International Journal of
Intelligent Systems and Applications in Engineering

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

An Investigation of the Effect of Meteorological Parameters on Wind Speed Estimation Using Machine Learning Algorithms

Cem Emeksiz*[1], Gulden Demir[2]

Accepted : 16/12/2018 **Published: 31/12/2018**

Abstract: Wind speed is the most important parameter of the wind energy conversion system. Therefore temperature, humiditiy and pressure data, which has significant effect on the wind speed, have become extremely important. In the literature, various models have been used to realize the wind speed estimation. In this study; Six different data mining algorithms were used to determine the effect of meteorological parameters on wind speed estimation. The data were collected from the measurement station established on the campus of Gaziosmanpaşa University. We focused on the bagging algorithm to determine the appropriate combination of wind speed estimates. The bagging algorithm was used for the first time in estimation of wind speed by taking into account meteorological parameters. To find the most efficiency method on such problem 10-fold cross validation technique was used for comparision. From results, It is concluded that bagging algorithm and temperature-humiditiy-pressure combination showed the best performance. Additionaly, temperature and pressure data are more effective in the wind speed estimation.

Keywords: *Bagging algorithm, Data mining, Renewable energy, Wind speed estimation.*

1. Introduction

With industrialization, basic energy resources emerge as an indispensable phenomenon to maintain the daily life of human beings. In industrial societies, fossil fuels are used as the main energy source. Nowadays, while a large part of energy needs are met by fossil fuels, it is a necessity to turn to alternative sources when the negative impact of these resources on the environment is taken into consideration. At the same time, global warming, seasonal changes and the political approaches of the countries have accelerated the search for alternative energy sources in the future where fossil fuel reserves can not meet their needs. This is especially the case for wind energy and solar energy, which are among the renewable energy sources.

Renewable energy sources can be renewed owing to the variety of energy that does not emit to the environment, which is thought to take its power from the sun and never to be consumed. When we look at the energy policies of developed and developing countries, we see that they are turning to renewable energy sources and accelerating their efforts to develop these resources. Solar energy, wind energy, hydrolic energy, biomass energy, hydrogen energy, geothermal energy, wave energy, tidal energy are the categories of renewable energy resources.

According to the 2017 Renewable Energy Global Situation Report, in 2016, a total of 161 GW renewable energy is installed in the power system. The total global capacity has increased by about 9% compared to 2015 and is almost 2.017 GW at the end of the year. In 2016, about 47% of the newly established renewable energy capacity was solar energy. This ratio was followed by wind energy with 34% and hydroelectric energy with 15.5% [1].

Wind energy is at the forefront of the renewable energy sources that are the most used owing to cost efficiency in the world. In the past, wind energy was used to move sails, run ships, windmills and in irrigation. Recently, the use of wind turbines operating in the world by wind energy in the production of electricity has become increasingly widespread. According to the 2016 Global Wind Statistics report on the situation of the global wind industry of Global Wind Energy Council (GWEC), countries with the highest installed power are shown in Table 1 [2].

Table1 The top 10 cumulative capacity DEC 2016

Country	MW	%Share
PR China	168.690	34.7
USA	82.184	16.9
Germany	50.018	10.3
India	28.700	5.9
Spain	23.074	4.7
United K.	14.543	3.0
France	12.066	2.5
Canada	11.900	2.4
Brazil	10.740	2.2
Italy	9.257	1.9
Rest of theworld	75.577	15.5
Total TOP 10	411.172	84
World Total	486.749	100

To benefit technologically from wind energy; it is very important to know the possibilities of utilization, to determine the zones with high wind energy potential, to be able to predict the wind characteristics and wind speeds. In particular, estimating wind speeds is necessary for estimating the energy expected to be generated from wind turbines in short, medium and long period. According to these estimation values, the profitability of the power generation plants can be calculated and it is determined whether or not the investment of wind energy in a region will be profitable. In this way, the operating and production costs can be calculated more accurately [3].

Wind energy estimations are generally produced by hybrid models

[1] *Electric-Electronic Eng. Tokat Gaziosmanpasa University, Tokat, Turkey*
[2] *Electric-Electronic Eng. Tokat Gaziosmanpasa University, Tokat, Turkey*
* *Corresponding Author: Email: cem.emeksiz@gop.edu.tr*

that are used physically, statistically, or both. The physical method is generally used for long-term estimates, while statistical methods are used for short-term estimates. In the physical approach, maps of the plant area to be used in the models, topographic maps and smoothness maps are prepared using Geographic Information System (GIS) software. The plant settlement is done on maps, and the flow of wind is modeled by computational fluid dynamics software. In the statistical estimation models, the relationship between the local winds and the numerical weather forecast results is established by using the meteorological data in the past. The statistical approach uses models such as Neural networks (NN), Support Vector Machine (SVM), Fuzzy Logic and Regression Trees.

When studies on wind speed estimations are examined; Monthly mean temperature measured from Arak weather station from 1960 to 2005, dew point in sunny hours, relative humidity, average wind speed, saturated vapor pressure data and monthly potential evaporation temperature estimation were studied. Bagging model was found to be functionally more successful as a result of Root Mean Square Error (RMSE), Mean Squared Error (MSE) and Correlation Coefficient (CC) values [4]. Cadenas et al. used an univariate Autoregressive Integrated Moving Average (ARIMA) model and a multivariate Nonlinear Autoregressive (NARX) model to estimate the wind speed. The data set consists of the parameters taken from the stations in two different regions. Mean Absolute Error (MAE) and Mean Square Error (MSE) values are used to analyze the results. Using the NARX model has proved to be more accurate. It has also been proposed to incorporate additional meteorological parameters in wind speed prediction models [5].

To estimate the wind speed in India, data from 31 provinces with differentiated geographical conditions from National Aeronautics and Space Administration (NASA) and the Generalized Regression Neural Networks (GRNN) model were used. 26 provincial data were used as training data and the other 5 provincial data were used as test data. Based on CC and MSE values, GRNN model was found to be very efficient for predicting long-term wind speed [6]. Zeng et.al. studied the effect of different sampling frequencies on short-term wind speed determination and power estimation with support vector machines [7]. Khandelwal et.al. have developed a time series forecasting model that combines discrete wavelet transforms with ARIMA and artificial neural network model [8]. Khanna et.al. studied the determination of time series properties in wind power generation [9]. For wind turbines in Korea, wind energy was estimated with ANFIS, Sequential Minimal Optimization (SMO), K-Nearest Neighbor (KNN) and Artificial Neural Networks (ANN) models using hourly and daily wind speed, wind direction, temperature and time intervals. ANFIS and SMO models have been shown to perform well in predicting wind energy [10]. Wanga et.al. used copula theory to estimate wind speed. They have determined that representative wind data can be obtained much more reasonably with the conditional distribution [11]. Velo et.al have presented a method for determining the annual average wind speed in a complex land area using neural networks, where only short term data are available. As neural network inputs, they used wind speed and direction data obtained from a single station [12]. Timur et.al. estimated the wind speed using LinearRegression, K-Nearest Neighbor (KNN), Bagging, DecisionTable, and REPTree classification algorithms for Istanbul Göztepe region. In the cross-validation option, 5 and 10 values were given to k in the cross validation option, and the Bagging algorithm was observed to be more successful in terms of the CC and RMSE (Root Mean Square Error) values of the most successful result [13]. In a study Zontul et al. Between 2001 and 2007, wind speed estimation with WEKA was performed by using cross-validation method with yearly, monthly, daily wind direction data were obtained from meteorological data of Kırklareli Province. The correlation coefficient between the true value and the estimated value was found to be a successful classification method of the Bagging model with 0.8154 [14]. Based on the Wavelet Packet Transformation (WPD), the Cross-Optimization (CSO) Algorithm and the Artificial Neural Networks, Meng et.al. have developed a new hybrid model to predict a short-term wind speed of 1 hour intervals up to a 5-hour period with two different wind speed ranges in the wind observation station in Rotterdam [15].

