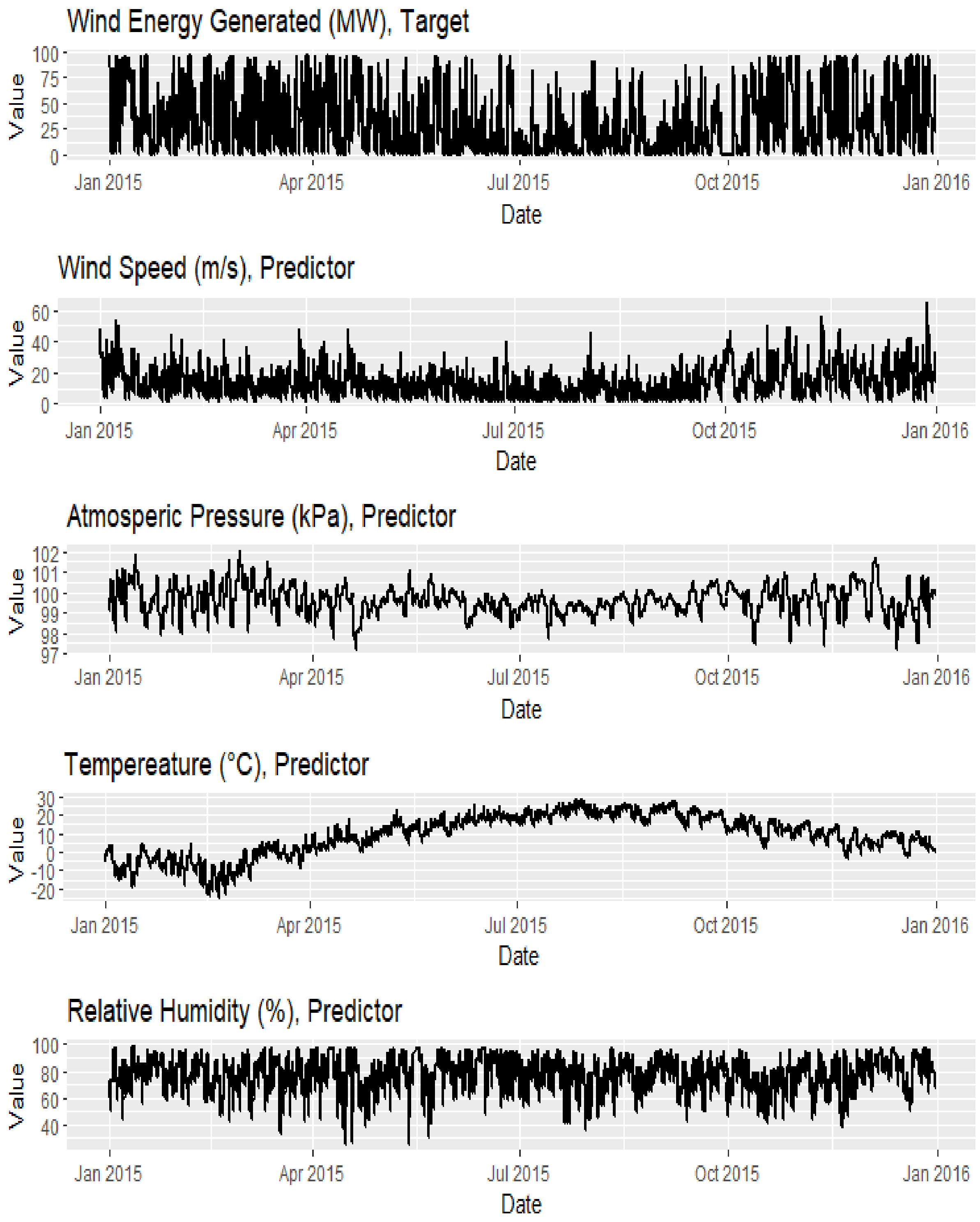
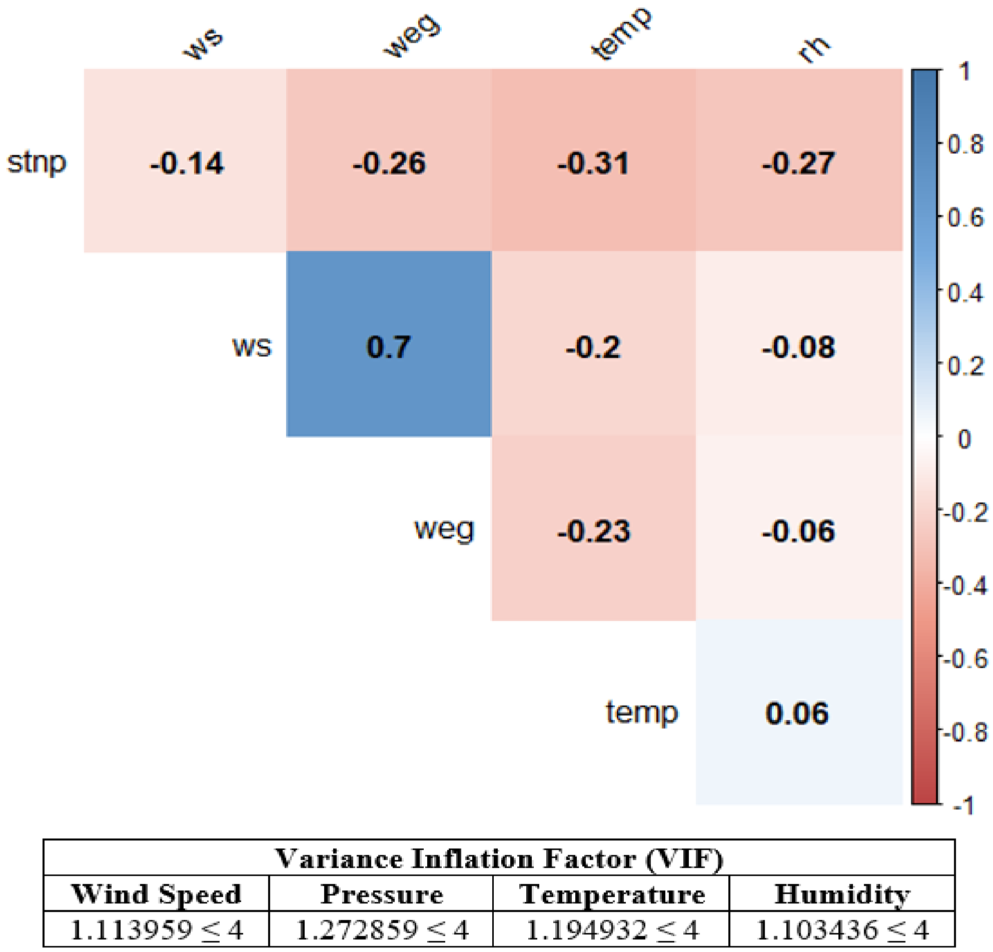


Figure 2: Target and Predictor Variables Over Time

Following, a correlation matrix (heatmap) and the variance inflation factors (VIF) of the independent variables are illustrated. According to the heatmap, there is a linear relation, associating the target with the predictor variables, while the most prominent is the one between wind speed and wind energy generation. This, is in line with the observations from the preceding graph of the different time series. Also, there is a linear relationship among the independent variables, indicating the presence of multicollinearity, which negatively effects the training process of the predictive models. However, as the VIF table designate, its current state is not severe enough, since the VIF values of every variable is below the relating threshold.