In this study, the wind speeds, which are accepted as the most important inputs of wind energy, were estimated by using machine learning algorithms. Six different algorithms were used in the estimation and Bagging algorithm realized the best estimation with the lowest error and highest correlation coefficient (CC) among these algorithms. The meteorological parameters that affect the estimation of wind speed in the study were also examined and results indicate that the combination of temperature-humidity-pressure combination realized a prediction with lower error rate.

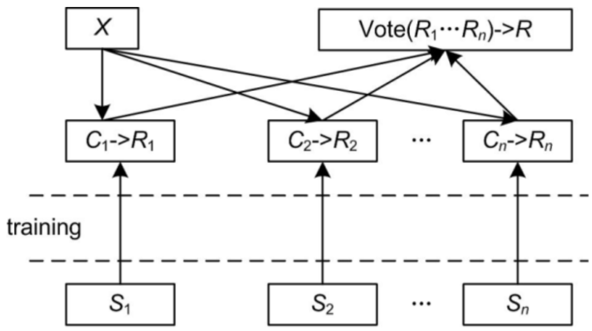
2. MATERIALS AND METHODS

In this study; the data mining algorithms used for wind speed estimation are Bagging Algorithm, SMOreg Algorithm, K-Star Algorithm, Multilayer Perceptron, REP Tree Algorithm and M5P algorithm.

2.1. Bagging Algorithm

Bagging algorithm is a method that is used to improve results are obtained from machine learning classification algorithms. Leo Breiman formulated the Bagging algorithm in 1994. This algorithm is an abbreviation of "bootstrap aggregating" [16]. The generalization capability of machine learning algorithms can be evolved by using the Bagging algorithm. The bagging algorithm is analyzing in a shorter time using parallel learning as an alternative to algorithms that have long analysis time, such as artificial neural networks. The process of the Bagging algorithm is shown in Figure 1 [17]. Sample sets (S_i) are created by using random return selection. The corresponding classifier C_i is trained based on the training set S_i, respectively. In this way, the final pattern recognition can be obtained by using the primary recognition results.

Fig. 1. Bagging algorithm description



The bagging method creates a sequence of classifiers C_i, i=1,…,M in respect to modifications of the training set. These classifiers are combined into a compound classifier. The prediction of the compound classifier is given as a weighted combination of individual classifier predictions [18]:

$$C(d_m) = sign\left(\sum_{i=1}^{M} \alpha_i C_i(d_m)\right) \quad (1)$$

The meaning of the above given formula can be interpreted as a voting procedure. An example d_m is classified to the class for which the majority of particular classifiers vote. Parameters α_i, i=1,…,M are determined in such way that more precise classifiers have stronger influence on the final prediction than less precise classifiers. Bagging decrease the variance of your single estimate so the result may be a model with higher stability.

2.2. SMOreg Algorithm

SMOreg Algorithm is used with Support Vector Machine (SVM) that is classification method that divides the data into two categories. There is a training data set for classification process. This training data belongs to any of the two categories. After that SVM estimates that new instance is in which category. It is aimed to create a hyperplane in a n-dimensional space in this process. Whereby two datasets can be separated which data is nearest to hyper plane is named support vector. Finally, it is checked that data is in whether right side or left side of the hyper plane in order to classify the new test data [19]. SMOreg uses the sequential minimal optimization algorithm for training a support vector classifier. The following equation is used in the optimization process [20]

$$\max \Psi(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j k(x_i, x_j)\alpha_i \alpha_j \quad (2)$$

Initial conditions; $\sum_{i=1}^{N} y_i \alpha_i = 0,\ \ 0 \le \alpha_i \le c \ and\ i = 1, \dots, n$
Where x_i is trainin sample, $y_i \in \{-1, +1\}$ is target value, α_i is Lagrange multiplier and c is actual cost parameter value. In this algorithm, the planar kernel function, $k(x_i, x_j) = x_i^T * x_j$, was used.
It preferes to use Gaussian or Polynomial kernels during this process [21]. In this study, we set the 'exponent' property to 2 in dialog of WEKA software for SMOreg algorithm.This is necessary, though, to force WEKA to use support vectors. While SMOreg is using, all the attributes are normalized into binary attributes [22].

2.3. K-Star Algorithm

K-star is an sample-based classifier that uses entropy based distance function. Entropy is useful method to evaluate distance [23]. An instance based algorithm made for symbolic attributes fails in features of real value thus lacking in incorporated theoretical base. Therefore this situation makes K-star algorithm that use entropy based distance function quite important.
Description of K-Star Algorithm [24]:
Let I : infinite set of instances
T is a finite set of transformations on I;
P set of all prefix codes from T* which are determined by σ(the stop symbol)
Members of T* uniquely define a transformation on

$$I: \bar{t}(a) = t_n(t_{n-1})(\dots . t_1(a) \dots .) \text{ Where } \bar{t} = t_1 \dots \dots . t_n \quad (3)$$

A probability function p is defined on T*.It satisfies the following properties:
$0 \le (\bar{t}\bar{u})p(t) \le 1$ (4)

$$\Sigma u\ p(\bar{t}\bar{u})\ \ p(\bar{t}) \quad (5)$$

$$p(\Lambda) = 1 \quad (6)$$

as a consequence, it satisfies the following

$$\Sigma p(\bar{t} \in p\bar{t}) = 1 \quad (7)$$

The probability function $P*$ is defined as probability of all paths from instance a to instance b

$$P^*(b\,|a)\ \Sigma\ \bar{t} \in p\!:\ \bar{t}(a) = b\mathrm{P}(\bar{t}) \quad (8)$$

$$\Sigma P * b(b|a) = 1 \quad (9)$$

$$0 \le P^*(b|a) = 1 \quad (10)$$

The K* function is then defined as

$$K^*(b|a) = -\log 2P*(b|a) \quad (11)$$

The basic assumption is that similar examples have similar classifications. For this reason, it is necessary to define "similar classifications" and "similar samples" well.

2.4. Multi-Layer Perceptron

The multilayer perceptron (MLP) is a feedback method used the artificial neural network. The Multi-Layer Perceptron consists nodes and neurons [25].

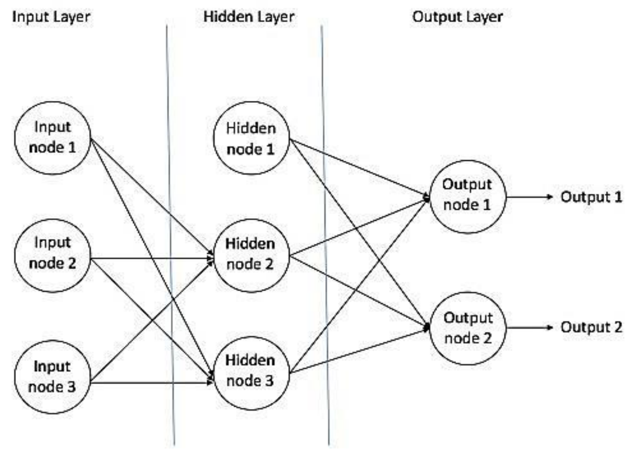


Fig. 2. Multilayer perceptron

Nodes and neurons are placed in three layer (input layer, hidden layer and output layer). The nodes are connected by weights. Each neuron has mathematical function. Each node outputs an activation function applied over the weighted sum of its inputs:

$$\boldsymbol{S_i = f(w_{i,0} + \sum_{j \in I} w_{i,j} \times s_j)} \quad (12)$$

MLP uses a nonlinear activation function to learn nonlinear function mappings between input and output. These activation functions :

Linear: y = x; (13)

Tanh: y = tanh (x) (14)

Logistic (or sigmoid): $y = 1/(1+e^{-x})$ (15)

The input information of a neuron is processed. Subsequently output information of neuron is used as input information for other neuron that is in the next layer. MLP weighs the last relation until the last node is in the last layer. After finding, the method calculates the error and send backward to remodify the model [26]. The neural network used in this study consists of 3 layers.These are the input layer, the hidden layer, and the output layer.The output layer has a single output neuron.When the studies in the literature were examined, it was observed that using 10 neurons in the hidden layer was good in estimating. The number of neurons in the input layer ranged from 1 to 3 according to the groups formed in Table 2 in the 3rd section.

2.5. REP Tree Algorithm

RepTree algorithm prefers to use the regression tree logic. This algorithm constitutes multiple trees iterations. At the end of the iterations the best tree is selected among generated trees. The accuracy of selection is tested by mean square error. Also the pseudocode of the basic operation of REPTree is shown below in algorithms 1 and 2 [27].

Data: Dataset D with a set of attributes A
Result: Decision tree using REPTree
if *use pruning* **then**
    Split D into training data Dt and pruning data Dp;
else
    Training data Dt = D;
end
Build a tree using Dt and A as shown in Algorithm 2;
if *use pruning* **then**
    Reduce error pruning using Dp;
end
Algorithm 1: REPTree algorithm

Data: Dataset D and a set of attributes A
Result: A Decision tree Tree
if *no stop condition is reached* **then**
    Compute splitting criterion, SC(D,ai), for each attribute ai ∈ A;
Find the best attribute ab according to the splitting criterion;
Using ab, split D in n subsets;
if *max* SC > *0* **and** n > *1* **then**
    foreach *of the* n *subset of* Di **do**
        Tree = BuildTree using Di and A;
            end
        end
else
        Tree = Create a leaf using D;
end

Algorithm 2: BuildTree algorithm (REPTree)

RepTree is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance [28]. Values of numeric attributes are classified once by RepTree algorithm. Reptree algorithm is used to solve both regression and classification problems.

2.6. M5P Algorithm

M5P is the improved model of M5 algorithm that was created by Quinlan. The main advantage of M5P algotrithm is that this algorithm can efficiently handle large number of data sets with high dimensions [29]. If training set is small, there could be high classification error rate when comparing with the number of classes. Parameter setting is not require for M5P algorithm. So this algorithm does not need knowledge discovery [30]. A M5P model tree is shown in Figure 3 [31].

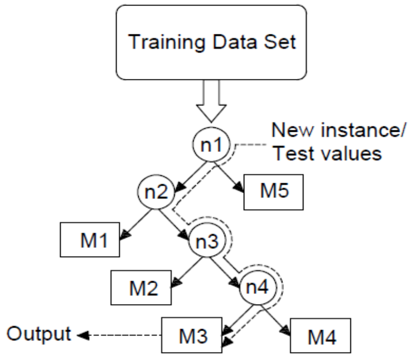


Fig. 3. A M5P model tree, *ni* are split nodes and *Mi* are the models

M5P algorithm is fast, simple and have a good accuracy during process. M5P creates classification and regression trees by using a multivariate linear regression model. So it can minimize the variation within a particular sub-space. These model trees resemble piecewise linear functions. M5P algorithm is also known as robust algorithm when dealing with missing data.

2.7. Description of Location and Dataset Measurement

The pressure, wind speed, temperature and humidity data that were used as inputs for both classification and estimation model were obtained from measurement station that was established. Measurement station was placed in latitude (N 40º19′58.73″) longitude (E 36º29′0.28″). Collage photo of measurement station is shown in Figure 4.

Fig. 4. Application region of measurement station

It is seen that the measurement site is located in a region with woodland and agricultural areas. The vegetation in the region consists of short herbaceous plants and rarely straightened trees. Measurement mast with a height of 12 meters was used in station. It is shown in Figure 5. Two wind speed sensor and one wind direction sensor were placed on mast. Pressure, temperature and humidity sensors were placed in power box. Also data logger was placed in power box, shown in Figure 6. Solar panel which has 10 W was used to meet energy needs of sensors.
The data (wind speed, pressure, humidity and temperature) were used in this study were collected at time interval of 10 minute throughout 2017. The wind speed, pressure, temperature and

humidity data which are "dat" format were converted into "arff" format to process in WEKA software. "WEKA" stands for the Waikato Environment for Knowledge Analysis, is developed by University of Waikato, New Zealand in 1993. WEKA is a collection of machine learning algorithms for solving real-world data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules and visualization [32].

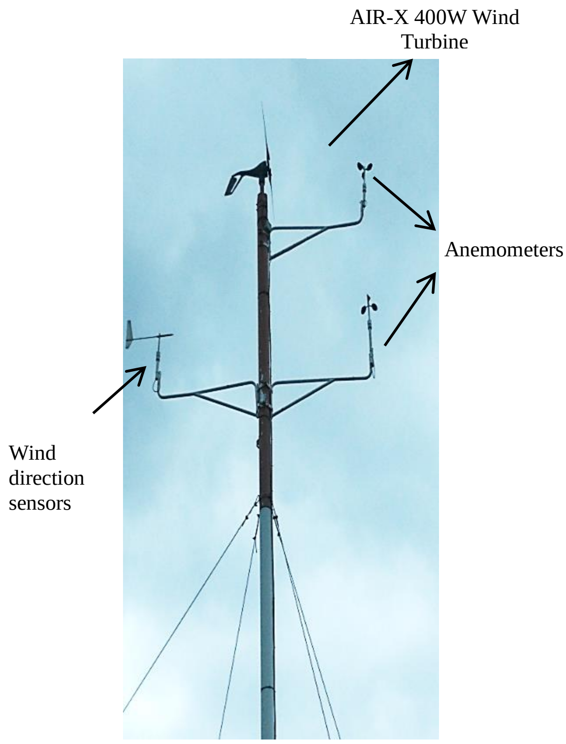


Fig. 5. Measurement station

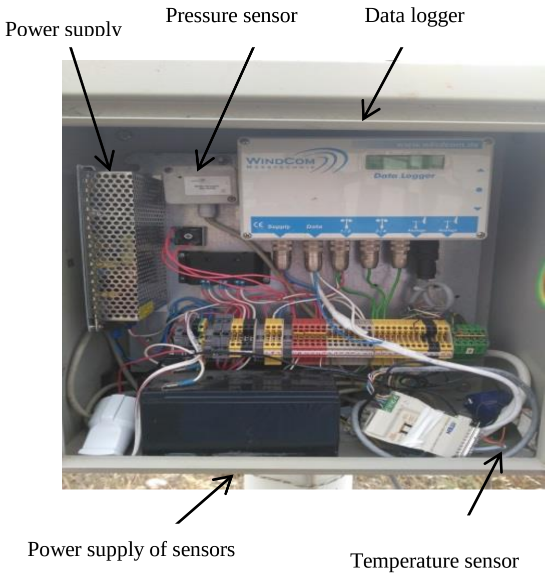


Fig. 6. Power box

3. IMPLEMENTATION OF MACHINE LEARNING ALGORITHMS

At first the missing or inconsistent data in the database were extracted because this kind of data must be extracted to obtain successful results. Also it is necessary to remove data repetitions from the database and increase the data consistency (correctness). So all data were normalized.

Seven combinations were created from the normalized data (pressure, humidity and temperature). Groups are given in Table 2.

Table2 Combinations

Combinations	
1. Combination	Temperature-Humiditiy-Pressure
2. Combination	Temperature-Humiditiy
3. Combination	Temperature- Pressure
4. Combination	Humiditiy-Pressure
5. Combination	Temperature
6. Combination	Humiditiy
7. Combination	Pressure

In order to find the most efficiency method we used 10-fold cross validation technique which divided data into ten sets of size n/10 (n is number of records). Every data set is tested by using the remaining sets as its training set. It is shown in Figure 7. The 52560 data were used in the estimation process.