Variance Inflation Factor (VIF)			
Wind Speed	Pressure	Temperature	Humidity
$1.113959 \leq 4$	$1.272859 \leq 4$	$1.194932 \leq 4$	$1.103436 \leq 4$

Figure 3: Correlation Matrix and VIF scores

The below complex of graphs intends to shed light to the underlying and hidden characteristics of the dataset through the inspection of the residuals' behaviour. Specifically, the graph (a) describes the nature of the relationship's form, between the target and the predictor variables, namely if it is linear or non-linear. In this case, linear regression did not capture every pattern in the data, hence indicating their non-linear nature. Next, graph (b) shows how the residuals of the linear regression are distributed. The residuals' distribution seems to be resembling the normal, yet it is not identical. Continuing, graph (c), also called spread-location plot, helps in examining how the residuals are spread along the range of the predictor variables. Here, the funnel-shaped dispersion of the samples unveils the presence of heteroskedasticity. Lastly, graph (d) helps to identify any influential cases in the dataset, i.e. outliers, leverage points etc, which might influence the determination of the best regressor. It is shown that not many influential samples exist in the dataset, apart from few cases, which are denoted by their index number. The issue of influential cases will be more thoroughly assessed later. These remarks divulged some of the particular characteristics of the data, which violate several assumptions of the linear statistical models that use ordinary least squares as an optimization technique. Hence, in such situations, other predictive models should be used, so that better understanding of the nature of the data is derived.

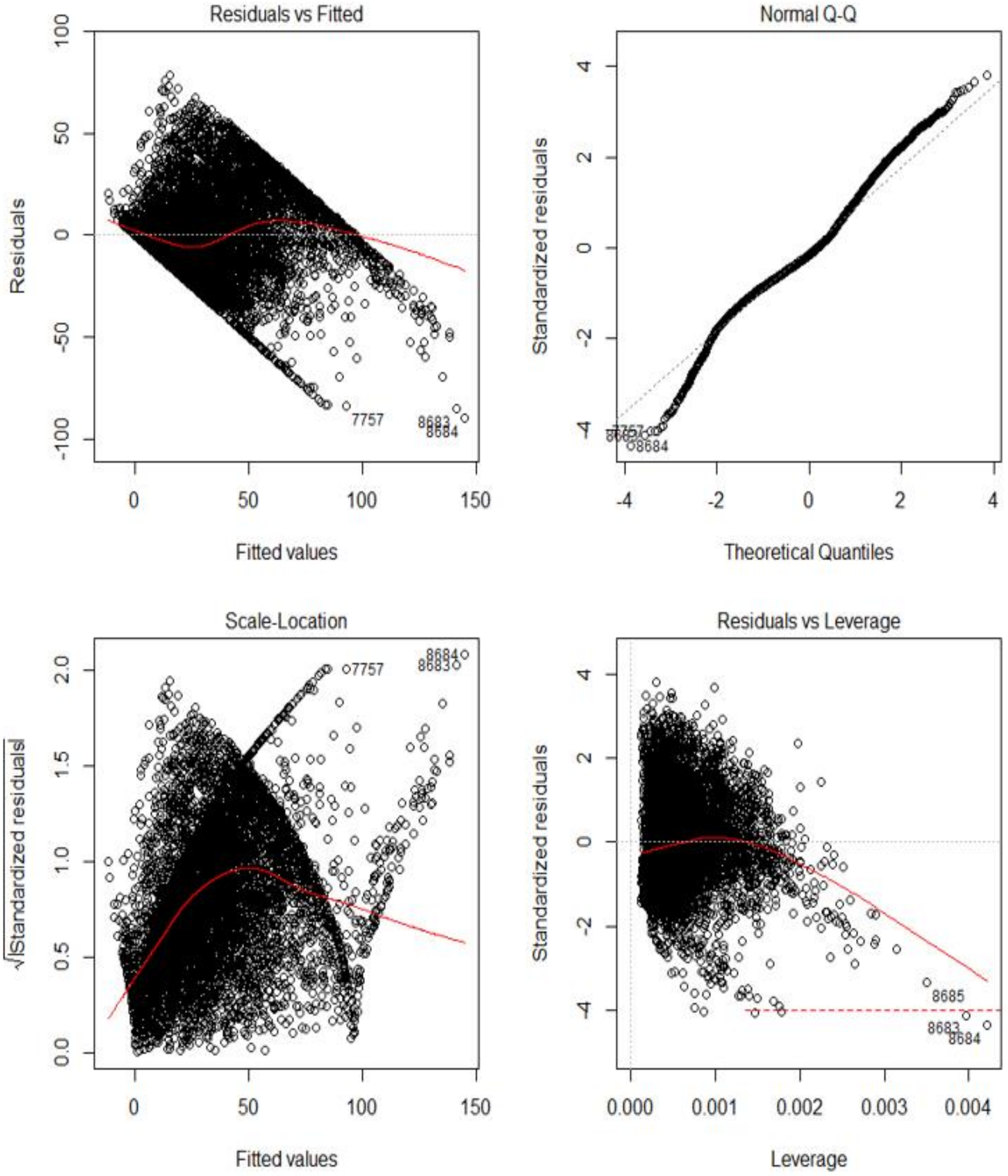


Figure 4: Diagnostic Plots of Multiple Linear Regression

Now, it would be wise to more technically investigate matters, concerning influential cases in the dataset. After calculating the Cook's Distance threshold (red line) for the regression and the same value for every sample (black bars), the comparison of them is conducted. Instantly, only a few cases exceeded the threshold, and thus, were further investigated. Specifically, two linear regressors were fitted into two different datasets, from which the one had every influential observation excluded from it. The adjusted coefficients of determination of the two regressors do not have great difference, thus no action is needed regarding the addressing of the outliers in the data.

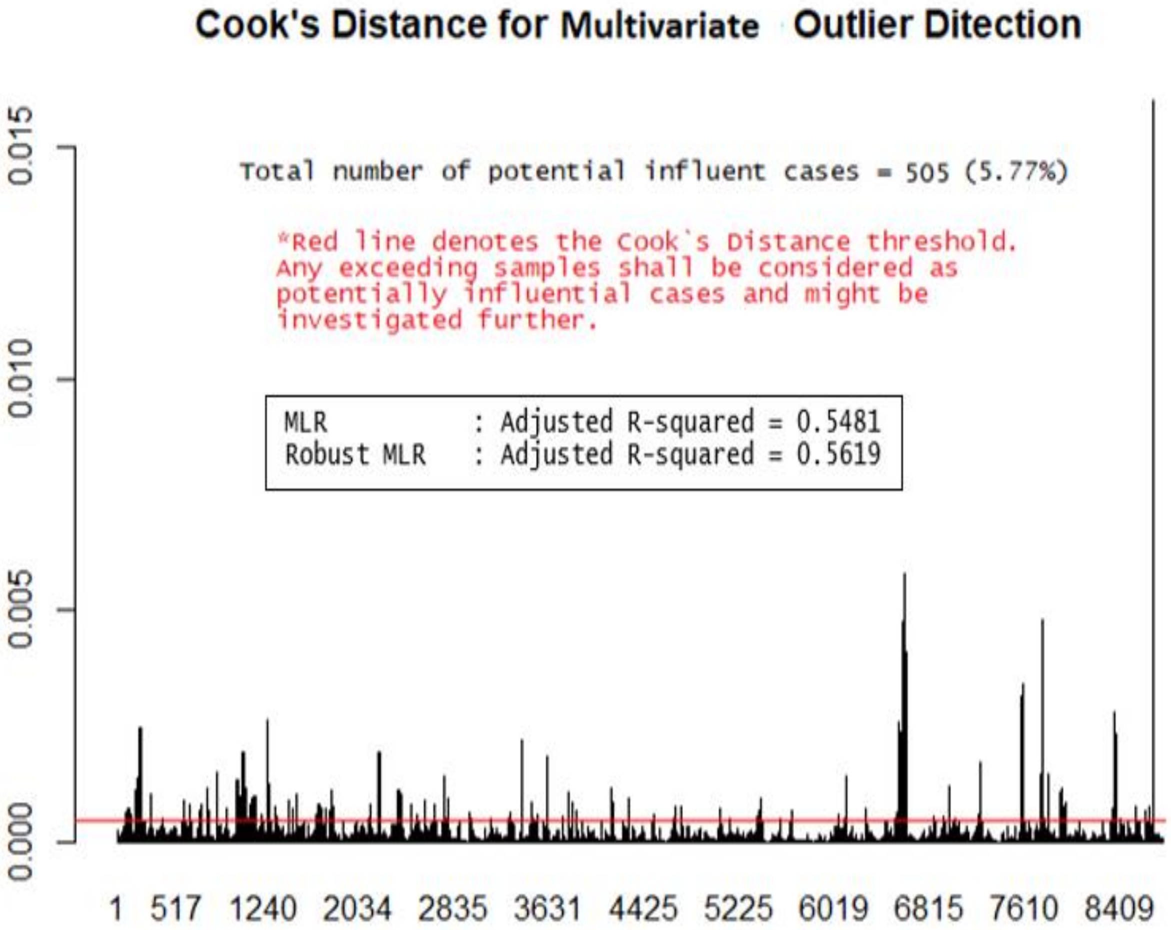
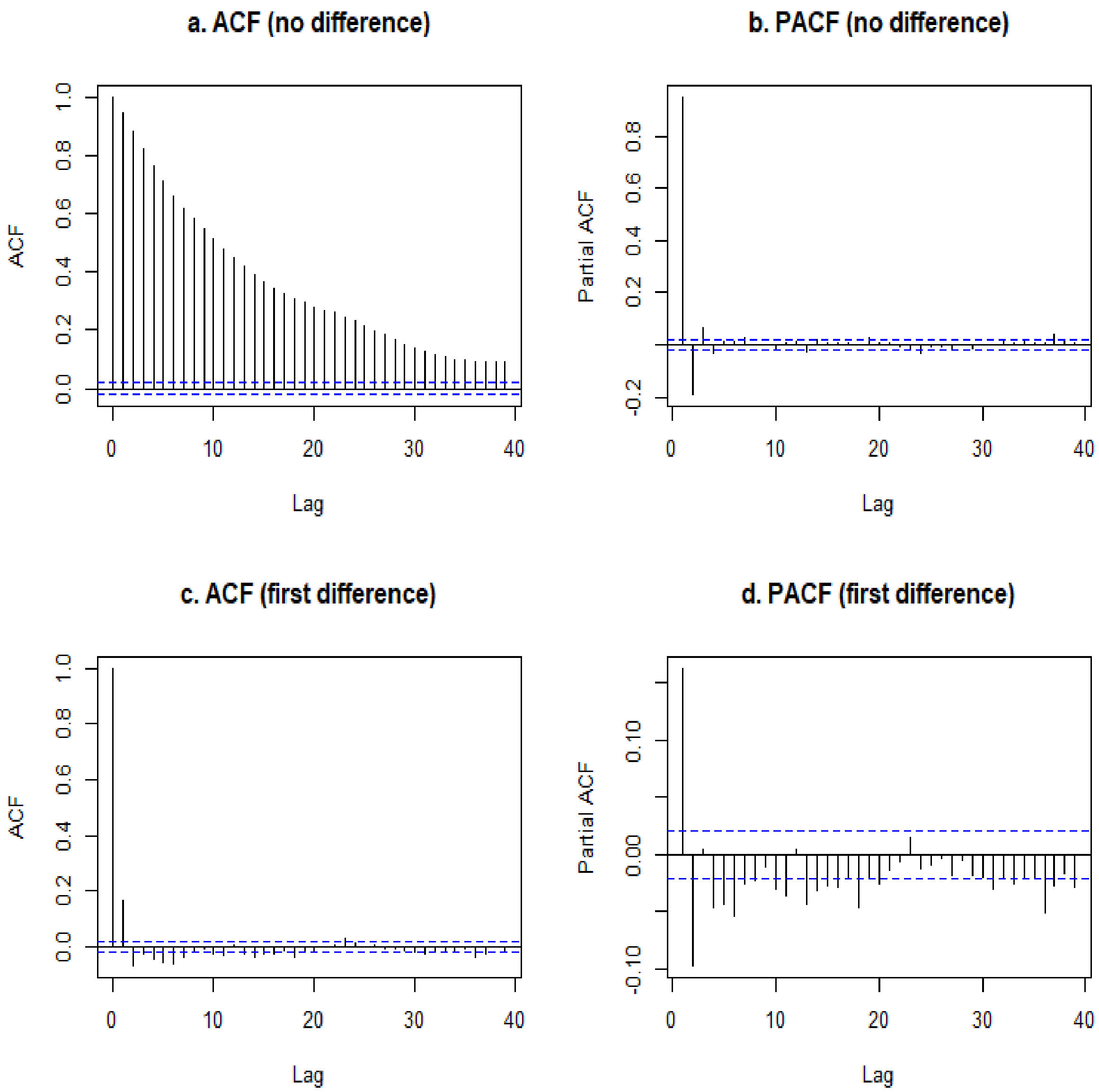


Figure 5: Cook's Distance for Multivariate Outlier Detection

Proceeding, in order to investigate the autocorrelation of the target's time series, the graph of the functions of the autocorrelation and partial autocorrelation are shown. More specifically, the ACF describes a gradually decreasing behaviour of the autocorrelation of the residuals, which is mostly be due to the propagation of the first order's autocorrelation to the next ones, as it is also ratified by the abrupt cut off of the undifferenced series' PACF (b) at the first lag. Furthermore, the persistent decaying over time of the ACF (a) indicates that non-stationarity is present. Furthermore, no signs of seasonality or cyclical movements are present, according to the graphs (a) and (b). It is evident that the derivation of the first differences of the original series has converted it into a stationary and random process, since it has almost every of its lagged autocorrelations statistically insignificant. Although, the differenced series' ACF and PACF can serve as a guide for selecting the ARIMA model's best hyperparameters, they are not indicative enough of the optimal setting that has to be used, presenting ambiguous results. Nevertheless, their concurrent sharp cut off at the first lag, reveal the process's joint AR and MA signature. Conclusively, the optimal hyperparameter setting has to be found through grid search methodology, which was described in the methodology part.



a. ACF (no difference)

b. PACF (no difference)

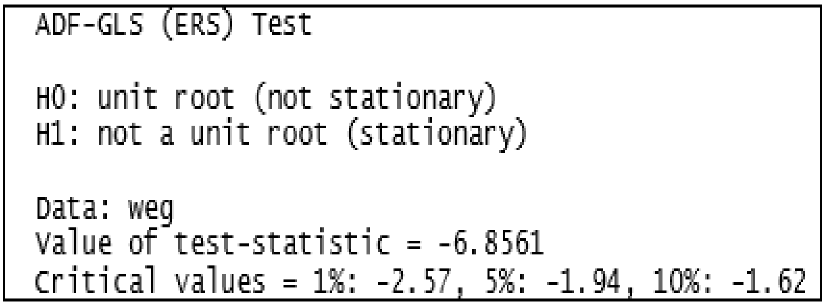
c. ACF (first difference)

d. PACF (first difference)

*blue dashed lines indicate 95% level of statistical significance

Figure 6: Autocorrelation (AC) and Partial Autocorrelation (PAC) Functions Before (a and b) and After (c and d) Differencing

Next, the validity of the stationarity's assumption, which has to hold true for timeseries data, is being formally checked. To that end, the ADF-GLS (ERS) test is employed, since, among the available test candidates, it produces the most accurate and reliable results. As it is evident, the underlying series has a unit root, namely is non-stationary, as the value of the test-statistic is lower than the minimum critical value. This, also confirms the previous visual inspections.



```
ADF-GLS (ERS) Test

H0: unit root (not stationary)
H1: not a unit root (stationary)

Data: weg
Value of test-statistic = -6.8561
Critical values = 1%: -2.57, 5%: -1.94, 10%: -1.62
```

Figure 7: Output of the ADF-GLS (ERS) Unit Root Test

Continuing, a histogram and a kernel density estimate (KDE) graph of the original series are shown. Seemingly, the series is differently distributed from the normal distribution, and resembles a positively skewed and long-tailed random variable. More specifically, most of the observations are centred around zero, while a part of them is dispersed along the right tail of the distribution.