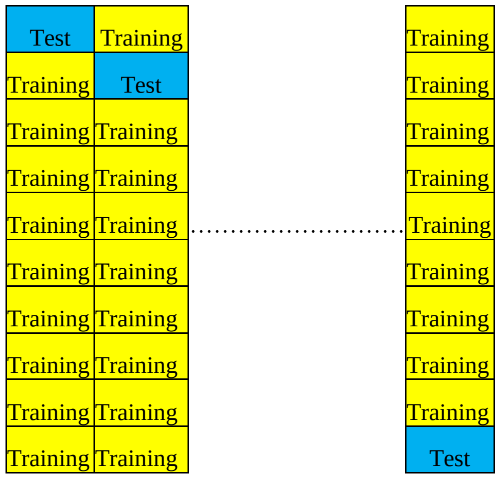


Fig. 7. Schematic representation of 10-fold cross-validation

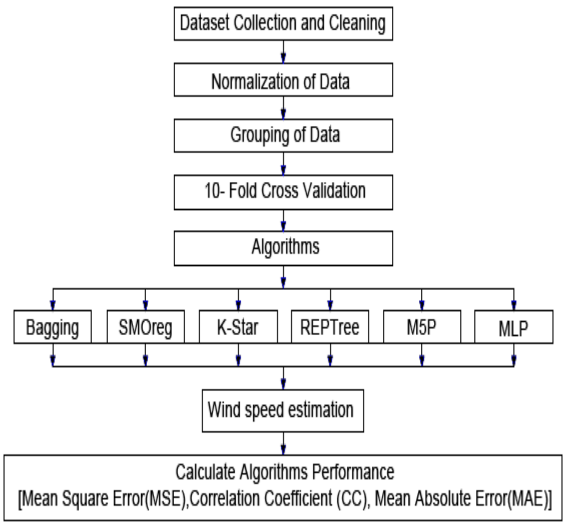
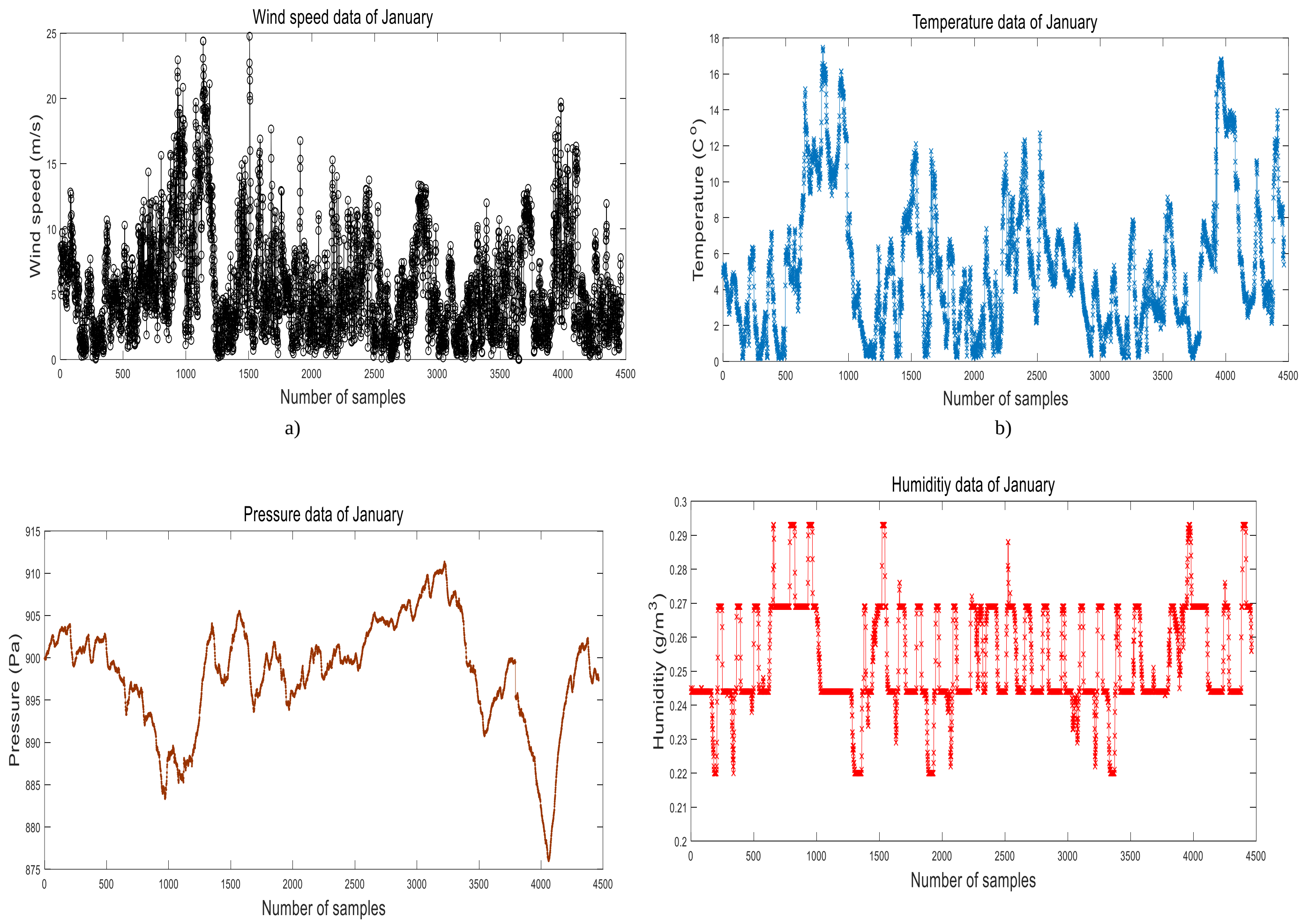


Fig. 8. Flow chart of estimation model

Cross-validation is a technique to evaulate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In 10 fold cross-validation, the orginal sample is randomly partitioned into 10 equal size subsamples.

Of the 10 subsamples, a single subsample is retained as the validation data for testing the model and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as validation data. The 10 results from the folds can then be averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and validation and each observation is used for validation exactly once. Our comparing methods that are used to evaluate algorithms performance are described in detail as follows chapter 4. The flow chart of estimation model that is established to examine effects of meteorological parameters is shown in Figure 8. Since the data set we used in this study is very large, only 4464 wind speed, temperature, humidity and pressure data of January are shown in Figure 9.



c) d)

Fig. 9. a) Wind speed b) Temperature c) Pressure and d) Humidity data of January

4. RESULTS AND DISCUSSION

As explained in Chapter 3, seven combinations were tested for all algorithms. The performances of the algorithms were evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Correlation coefficient (CC) statistical parameters. Root mean square error is defined as square root of sum of squares error divided number of predictions. It is a frequently used to measure the differences between values predicted by a model and the values actually observed. It is formulated as given in Equation 2.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \quad (2)$$

x_i is actual wind speed and y_i is estimated wind speed.

Mean absolute error can be defined as sum of absolute errors divided by number of predictions. It is measure set of estimated value to actual value i.e. how close a estimated model to actual model. MAE is calculated using Equation 3.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \quad (3)$$

Small value of RMSE means better accuracy of model. So, minimum of RMSE & MAE is better estimation and accuracy. Correlation coefficient measures the statistical relationship between the two variables (the actual value and the estimated value). The correlation coefficient takes values between -1 and +1. A value close to 1 indicates a good relationship. A value close to -1 indicates a weak relationship. If the result is 0, it indicates that there is no relation between the two variables. Equation of correlation coefficient is given Equation 4.

$$Correlation\ coefficient = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \qquad (4)$$

There is a direct relationship between the correlation coefficient and RMSE. If the correlation coefficient is one, the RMSE will be zero because all points are in the regression line.

The analyzes were carried out in three stages. These are monthly, annual and seasonal analysis. Firstly, monthly analysis were performed. Analysis results of first combination are given Table 3 since more lower error values were obtained with estimation analysis that was made with first combination (temperature-humiditiy-pressure). Based on the obtained analysis results, bagging algorithm showed the best performance in all months. Mean values of MAE, RMSE and CC parameters that were calculated by using Bagging algorithm are 0.084,0.112 and 0.702 respectively.