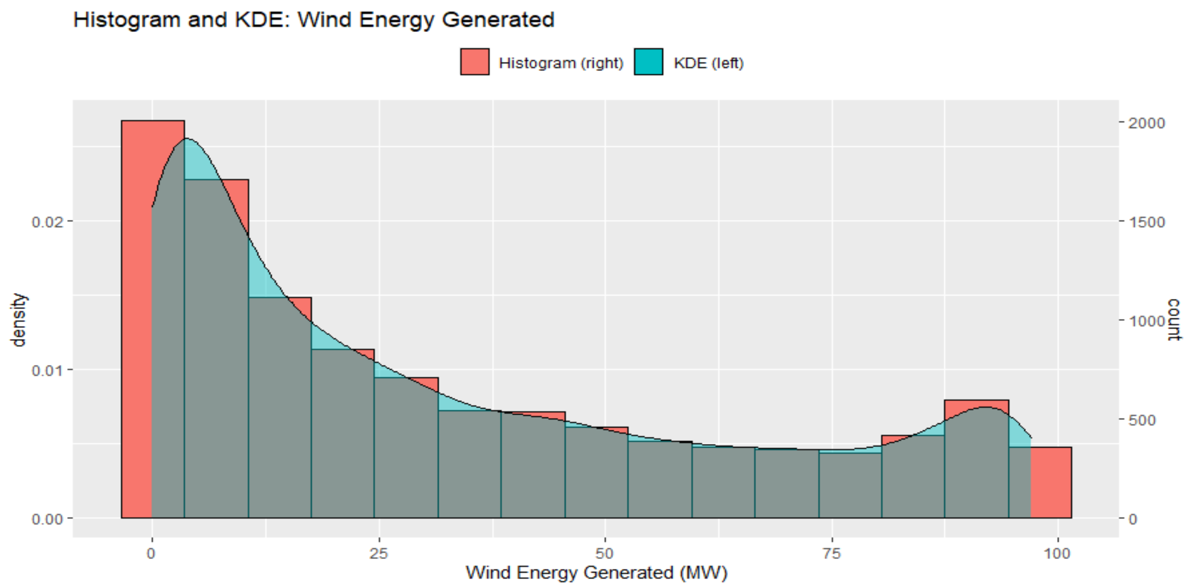
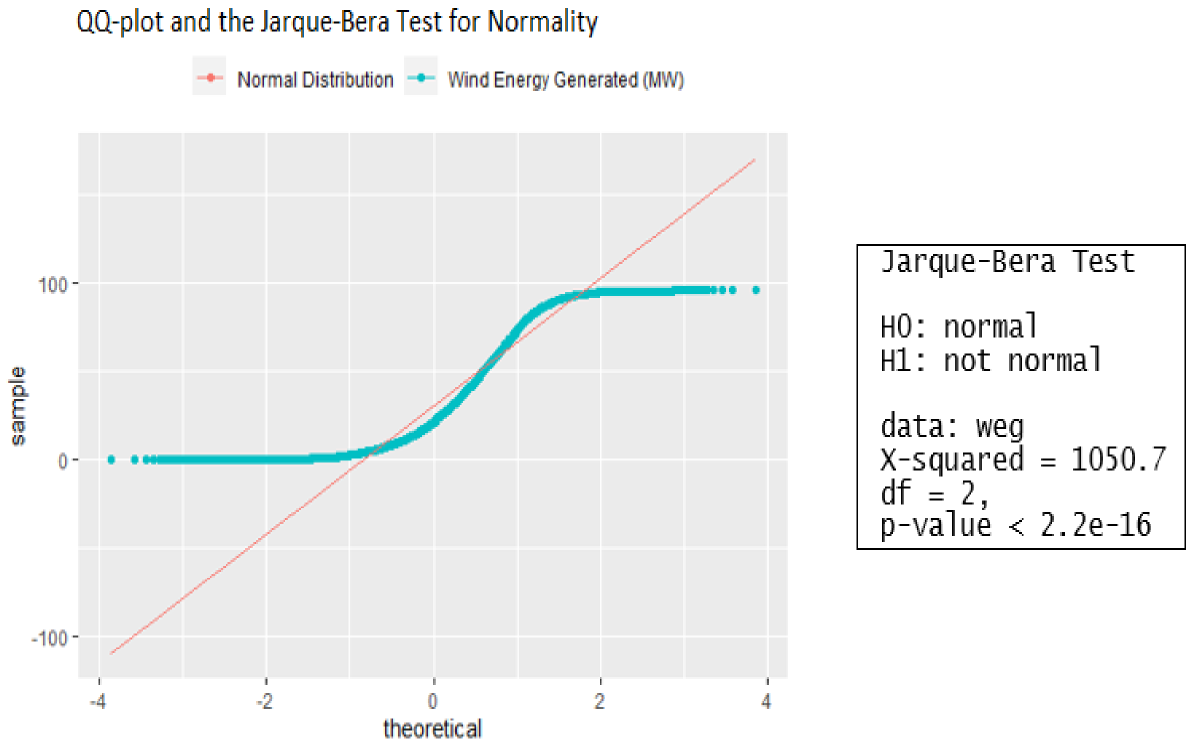


Figure 8: Histogram and KDE of Wind Energy Generation Series

Finally, the quantile-quantile plot of the time series is demonstrated, along with the Jarque-Bera test for normality. As the graph indicate, the data does not follow the normal distribution, a fact which was implied by the previous histogram, and is now strictly validated by the statistical test's results.



```
Jarque-Bera Test

H0: normal
H1: not normal

data: weg
X-squared = 1050.7
df = 2,
p-value < 2.2e-16
```

Figure 9: QQ-plot and the Jarque-Bera Test for Normality

3.6 Results

In this section, the results of the first part of this study will be discussed. Below, there is a table, in which the generalization errors (testing or out-of-bag) of the different predictive models are listed. The persistence model is shown first, since it serves as the baseline for the comparison among the rest of the models. Generally, the comparison consists of two parts. Initially, is measured the difference in the accuracy between some models, in order to establish, which one of them is the best. Then, the accuracy of the best model is juxtaposed with that of the persistence model, and it is examined whether or not it is able to outperform it. If so, it is concluded that it can yield predictions that are more valuable for the achievement of the underlying objective than those produced by the naïve option, which in this case is the persistence model. Furthermore, the constructed models were optimally calibrated by selecting the best setting of hyperparameters for the study's purpose, and so none of them did overfit or underfit the data. Hence, the training errors and the best set of hyperparameters of the models are not cited, because they add no value to the analysis of the results.

Model	**MAE (MW)**	**RMSE (MW)**
Persistence Model	8.57590	13.81597
Linear Regression	11.75929	15.94240
Elastic Net Regression	7.07855	8.48146
Polynomial Regression	10.02260	12.04853
ARIMA	10.67559	12.22158
k-Nearest Neighbours Regression	6.63282	8.55386
Random Forest Regression	6.93581	8.77609
Support Vector Regression	**6.45846**	8.88295

Table 2: Final Generalization Errors

The first main goal of this article was to establish a comparative relationship, between various traditional and machine learning methods, which will render superior only one of the two groups of models, in terms of accurately predicting wind power production. According to the table above, this relationship is clearly formed, since almost every machine learning model yields better predictions, both from every other traditional predictive method, and from the persistence model. By referring to MAE, it is safe to infer that the study's best model for predicting wind energy generation is the Support Vector Regression (SVR), followed by Random Forest Regression (RFR) and k-Nearest Neighbours Regression (k-NNR). Next, the best traditional method is Elastic Net Regression (ENR). Also, it is concerning that some of the models did not surpass the persistence model's accuracy, thus indicating that the task of predicting wind power is demanding, yet attainable, as has also been proven by the added complexity of machine learning models.

Now, given the conclusion made by the above discussion, it would be interesting to examine how recent technological advancements, such as machine learning, could benefit the environment, economy, and the society, in general. On top of that, it would be quite useful to convert these benefits into monetary terms, if possible, in order to better inspect their economic implications. To that end, a theoretical blueprint has been structured. What it does regard, is how could artificial intelligence benefit the functioning of Nord Pool's energy market. The inclusion of Nord Pool's case is capable, not only of capturing the benefits that machine learning entails for the environment, and wind energy producers, by means of greenhouse gasses emission mitigation and cost reduction, respectively, but also of converting them into monetary terms. In the next section of this study, the theoretical setup, and the results of this procedure will be reported and discussed.

4 Evaluation of Wind Energy Generation Forecasts of Machine Learning

4.1 What is Nord Pool?

Nord Pool is one of Europe's leading power markets. In effect, it operates through the day-ahead and intraday markets, which are the main ways of trading energy. Specifically, the day-ahead market creates a joint pool of selling and buying bids, which are placed by the producers and consumers, respectively. Note that the day-ahead is a spot market, namely agents retrospectively, one day (24 hours) before the actual time delivery, designate their position for the next day at a known and certain future price. Afterwards, these bids are aggregated, in order to calculate supply and demand for energy. On the other side, the intraday supplements the day-ahead market by serving as a regulatory tool for suppliers, so that balance is secured between supply and demand. Lastly, there exists the regulatory market, which clears any imbalances incurred, between supply and demand, after the closure of the day-ahead and intraday markets. For example, if demand exceeds supply, then conventional energy producers provide the needed energy to eliminate the deficit. Similarly, these urgent loads of energy are traded in the regulatory market an hour prior to the time of energy delivery.

4.2 How can Machine Learning help?

4.2.1 The benefits associated with the Wind Energy Producers

Since wind power production is by nature volatile, deviations between the placed selling bids and the actual energy delivery are subject to merits and penalizations, according to the current imbalance settlement system. There exist two general pricing settings, which are the single-price and double-price. For instance, in a double-price setting, if the market is in deficit and a wind energy producer delivers more power than the initially stated, thus helping in the upregulation of the energy deficit, then he is benefited for the excessive power produced. The same applies, if the market is in surplus and a producer delivers less power than expected, because he or she helps in the downregulation of the market's surplus. Contrastingly, however, if the market is in energy surplus/deficit and a producer delivers more/less energy than expected, then he is penalized for exacerbating the current energy imbalance in the market. In such situations, machine learning can benefit producers, since with its aid they can more accurately predict their future production, thereby minimizing their expected loss, which refers to penalties imposed, due to imbalances in the expected and actual energy delivery.

4.2.2 The benefits associated with the Environment

The provision of generated wind energy without deviations between the contracted and actual energy deliveries, not only upholds the better integration of wind power to the market, but it also mitigates the mediation of the regulatory market, because less energy imbalances occur over time in the day-ahead market. As previously stated, the regulatory energy mostly comes from conventional producers, who produce it, using coal, thus leading to higher emissions of greenhouse gasses (GHG) to the Earth's atmosphere. Consequently, should less energy-supplying interventions by the regulatory market happen, less GHG will be emitted. Hence, a positive relation exists, between the utilization of machine learning in the energy market and greenhouse effect's alleviation.

4.3 How can these benefits be quantified in monetary terms?

This analysis will be conducted, assuming that the market functions upon a double-pricing imbalance settlement system. In their paper, Mazzi and Pinson (2017), have formulated the profit's function of a wind power producer under the effect of certain assumptions, in relation to the loss of profit, applied due to discrepancies from the contracted energy delivery. In addition, they also formulate that loss's function, and in conjunction with the function of the profit, they construct the maximization problem of the wind energy producer. Mathematically:

$$\pi_{k=t-t_0} = p_k^{DA} \times Q_k - E(L_k), \forall\, k \in Z \qquad (13)$$

k period between the bidding (t_0) and delivery time (t) (equals 24 hours)
π profit of the wind energy producer
p^{DA} price of energy per unit in the day-ahead market
Q actual wind energy production
$E(L)$ expected loss of profit, due to imperfect information, i.e. penalties

$$L_k = \begin{cases} \left(p^{UR} - p_k^{DA}\right) \times \left(q_k^{DA} - Q_k\right), & \text{if } q_k^{DA} \geq Q_k \\ \left(p^{DR} - p_k^{DA}\right) \times \left(q_k^{DA} - Q_k\right), & \text{if } q_k^{DA} < Q_k \end{cases}, \forall\, n \in k \qquad (14)$$

p^{UR} price of energy per unit, when upregulation is required (= p^R) or not (= p^{DA})
p^{DR} price of energy per unit, when downregulation is required (= p^R) or not (= p^{DA})
p^R price of energy per unit in the regulatory market
q^{DA} contracted amount of wind energy for delivery

In equation (1), the expected loss: $E(L_k)$, is used, because it is not priorly known what amount of energy the producer is going to offer as his/her bid in the day-ahead market. That is, the expectation of the loss is calculated with regard to the forecasted amount of energy that will have produced over the fixed interval of $k = t - t_0$ hours, $E(Q_k)$. Thus, the final maximization problem is the following:

$$\max_{q_k^{DA}} \ \pi_{k=t-t_0} = p_k^{DA} \times Q_k - E\left[\, L_k \mid q_k^{DA} = E(Q_k) \,\right], \forall\, n \in k \qquad (15)$$

In the above maximization problem, the $E(Q_k)$ is the prediction of future wind energy production created by the machine learning models that had been previously described in this article. These predictions serve as a bidding guide for the producers, as they will now use them to pinpoint the optimal amount of energy that has to be contracted in the day-ahead market, which will minimize their expected loss. The monetarized benefits for the producers can be conceived as the difference between machine learning forecasts and naïve forecasts, expressed in deviancy reduction terms, between the contracted and actual energy output. These naïve forecasts are perceived as the worst-case scenario, when trying to minimize the losses in profit, associated with enforced penalties for unbalanced energy deliveries. In other words, it is the difference in the expected loss, which is calculated, using the aforementioned methods:

$$BP_k^{ML} = \sum_{k=1}^{T} \left\{ E\left[\, L_k \mid q_k^{DA} = E\left(Q_k^{ML}\right) \right] - E\left[\, L_k \mid q_k^{DA} = E\left(Q_k^{naive}\right) \right] \right\}, \forall\, n \in k \qquad (16)$$

On the other side, the benefits for the environment arise, when the energy market is in energy deficit, and the producer also delivers less energy than the expected, hence extending the already formed market deficit further. In order to effectively record the benefit that machine learning has to offer in such cases, the same naïve forecasts are again assumed the worst-case scenario. The maximum number of megawatts of wind energy, up to which machine learning methods reduce the declination that occurs by these naïve options, given that the imbalances of the market and the producer are both of negative direction, consists the benefit associated with the environment. Particularly, when these deviations, which create the market's energy deficit, are curtailed, less supplements of conventional energy have to be provided by the regulatory market into the energy grid. Hence, the less these supplements are, the lower are also the emissions of GHG into the atmosphere. Concretely:

$$\mathrm{BE}_{\mathrm{k}}^{\mathrm{ML}}\Big|_{\mathrm{q}^{\mathrm{DA}}>\mathrm{Q}}^{\mathrm{AD}>\mathrm{AS}} = \left|\sum_{\mathrm{k}=1}^{\mathrm{T}} \left(\mathrm{q}_{\mathrm{k,ML}}^{\mathrm{DA}} - \mathrm{q}_{\mathrm{k,naive}}^{\mathrm{DA}} \right)\right| \times \mathrm{p}_{\mathrm{CO}_2}, \forall\, \mathrm{n} \in \mathrm{k} \qquad (17)$$

In equation (5), the $\mathrm{p}_{\mathrm{CO}_2}$ is the marginal cost saved from the avoidance of the total amount of CO_2 that is emitted per MWh of conventional energy produced. Specifically, USEIA (US Energy Information Administration) has calculated the borderline amount of CO_2 emitted into the atmosphere per kWh of conventional energy generated, using coal, natural gas, or petroleum, for the year of 2018 in the US. That number is approximately 2.21 pounds of CO_2 per kWh. Since the volume of GHG emissions from the combustion of fossil fuels is not to be altered, if realized in a different place on the Earth, it is also deemed to be promisingly resembling that of the EU. Additionally, Richard S.J. Tol, in his recently conducted study (2019), has measured that the EU's social cost of carbon (SCC) is $0.33 per ton of carbon emitted. Hence, by algebraically manipulating the above pieces of information, it is possible to be shown that the SCC (pCO_2) per MWh of conventional energy generated is $0.3315/MWh. Thus, $\mathrm{p}_{\mathrm{CO}_2}$ equals $0.3315/MWh. Lastly, any required data for the calculation of the benefits of ML for the wind energy producers and the environment, had been collected from the official website of Nord Pool.