Table 3 Statistical analysis results of months

Algorithms	**MAE, RMSE and CC Values**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
Multilayer Perceptron	MAE	0.1341	0.1334	0.1458	0.1433	0.1388	0.1390
	RMSE	0.1676	0.171	0.1805	0.1808	0.1769	0.1772
	CC	0.2374	0.1094	0.216	0.0705	0.1147	0.1612
SMOreg	MAE	0.1116	0.1122	0.1312	0.1333	0.1249	0.1295
	RMSE	0.1515	0.1543	0.1705	0.1722	0.1644	0.1699
	CC	0.3839	0.2229	0.2225	0.0354	0.1322	0.2225
Kstar	MAE	0.0915	0.098	0.1102	0.1111	0.1041	0.1059
	RMSE	0.118	0.1255	0.139	0.1417	0.135	0.1372
	CC	0.6985	0.6049	0.6053	0.5948	0.5827	0.6237
Bagging	MAE	**0.0757**	**0.0811**	**0.0959**	**0.0907**	**0.0855**	**0.0861**
	RMSE	**0.1011**	**0.1103**	**0.126**	**0.1216**	**0.1184**	**0.1185**
	CC	**0.7774**	**0.6982**	**0.6802**	**0.7032**	**0.6874**	**0.7215**
M5P	MAE	0.0858	0.0929	0.1076	0.1036	0.1001	0.1019
	RMSE	0.1122	0.1216	0.1386	0.1337	0.1321	0.1333
	CC	0.717	0.6145	0.5909	0.6236	0.5846	0.6279
REPTree	MAE	0.0843	0.0906	0.103	0.1001	0.0983	0.0941
	RMSE	0.1146	0.1227	0.1379	0.1371	0.1368	0.1324
	CC	0.7087	0.6148	0.6087	0.6111	0.5691	0.6445

Table 3 Continued

Algorithms	**MAE, RMSE and CC Values**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Multilayer Perceptron	MAE	0.1693	0.1478	0.1178	0.0908	0.1157	0.1040
	RMSE	0.2072	0.1834	0.1468	0.112	0.1491	0.1451
	CC	0.3509	0.3542	0.3253	0.1292	0.2653	0.2187
SMOreg	MAE	0.1573	0.1416	0.1165	0.0751	0.0992	0.1008
	RMSE	0.1880	0.1742	0.145	0.0974	0.1331	0.1445
	CC	0.4486	0.3686	0.3464	0.1448	0.4089	0.2505
Kstar	MAE	0.1280	0.1186	0.1032	0.0654	0.0874	0.0851
	RMSE	0.157	0.1458	0.128	0.084	0.1136	0.1146
	CC	0.676	0.6361	0.5669	0.5289	0.6274	0.6379
Bagging	MAE	**0.1044**	**0.1032**	**0.0899**	**0.0588**	**0.0721**	**0.069**
	RMSE	**0.1396**	**0.1332**	**0.1164**	**0.0775**	**0.0983**	**0.0981**
	CC	**0.7468**	**0.6991**	**0.6522**	**0.6064**	**0.7289**	**0.7363**
M5P	MAE	0.1207	0.1145	0.0994	0.0642	0.0831	0.0789
	RMSE	0.1548	0.1439	0.1253	0.0826	0.1094	0.1081
	CC	0.6754	0.6346	0.578	0.5312	0.6467	0.6673
REPTree	MAE	0.1153	0.1104	0.0991	0.0641	0.0797	0.0761
	RMSE	0.1557	0.1436	0.1288	0.0848	0.1108	0.1092
	CC	0.6779	0.6462	0.5612	0.5214	0.6461	0.6678

It is possible to say that the Bagging algorithm and the first combination are most efficient (the highest correlation coefficient, the lowest RMSE and MAE values) in the monthly estimation. Apart from the Bagging algorithm, also favorable results were obtained by using the MP5 algorithm in monthly estimation.

According to months the variations of correlation coefficient and RMSE values that were calculated by using the first combination are shown in Fig 10-11. It is seen in Fig. 10-11 that the highest correlation coefficient value and the lowest RMSE values were obtained in the bagging algorithm

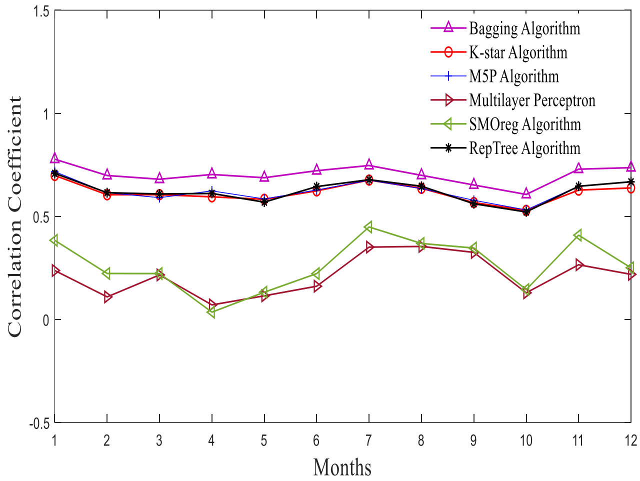


Fig. 10. The variations of correlation coefficient according to months

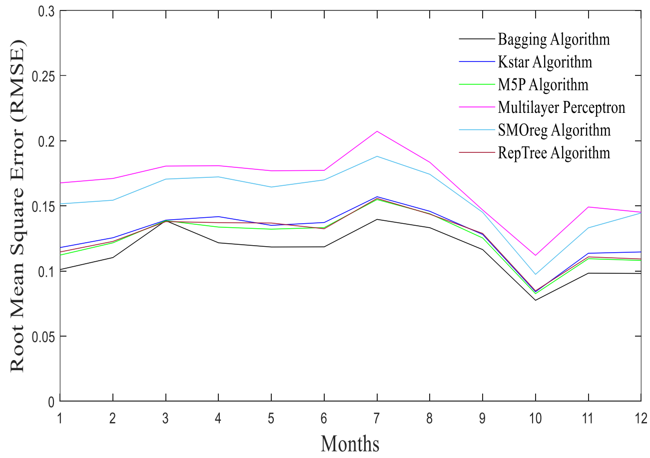


Fig. 11. The variations of RMSE values according to months

In the second stage, annual analysis were carried out. Analysis results according to combinations are shown in Table 4. In the annual analysis, the lowest RMSE value was obtained in the first combination (Temperature-Humidity-Pressure). In the three algorithms (MultilayerPerceptron, SMOreg, Kstar) different values were obtained in different combinations but the values that were obtained in the combinations of these algorithms are worse than the values that were obtained in the first combination of the bagging algorithm. Good results were obtained again with the bagging algorithm in all combinations. The worst results were obtained by multilayer perceptron algorithm.

In the third stage, seasonal analyzes were carried out. The results of analyzes of all seasons are shown in Table 5-8. In the anaysis for all seasons, better results were obtained with the Bagging algorithm and the first combination.

Table 4 Statistical results of annual analysis according to combinations

Algorithms	MAE, RMSE and CC Values	Temperature Humidity Pressure	Temperature Humidity	Temperature Pressure	Humidity Pressure	Temperature	Humidity	Pressure
Multilayer Perceptron	MAE	0.1041	0.1035	0.1041	0.1047	**0.1029**	0.1035	0.1056
	RMSE	0.1327	0.1322	0.1328	0.1335	**0.1318**	0.1323	0.1348
	CC	**0.144**	0.1321	0.1348	0.1328	0.1317	0.1311	0.0815
SMOreg	MAE	**0.0874**	0.0892	0.0956	0.0941	0.0941	0.0985	0.0978
	RMSE	**0.1225**	0.1232	0.1226	0.1233	0.1233	0.1234	0.1252
	CC	0.2096	0.2704	**0.2818**	0.2668	0.2668	0.2674	0.2164
Kstar	MAE	0.0916	**0.0836**	0.0922	0.0978	0.0879	0.0898	0.0947
	RMSE	**0.1153**	0.1202	0.1175	0.119	0.1208	0.1212	0.1225
	CC	**0.4134**	0.2917	0.3695	0.3305	0.2764	0.2697	0.2319
Bagging	MAE	**0.0716**	0.0883	0.0767	0.085	0.0914	0.0955	0.0943
	RMSE	**0.0969**	0.1149	0.1026	0.1112	0.1179	0.1205	0.1209
	CC	**0.6358**	0.4099	0.5769	0.4676	0.3567	0.2801	0.2958
M5P	MAE	**0.0831**	0.0936	0.0869	0.0913	0.0953	0.0959	0.0964
	RMSE	**0.1076**	0.1185	0.1115	0.1159	0.1204	0.1208	0.1216
	CC	**0.5155**	0.3303	0.4598	0.3857	0.283	0.272	0.2491
REPTree	MAE	**0.078**	0.0916	0.0828	0.0886	0.0941	0.0956	0.096
	RMSE	**0.1062**	0.1186	0.111	0.1157	0.1213	0.1206	0.1228
	CC	**0.5502**	0.3523	0.488	0.4107	0.2971	0.2785	0.2599

Table 5 Statistical analysis results of spring

Algorithms	MAE, RMSE and CC Values	Temperature Humidity Pressure	Temperature Humidity	Temperature Pressure	Humidity Pressure	Temperature	Humidity	Pressure
Multilayer Perceptron	MAE	0.1424	0.1442	0.1426	0.1435	0.1441	0.1451	0.1424
	RMSE	0.1771	0.1786	0.1772	0.1779	0.1786	0.1796	0.1778
	CC	0.1367	0.1017	0.1007	0.0978	0.0982	0.095	0.064
SMOreg	MAE	0.124	0.1248	0.124	0.1244	0.1249	0.126	0.1253
	RMSE	0.1622	0.1631	0.1621	0.1624	0.1629	0.1637	0.164
	CC	0.2319	0.2114	0.2344	0.2269	0.2167	0.1991	0.1804
Kstar	MAE	0.1151	0.1254	0.1206	0.1215	0.1267	0.1272	0.1261
	RMSE	0.1446	0.1571	0.1513	0.1521	0.159	0.1587	0.1588
	CC	0.4959	0.2897	0.4049	0.3893	0.2475	0.2547	0.2607
Bagging	MAE	**0.0926**	0.1208	0.101	0.1095	0.1259	0.1252	0.1249
	RMSE	**0.1238**	0.1536	0.1338	0.1429	0.1594	0.1575	0.1597
	CC	**0.6555**	0.3629	0.5771	0.4944	0.2688	0.2768	0.272
M5P	MAE	0.1051	0.1224	0.1126	0.1143	0.1253	0.1255	0.1241
	RMSE	0.1356	0.1539	0.1444	0.1459	0.1575	0.1579	0.1567
	CC	0.5618	0.3431	0.4733	0.4551	0.2764	0.2681	0.2926
REPTree	MAE	0.101	0.1227	0.109	0.1136	0.1261	0.1254	0.1256
	RMSE	0.1366	0.1556	0.1449	0.1472	0.1593	0.1577	0.1595
	CC	0.5685	0.3346	0.4921	0.4537	0.2522	0.2718	0.2546

Table 6 Statistical analysis results of summer

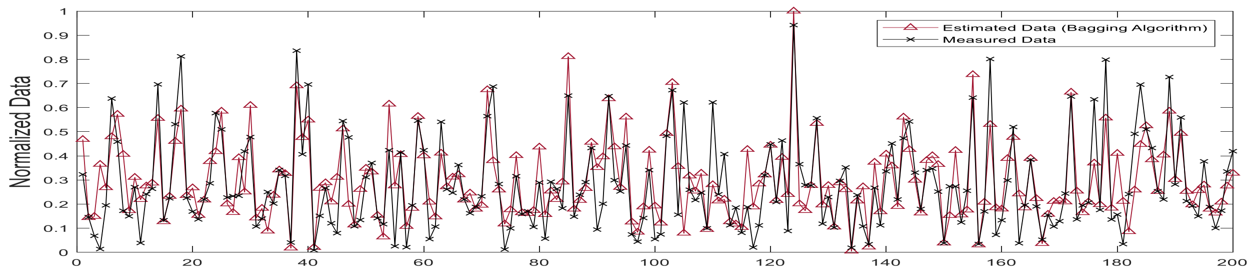
Algorithms	MAE, RMSE and CC Values	Temperature Humidity Pressure	Temperature Humidity	Temperature Pressure	Humidity Pressure	Temperature	Humidity	Pressure
Multilayer Perceptron	MAE	0.1441	0.1441	0.1455	0.1442	0.1455	0.1442	0.1492
	RMSE	0.1824	0.1823	0.1839	0.1825	0.1838	0.1824	0.1888
	CC	0.156	0.1522	0.13	0.1502	0.13	0.1517	0.0068
SMOreg	MAE	0.1342	0.1344	0.1345	0.1342	0.1347	0.1344	0.1382
	RMSE	0.1699	0.1702	0.17	0.1701	0.1703	0.1703	0.1775
	CC	0.2711	0.2704	0.2667	0.2675	0.266	0.268	-0.0207
Kstar	MAE	0.116	0.1326	0.1217	0.1288	0.1051	0.1363	0.1377
	RMSE	0.1469	0.1641	0.154	0.1609	0.1422	0.1675	0.1711
	CC	0.5642	0.3353	0.4981	0.3978	0.5755	0.2728	0.2045
Bagging	MAE	**0.0797**	0.1099	0.0874	0.1051	0.1164	0.1356	0.1267
	RMSE	**0.1133**	0.1457	0.1243	0.1422	0.153	0.1671	0.1644
	CC	**0.761**	0.546	0.6998	0.5755	0.4771	0.2781	0.3415
M5P	MAE	0.1068	0.1283	0.1124	0.1202	0.1326	0.1362	0.135
	RMSE	0.1379	0.1604	0.1452	0.1202	0.1646	0.1672	0.1688
	CC	0.6099	0.3866	0.5509	0.4719	0.3231	0.2751	0.2417
REPTree	MAE	0.0873	0.1184	0.0939	0.113	0.1247	0.1358	0.1312
	RMSE	0.1267	0.1574	0.1353	0.1523	0.1625	0.1674	0.1689
	CC	0.6906	0.4515	0.6374	0.4979	0.3879	0.2724	0.2917