4.3 Methodology: The Technical Procedure

First, a new dataset was created, using all the available data, which had been collected from Nord Pool's official website. This new set of data was created, as if wind energy producers had been bidding every hour in the day-ahead market. After their bids are placed, the market stops receiving offers (usually at 12 a.m.) and the market operator clears any imbalances occurred, between supply and demand. Therefore, for every bidding hour, the dataset includes the corresponding information that is needed to calculate the deviation of the actual wind energy output from the contracted one, the net difference among market's aggregate supply and demand (deficit or surplus), the relating losses for the discrepancies from the energy delivery target, and lastly, the total amount of conventional energy avoided, because of machine learning forecasts' superiority against the ones of the naïve methods.

Continuing, the benefits for the producers and the environment, are deduced against three baseline forecasts, which are based on: the persistence model, the median, and the average of the wind power production time series. Technically, these forecasts produce predictions, based on the available past information. For instance, the persistent forecasts assume that the produced wind energy of every next hour up until the delivery time, will be the one that was recorded at the bidding time. Contrastingly, the forecasts, which are based on the average value of the production, conceive the production of every next hour to be the average production of the previous hours. Lastly, the forecasts that are generated, using the median value of the production, suppose the production of every next hour to be the median value of the production. The baseline forecasts are graphically illustrated in the plot below:

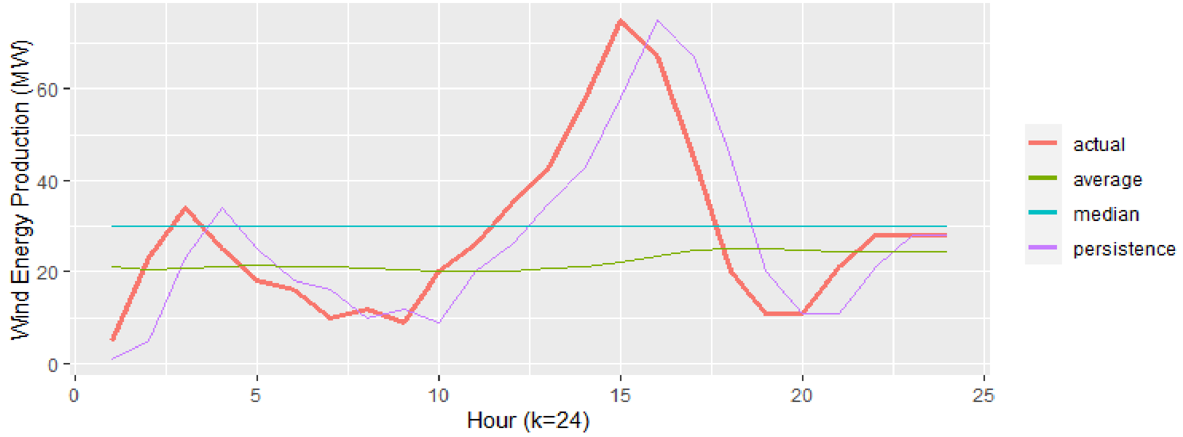


Figure 10: Hourly Forecasts of the Baseline Models

Once every baseline forecast has been constructed, their final proposals, upon which the wind energy producers will base their offers in the day-ahead market, are measured. In effect, the final proposal of a baseline model equals the summation of every hour's forecast up to and the hour of the energy delivery, which is arranged for exactly 24 hours after the bid has been placed. Thus, the proposals are as follows:

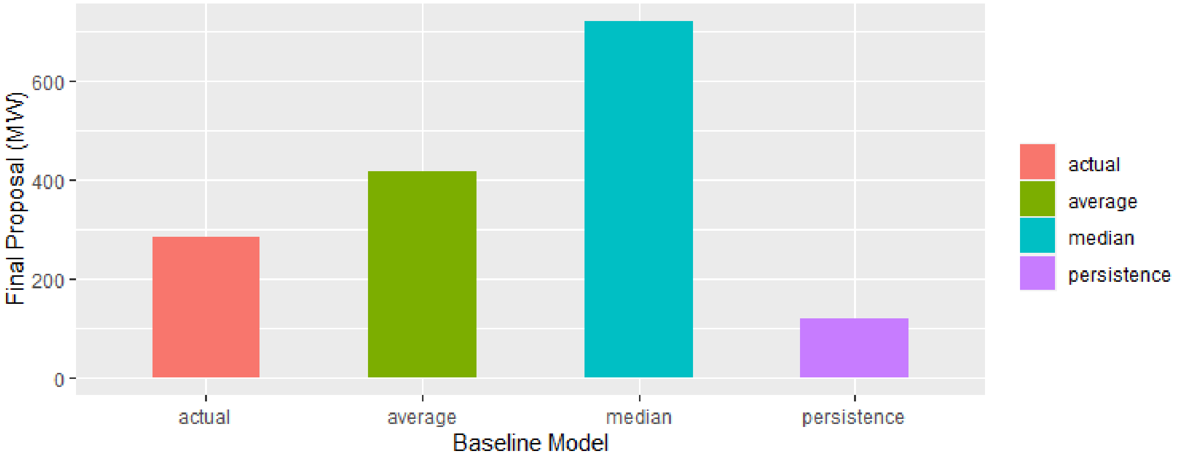


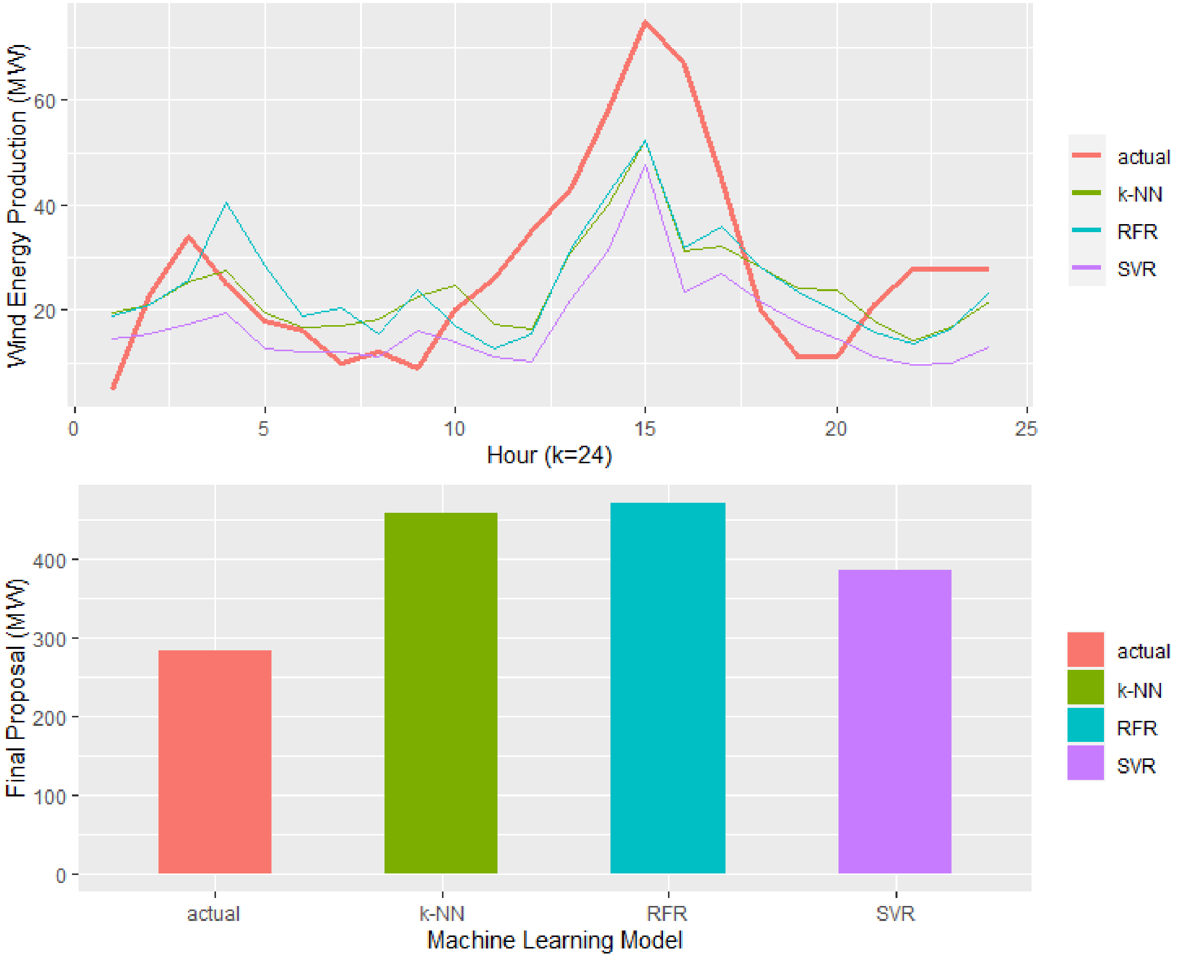
Figure 11: Final Proposals of the Baseline Models

Note that the above graph is demonstrating the final proposal of the baseline models, only for a random auction, during not a specific day of the year, and it just grants a visual representation of their function. Thus, it is not to be considered as a general rule of the effectiveness of the selected naïve models. It is evident that the selected baseline forecasts, which cover a wide variety of the producers' basic strategies for configuring the total amount of wind energy to be offered at every available auction of the day, can serve well enough their purpose, deeming from the above instance. Specifically, according to the above graph, the average of the production has offered an effective proposal, while the rest of the strategies did the same, but with some moderate amount of deviancy, which will, in conjunction with the market's clearing state, also be indicative of the imposed penalties to the wind energy producer.

Contrary to the aforementioned baseline models, another useful option that can generate bidding proposals for the producers, relies on "point-to-point" forecasts. That is, they utilize past observations to predict wind energy production, and not just guess upon its future nature. Such an option are also the forecasts, which are created, using machine learning methods. According to the analysis of the first part of this exercise, the best data-driven models for predicting wind energy generation, come from the machine learning family. More specifically, the SVR, RFR, and k-NN models have produced the best generalization errors. Therefore, these three models, along with their optimal configuration settings, will be used to quantify the benefits that are created for the environment and wind energy producers.

The training/testing procedure of the models is now different than the one followed during the first part. Here, a single split of the data is performed, in order to create a training and a testing set, where 70% of the samples are attributed to the training set and the rest 30% to the testing set. The split's weights were set in such a manner, in order for a complete number of days to be part, both of the training set, and the testing set. More specifically, a complete number of 255 and 110 days, were added to the training set and the testing's set, respectively. In other words, 255 × 24 = 6120 and 110 × 24 = 2640 auctions have been included to the two datasets, correspondingly. As it is already mentioned, the initial dataset has been calibrated in such a way, where all the needed information is granted for every auction. Next, after every necessary information is gathered, using both groups of models, BP and BE for every auction are calculated, and the mean of all of the auctions is reported back as the final result.

Next, the forecasts of the three selected machine learning models are visually illustrated. Furthermore, there is a bar chart, in which their final bidding proposals are recorded and shown. More elaboratively,



Figures 12 and 13: Hourly Forecasts and Final Proposals of the selected Machine Learning Models

in the case of the machine learning models, it is quite obvious that the regarding predictions and final proposals of theirs, converge much more to the actual wind energy generated in the course of the day, than those generated by the baseline models. A clear, consistent pattern is kept between the actual wind energy produced and the forecasts, implying the effectiveness of the models to even capture long-term relationships in the data. Although every model overestimates the actual energy production, the SVR's forecasts seem to be associated the most with the lowest incurred error among the three candidates. The above sets of charts represent only one out of the total number of available auctions for evaluation in the dataset, and so, an overall measurement of all the instances has to be derived. Lastly, the final monetarized benefits for the producers and the environment are measured in euros saved per day (or auction), due to the integration of machine learning technologies into the energy market.

4.4 Results

Indeed, the results of the second part of this study, similarly to these of the first part's, designate an apparent dominance of the machine learning over benchmark forecasts, employed by the wind energy producers, to successfully predict future wind power production. For instance, in the case of the final monetarized benefit for the producers (FBP_{ML}), the capitalization of advanced technologies can yield on average a 52.4% decrease in the daily (per auction) expected losses of the average wind energy producer. In this exercise, a wind farm of total capacity of 100 MW have been utilized, whose summary statistics are shown below, in order for a comparative nature of the results to be revealed:

Wind Farm (MW)					
Min	1Q	Median	Mean	3Q	Max
0.00	6.00	22.00	32.51	55.00	97.00
Expected Loss Using Baseline Forecasts (€/auction)					
Min	1Q	Median	Mean	3Q	Max
0.00	0.00	23875	**34616**	52115	385899
Expected Loss Using Machine Learning Forecasts (€/auction)					
Min	1Q	Median	Mean	3Q	Max
0.00	0.00	12722	**18138**	28350	238374

Table 3: Summary Statistics of the Final Benefit for the Wind Energy Producers

$$\mathrm{FBP_{ML}} = |\mathrm{Loss_{ML}} - \mathrm{Loss_{naive}}| = |18{,}138 - 34{,}616| = \mathbf{16,478\,€/auction\ (day)/100MW}$$

More simply, a producer, who occupies a wind farm of total capacity of 100 MW, can on average save 16,478 euros per submitted offer (auction) in the day-ahead market, should he or she utilize machine learning forecasts, instead of those of the standard options. It would be possible to extrapolate this expectation of the final benefit also for wind farms of different energy capacities, by analogically manipulating the above summary statistics. However, these analogies will be a part of the realm of speculation and should be taken with a grain of salt.

Moving onto the benefits, linked to the environment, one should take a more cautious glance to equation (5). Specifically, it is comprised of two terms, of which the second one is the marginal cost, $\mathrm{p_{CO_2}}$, of a pound of CO_2 emitted per MW of conventional energy produced. Now, let's rule that term out, and only focus on the first one. The first term of the equation, neglecting for now its absolute value, is responsible, for tracking down of how much better do machine learning options can predict the optimal final offers to be submitted in the day-ahead market. Algorithmically, the construction of this metric was done by using only the instances, where: $\mathrm{q^{DA}_{k,ML}} < \mathrm{q^{DA}_{k,naive}}$, in order to reveal the worth that machine learning adds to the whole procedure. Now, moving further, the difference: $\mathrm{q^{DA}_{k,ML}} - \mathrm{q^{DA}_{k,naive}}$, shows the portion of the baseline models' error that was curtailed, because of the deployment of the new forecasts. Evidently, this difference is always negative, thereby demanding the calculation of its absolute value.