Table 7 Statistical analysis results of autumn

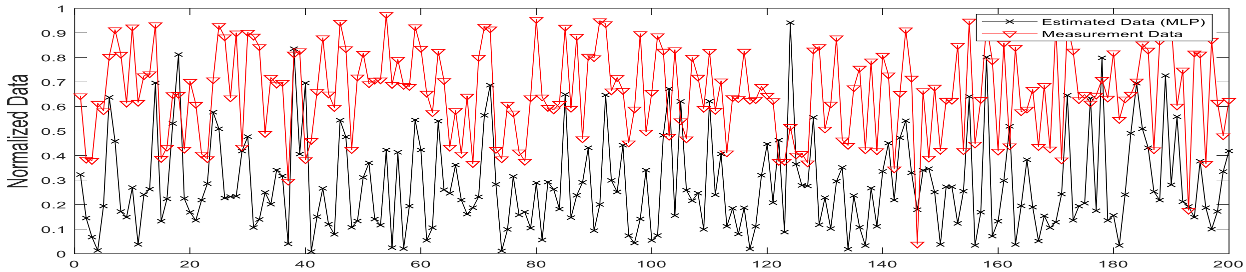
Algorithms	MAE, RMSE and CC Values	Temperature Humidity Pressure	Temperature Humidity	Temperature Pressure	Humidity Pressure	Temperature	Humidity	Pressure
Multilayer Perceptron	MAE	0.1032	0.1038	0.1037	0.1045	0.1037	0.1044	0.1092
	RMSE	0.1269	0.1273	0.1273	0.1281	0.1274	0.128	0.1358
	CC	0.1909	0.1742	0.17	0.1641	0.1702	0.1645	0.0203
SMOreg	MAE	0.0897	0.0898	0.0898	0.09	0.0898	0.09	0.093
	RMSE	0.1135	0.1135	0.1134	0.1141	0.1135	0.1141	0.1201
	CC	0.3263	0.3252	0.3269	0.3107	0.326	0.3111	0.1087
Kstar	MAE	0.083	0.0893	0.0878	0.0878	0.0901	0.0906	0.0944
	RMSE	0.1035	0.1102	0.1087	0.1087	0.1114	0.1119	0.1173
	CC	0.4942	0.3679	0.4037	0.4037	0.3452	0.333	0.1456
Bagging	MAE	**0.0737**	0.0888	0.0848	0.0848	0.0918	0.0902	0.0955
	RMSE	**0.0956**	0.1113	0.1077	0.1077	0.1146	0.1115	0.1199
	CC	**0.5902**	0.3605	0.4273	0.4273	0.2915	0.3379	0.1313
M5P	MAE	0.0795	0.0882	0.086	0.086	0.0895	0.09	0.0943
	RMSE	0.1011	0.1096	0.1075	0.1075	0.1109	0.1113	0.1171
	CC	0.5212	0.379	0.4195	0.4195	0.3506	0.3413	0.1492
REPTree	MAE	0.0788	0.0878	0.0856	0.0856	0.0898	0.0901	0.0949
	RMSE	0.1027	0.1098	0.1083	0.1083	0.1114	0.1113	0.1186
	CC	0.5143	0.379	0.4143	0.4143	0.3416	0.342	0.1276

Table 8 Statistical analysis results of winter

Algorithms	MAE, RMSE and CC Values	Temperature Humidity Pressure	Temperature Humidity	Temperature Pressure	Humidity Pressure	Temperature	Humidity	Pressure
Multilayer Perceptron	MAE	0.1211	0.1223	0.1222	0.1222	0.1236	0.1226	0.123
	RMSE	0.1551	0.1583	0.1559	0.1562	0.1593	0.1586	0.1568
	CC	0.1325	0.0735	0.1277	0.1279	0.0588	0.0679	0.1163
SMOreg	MAE	0.0972	0.0993	0.0974	0.0972	0.0996	0.0995	0.099
	RMSE	0.1341	0.1389	0.1342	0.1341	0.1392	0.1393	0.1345
	CC	0.3395	0.2367	0.3373	0.3397	0.2278	0.2301	0.3308
Kstar	MAE	0.0894	0.1011	0.0924	0.0966	0.1017	0.1029	0.1002
	RMSE	0.1179	0.1334	0.1212	0.1271	0.1345	0.1355	0.1309
	CC	0.5594	0.2906	0.5176	0.4352	0.2619	0.2367	0.3617
Bagging	MAE	**0.0747**	0.1004	0.0793	0.0905	0.1036	0.1021	0.1000
	RMSE	**0.1021**	0.1339	0.1073	0.121	0.1371	0.1348	0.1311
	CC	**0.68**	0.2987	0.6372	0.499	0.2329	0.2499	0.3564
M5P	MAE	0.084	0.0998	0.0874	0.0929	0.1016	0.1029	0.0989
	RMSE	0.1116	0.1321	0.1148	0.1222	0.1338	0.1354	0.1285
	CC	0.5974	0.3157	0.5659	0.4788	0.2774	0.2309	0.3846
REPTree	MAE	0.0813	0.1004	0.0852	0.0922	0.1026	0.1021	0.0998
	RMSE	0.1122	0.1335	0.1157	0.1231	0.1354	0.1349	0.1303
	CC	0.604	0.2984	0.5678	0.4772	0.2459	0.2484	0.3615



a)



b)

Fig. 12. The changes of between estimated and measurement values according to algorithms. **a)** Bagging Algorithm **b)** Multilayer Perceptron Algorithm

Summer was chosen to see graphically the variation between the estimated and measured data. The changes of the 200 measured and estimated data that were randomly selected for summer are shown in Fig 12. According to bagging and multilayer perceptron algorithms, which have the best and lowest performance, the differences of between Bagging and MLP algorithms are clearly seen.

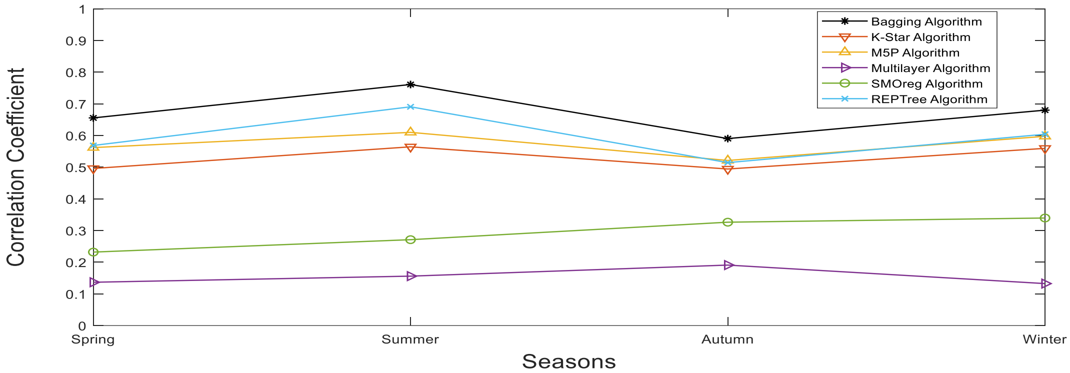


Fig. 13. The variations of correlation coefficient according to seasons

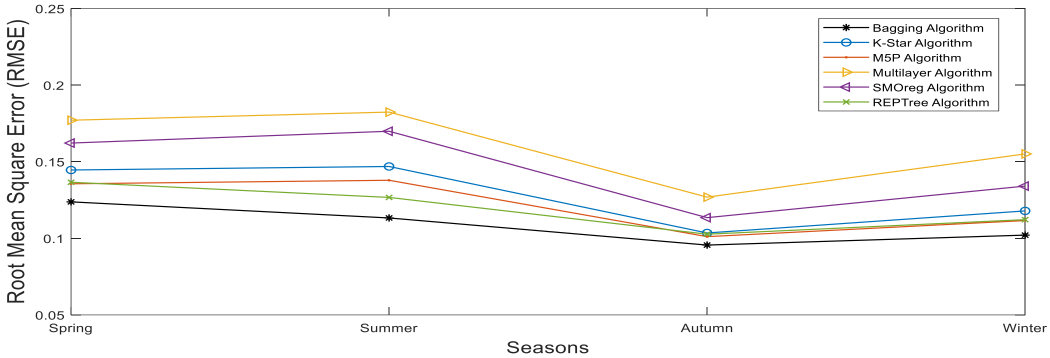


Fig. 14. The variations of RMS values according to seasons

According to seasons the variations of the RMSE values and correlation coefficient that were calculated by using the first combination are clearly shown in Fig 13-14. In seasonal analysis, the highest correlation coefficient value and the lowest RMSE values were all over obtained in the bagging algorithm.