The results have shown that smart technologies, such as machine learning, can indeed mitigate the extensive emission of GHG, when applied to the energy market, and overall, help in alleviating many of climate change's devastating consequences, due to high amounts of conventional energy produced. Practically, it is proven that a 160% reduction in GHG emissions can take place, if artificial intelligence be used for building a smart energy grid, which is not entirely relying on basic strategies for predicting future wind energy production, like the naïve ones. By using a part of eq. 5, it is relatively easy to demonstrate that a total number of 389,547 MWh of conventional energy can be avoided, during a time span of 4 months. The same number, if one use only the baseline forecasts, is 149,826 MWh. According to the previously mentioned information (see pg. 25), the total benefit, in terms of avoided GHG emissions, is translated into a volume of 2,210 × 389,547 = 861 × 10^6 pounds of CO_2 that is emitted into the Earth's atmosphere, in case, where the machine learning option is neglected. This, in monetary terms, writes:

$$\mathrm{FBE_{ML}} = 389{,}547 \times 0.3315 = \mathbf{129,135\,€/4\ mos./100MW}$$

Please, note that the final benefits for the producers and the environment are measured per a capacity of 100 MW of wind energy. That is, since the current study has been conducted, using a wind farm of a total capacity of 100 MW, the amount of the conclusive benefits is assuredly associated only with that certain capacity. Therefore, for different energy capabilities, e.g. the wind energy potential of a whole country, continent etc, separate studies have to be taken into consideration before these results are ensured to be able to be generalized, for instance, for any other energy capacity.

4 Conclusion and Discussion

Summarizing, the two main questions under consideration of this study were, on the one hand, if machine learning can be used in the energy sector for predicting future hourly wind energy production by wind farms, and on the other hand, if that is the case, what would the associated monetarized benefits be for the environment and the economy. The results of the first part's answers (see paragraph 3.6) indicate that machine learning is eligible for the accomplishment of such a purpose, as it can yield better results than the naïve forecasts. More specifically, it is found that machine learning algorithms, such as the Support Vector Regression, k-Nearest Neighbours, and Random Forest Regression, when their hyperparameters are optimally tuned, can reduce the persistence model's Mean Absolute Predictive Error (MAE) up to 24.7%.Similarly, the results of the second part (see paragraph 4.5), given the stipulated benefits for the wind energy producers and the environment, described in paragraphs 4.2.1 and 4.2.2, respectively, support that machine learning can produce tangible capitalized added value for both agents. For instance, wind energy producers, who utilize a wind farm of total capacity of 100MW, can, on average, reduce their expected losses, due to deviances between contracted offer and actual energy delivery, up to 16,478€ per submitted offer in the day-ahead energy market of Nord Pool. On the other side, the environment can get rid of itself approximately 861 million pCO_2, in a total time span of 4 months per 100MW of wind energy introduced into the energy grid. That is, expressed in monetary terms, 129,135€ saved per 4 months per 100MW of wind energy.

In line with the initial hypothesis that machine learning can be successfully employed for forecasting wind power generation (Chang, 2014), are also the above-mentioned findings. Moreover, as Treiber et al. (2016) have shown, the tuned SVR algorithm can outperform the persistence model, in terms of Mean Absolute Error (MAE), by 24%, while only using wind speed, as a predictor, and wind energy generated, as a target variable. Contrastingly, the current study has jointly selected wind speed, air pressure, temperature, and relative humidity, as the features, comprising its set of target variables. This, could indeed mean that wind speed is the most prominent predictor, when forecasting of future wind power generation is the underlying task, and that any other of the aforementioned variables are of minor significance. Also, in contrast with Chaudhary et al. (2020), who showed that Decision Trees (a Random Forest with one tree) can greatly outperform Support Vector Regression, this study found that both algorithms have almost the same predictive accuracy, but with the former to be generating slightly better results than the latter. Overall, however, as the rest of the studies have found (see, for example, Zendehboudi et al., 2018), the machine learning algorithms, which were deployed in this exercise, are yielding much better results than the traditional models, e.g. Linear Regression, Polynomial Regression, and ARIMA. Lastly, it is not easy enough to contextualize this paper, with respect to its second part, as no studies exist that intend to examine the same, or similar, field.

The conclusions from this article, should be taken into account, when considering to implement smart technologies, related to artificial intelligence, into a structure that produces clean energy, through an energy generator, which utilizes wind resources to do so, i.e. mostly wind farms. The results, not only provide insights for the expected economic feasibility of this venture for the wind energy producers, but also for the environment, which equally renders them suitable to be taken advantage of by policy makers, so that they speculate over the results of introducing machine learning, as a broader regime, in the energy sector, in order to further precipitate the current trajectory of world decarbonization.

5 Bibliography

Alencar, David & Affonso, Carolina & Oliveira, Roberto & Moya Rodríguez, Jorge & Leite, Jandecy & Reston Filho, José Carlos. (2017). Different Models for Forecasting Wind Power Generation: Case Study. Energies. 10. 1976. 10.3390/en10121976.

Barbounis, Thanasis & Theocharis, John & Alexiadis, Minas & Dokopoulos, Petros. (2006). Long-Term Wind Speed and Power Forecasting Using Local Recurrent Neural Network Models. Energy Conversion, IEEE Transactions on. 21. 273 - 284. 10.1109/TEC.2005.847954.

Bitar, Eilyan & Rajagopal, Ram & Khargonekar, P. & Poolla, Kameshwar & Varaiya, Pravin. (2012). Bringing Wind Energy to Market. Power Systems, IEEE Transactions on. 27. 10.1109/TPWRS.2012.2183395.

Catalão, João & Pousinho, H.M.I. & Mendes, V.M.F.. (2009). An Artificial Neural Network Approach for Short-Term Wind Power Forecasting in Portugal. Engineering Intelligent Systems. 17. 1 - 5. 10.1109/ISAP.2009.5352853.

Cadenas, Erasmo & Rivera, Wilfrido. (2010). Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA-ANN model. Renewable Energy. 35. 2732-2738. 10.1016/j.renene.2010.04.022.

Chaâbane, Najeh. (2014). A hybrid ARFIMA and neural network model for electricity price prediction. International Journal of Electrical Power & Energy Systems. 55. 187–194. 10.1016/j.ijepes.2013.09.004.

Chang, Wen-Yeau. (2014). A Literature Review of Wind Forecasting Methods. Journal of Power and Energy Engineering. 02. 161-168. 10.4236/jpee.2014.24023.

Colak, Ilhami & Sagiroglu, Seref & Yesilbudak, Mehmet. (2012). Data mining and wind power prediction: A literature review. Renewable Energy. 46. 241–247. 10.1016/j.renene.2012.02.015.

Kristiansen, Tarjei. (2014). A time series spot price forecast model for the Nord Pool market. International Journal of Electrical Power & Energy Systems. 61. 20–26. 10.1016/j.ijepes.2014.03.007.

Mazzi, Nicoló & Pinson, Pierre. (2017). Wind power in electricity markets and the value of forecasting. 10.1016/B978-0-08-100504-0.00010-X.

Tol, Richard. (2019). A social cost of carbon for (almost) every country. Energy Economics. 83. 10.1016/j.eneco.2019.07.006.

Treiber, Nils & Heinermann, Justin & Kramer, Oliver. (2015). Wind Power Prediction with Machine Learning (chapter). 10.1007/978-3-319-31858-5_2.

Zendehboudi, Alireza & Abdul Baseer, Mohammed & Saidur, R. (2018). Application of support vector machine models for forecasting solar and wind energy resources: A review. Journal of Cleaner Production. 199. 272-285. 10.1016/j.jclepro.2018.07.164.

Government of Canada (Meteorological Resources):
https://climate.weather.gc.ca/historical_data/search_historic_data_e.html

Ontario’s Independent Electricity System Operator (IESO) (Wind Energy Generation Resources):
http://www.ieso.ca/power-data/data-directory