5. CONCLUSION

In this study, Bagging algorithm is proposed to examine the effect of meteorological parameters on wind speed. Bagging algorithm has many advantages than the other machine learning algorithms. Bagging reduces variance or model inconsistency over diverse data sets from a given distribution, without increasing bias, which results in a reduced overall generalization error and enhanced stability.The other benefit of using bagging is related to the model selection. Since bagging transforms a group of over fitted neural networks into a better-than perfectly fitted network, the tedious time consuming model selection is no longer required. This could even offset the computational overhead needed in bagging that involves training many neural networks. Also Bagging is very robust to noise. For that reasons when the bagging algorithm is compared with other conventional machine learning algorithms, more efficient results were obtained with this algorithm in this study. The measured data were divided into seven combinations to examine the effect of meteorological parameters and the statistical performances of the algorithms used with each combination were evaluated according to RMSE, CC and MAE criteria. The best results were obtained in combination that were used temperature pressure and humidity data for wind speed estimation. It was obtained that temperature and pressure parameters are more effective in wind speed estimation.

The RMSE value was reduced % 17.81 respect to the temperature parameter, %19.85 respect to the pressure parameter and 19.58% respect to the humidity parameter with the combination used in the annual analyzes. As a result, all meteorological parameters should be used in correct estimation of the wind speed. It is aimed to examine the effect of other meteorological parameters such as altitude, wind direction, air density in the subsequent studies.

References

[1] REN21, RENEWABLES 2017 GLOBAL STATUS REPORT, 2017, [Online]. Available: http://www.ren21.net/wp-content/uploads/2017/06/178399_GSR_2017_Full_Report_0621_Opt.pdf.; Accessed on June 28 2018.

[2] WEC, Global Wind Statistics, Global Wind Energy Councıl,2016, [Online].Available:http://www.gwec.net/wpcontent/uploads/vip/GWEC_PRstats2016_EN_WEB.pdf.; Accessed on June 28 2018.

[3] I. Kırbas, "Short-term multi-step wind speed prediction using statistical methods and artificial neural networks," SAKARYA UNIVERSITY JOURNAL OF SCIENCE, 22 (1), 24-38, 2018.

[4] S. H. Mirhashemi, M. Tabatabayi, "Evaluation ten models of weka software to predict the potential evapotranspiration month for the next month in the synoptic weather station Arak," International Journal of Agriculture and Crop Sciences, 8(1), 5, 2015.

[5] E. Cadenas, W. Rivera, R. Campos-Amezcua, C. Heard, "Wind speed prediction using a univariate ARIMA model and a

multivariate NARX model," Energies, 9(2), 109, 2016.
[6] M. A. Ansari, N. S. Pal, H. Malik, "Wind speed and power prediction of prominent wind power potential states in India using GRNN, In Power Electronics, Intelligent Control and Energy Systems (ICPEICES)," IEEE International Conference on IEEE , 1-6, 2016.
[7] M. Zeng, X. Zhang, J. Li, Q. Meng, "Multifractal analysis of short-term wind speed time series with different sampling frequencies," 12th World Congress on Intelligent Control and Automation (WCICA), pp. 3213–3218, 2016.
[8] I. Khandelwal, R. Adhikari, G. Verma, "Time Series Forecasting Using Hybrid ARIMA and ANN Models Based on DWT Decomposition," Procedia Comput. Sci., vol. 48, pp. 173–179, 2015.
[9] M. Khanna, N. K. Srinath, J. K. Mendiratta, "Feature Extraction of Time Series Data for Wind Speed Power Generation," 6th International Conference on Advanced Computing (IACC), pp. 169–173, 2016.
[10] H. G. Lee, M. Piao, Y. H. Shin, "Wind Power Pattern Forecasting Based on Projected Clustering and Classification Methods," Etri Journal, 37(2), 283-294, 2015.
[11] Y. Wanga, H. Mab, D. Wanga, G. Wangc, J. Wua, J. Biand, J. Liud, "A new method for wind speed forecasting based on copula theory," Environmental Research, 160, 365–371, 2018.
[12] R. Velo, P. López, F. Maseda, "Wind speed estimation using multilayer perceptron," Energy Conversion and Management, 81, 1–9, 2014.
[13] M. Timur, F. Aydın, T.Ç. Akıncı, "İstanbul Göztepe Bölgesinin Makine Öğrenmesi Yöntemi ile Rüzgâr Hızının Tahmin Edilmesi," Makine Teknolojileri Elektronik Dergisi 8(4), 75-80, (in Turkish) 2011.
[14] M. Zontul, F. Aydın, G. Doğan, S. Şener, O. Kaynar, "Wind Speed Forecasting Using REPTree and Bagging Methods in Kirklareli-Turkey," J. Theor. Appl. Inf. Technol., 56, 17–29, 2013.
[15] A. Meng, J. Ge, H. Yin, S. Chen, "Wind speed forecasting based on wavelet packet decomposition and artificial neural networks trained by crisscross optimization algorithm," Energy Conversion and Management,114, 75-88, 2016.
[16] L. Breiman, "Bagging predictors," Univ. California, Dept. Stat., Berkeley, Tech. Rep. 421, 1994.
[17] T. Jiang, J. Li, Y. Zheng, C. Sun, "Improved Bagging Algorithm for Pattern Recognition in UHF Signals of Partial Discharges," Energies, 4, 1087-1101; doi:10.3390/en4071087, 2011.
[18] K. Machová, F. Barčák, P. Bednár, "A Bagging Method using Decision Trees in the Role of Base Classifiers," Acta Polytechnica Hungarica, Vol. 3, No. 2, 2006.
[19] C. Doukas, I. Maglogiannis, P. Tragas, D. Liapis, G. Yovanof, "Patient fall detection using support vector machines," International Federation for Information Processing, vol. 247, pp. 147-156, 2007.
[20] T. W. Kuan, J. F. Wang, J. C. Wang, P. C. Lin, et al., "VLSI Design of an SVM Learning Core on Sequential Minimal Optimization Algorithm," IEEETransactions on Very Large Scale Integration (Vlsi) Systems, 20(4), 673-683, 2012.
[21] J. Platt, "Fast training of support vector machines using sequential minimal optimization," In B. Schölkopf, C. Burges, and A. Smola, editors, Advances in kernel methods: Support vector learning. Cambridge, MA: MIT Press. 1998.
[22] A. J. Smola, B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing, 14(3), 199-222, 2004.
[23] C. K. Madhusudana, H. Kumar, S. Narendranath, "Condition monitoring of face milling tool using K-star algorithm and histogram features of vibration signal," Engineering Science and Technology, an International Journal, 19 1543–1551, 2016.
[24] R. Satishkumar, V. Sugumaran, " Remaining Life Time Prediction of Bearings Using K-Star Algorithm – A Statistical Approach", Journal of Engineering Science and Technology Vol. 12, No. 1 168 – 181, 2017.
[25] A. W. Harley, "An Interactive Node-Link Visualization of Convolutional Neural Networks," in ISVC, pages 867-877, 2015
[26] Y. Kumar, G. Sahoo, "Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA," I.J. Information Technology and Computer Science, 7, 43-49, 2012.
[27] D.P.Alonso, S.P. Tejedor, M. Navarro, C. Rad, Á.A. González, J.F.D.Pastor, "Decision Trees for the prediction of environmental and agronomic effects of the use of Compost of Sewage Sludge (CSS)," Sustainable Production and Consumption 12 119 – 133, 2017.
[28] S. Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News," International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February, 2015.
[29] A. Behnood, V. Behnood, M. M. Gharehveran, K. E. Alyamac, "Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm," Construction and Building Materials, 142, 199–207, 2017.
[30] R. Sharma, S. Kumar, R. Maheshwari, "Comparative Analysis of Classification Techniques in Data Mining Using Different Datasets," International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 12, December, pg.125 – 134, 2015.
[31] E. K. Onyari, F. M. Ilunga, "Application of MLP Neural Network and M5P Model Tree in Predicting Streamflow: A Case Study of Luvuvhu Catchment, South Africa," International Journal of Innovation, Management and Technology, Vol. 4, No. 1, February 2013.
[32] R. Meenal, A. I. Selvakumar, "Assessment of SVM, Empirical and ANN based solar radiation prediction models with most influencing input parameters," Renewable Energy, Volume 121, June, Pages 324-343, 2018.