EJT

INESEG

Research Article

APPLICATION OF DECISION TREE METHODS FOR WIND SPEED ESTIMATION

H. Selcuk NOGAY[1*], T. Cetin AKINCI[2]

An area's wind speed forecasting is very important to investigate whether the area is available for wind power. The wind speed estimation has been carried out by means of other machine learning methods, mostly artificial neural networks. Because, in such methods, it is aimed to estimate the wind speed with the highest accuracy along with the decimal. However, if a wind farm is to be installed, the wind speed, which is a variable in the range of 0-20 m / s, can easily be estimated with round values. If the wind speed values obtained with round values are forecasted with a high accuracy rate, the wind speed that is necessary for the establishment of a wind power plant in a region is obtained by a shorter and easier way. In this study, the decision tree method was used in order to reach wind speed information with an easier method and with a very short training period. Decision tree methods were examined in three different structures and three different decision tree models were designed. Additionally the estimation results of all three methods were very high, the most accurate estimation was obtained by the "Coarse Decision Tree" method which is much simpler and faster than the others.

Key words*: Wind speed, Coarse, Fine , Medium, Decision Tree, Forecasting*

1. Introduction

Decision tree method is an effective method used to solve classification problems. As can be seen from many publications in the literature, the decision tree method has so far only been used for the classification of two-variable data [1]. In other words, usually the respond have two values, such as 1 and 0. Therefore, decision tree method is not generally preferred for the purpose of predicting any parameters in renewable energy sources. However, if the output variable in the data set is divided into classes, the decision tree method can be used to estimate and classify different parameters such as power quality in the field of renewable energy [2]. Because wind speed values in renewable energy sources can not be classified as too many varieties. However, in some studies, decision tree method has been used for wind speed classification [3]. Evaluation of wind sources and approvals to be taken for a wind farm are usually the longest activities in the development of the wind energy project. These can take up to 4 years for a large wind farm, which requires extensive environmental impact research. The installation of the wind farm can normally be completed roughly within a year. Precise determination of

[1] Kayseri University, Mustafa Çıkrıkçıoğlu Vocational Collage, Turkey, (nogay@erciyes.edu.tr)

[2] Department of Electrical Engineering, Istanbul Technical University, Turkey, (akincitc@itu.edu.tr)

Received: 29 April 2019; Accepted: 27 May 2019 Doi: https://doi.org/10.36222/ejt.558914

ISSN 2536-5010 | e-ISSN 2536-5134

the wind source at a given site is the biggest and first step in realizing a wind energy project. Because the wind source can significantly affect the cost of the wind farm. It is recommended that wind speed measurement be carried out for at least one year before the feasibility studies required to establish a wind farm in a field [4]. It is a controversial issue whether the estimation of the wind velocity with its the decimal expression is required. However, the estimation of round wind velocity values with a high accuracy rate may be sufficient to decide on the establishment of the wind farm [4]. If the wind speed can be estimated in a fast and simple way, rather than simply high accuracy, then the fastest and simplest method can be preferred over other machine learning methods. In this study, wind speed estimation was performed by decision tree methods. Estimation was made by three decision tree approach. The estimation results of these three models were compared with each other and the fastest and most attractive method with the highest accuracy rate was tried to be found.

After the introduction in the first part of the study, in the second part, the theoretical framework of the study is briefly explained. In this section, the structure and approach methods of decision trees are given.

In the third part of the study, the preferred method for performing the study was introduced. The stages of the study are explained.

In the results section, numerical results and graphs obtained from the study are given.

In the last section, evaluation of the results is given. It is discussed which method is more advantageous than others and how the results can be evaluated from the obtained graphs.

2. Theoretical Framework

Decision Trees (DT) are a classification and pattern recognation algorithm that has been widely used in literature in recent years. The most important reason for the widespread use of this method is that the rules used to create DT structures are understandable and simple. A multi-step or sequential approach is used in performing the DT classification process. Although other methodologies such as neural networks can be used for classification, decision trees provide an advantage for decision makers in terms of easy interpretation and intelligibility [5]. DT have some advantages compared to other classification methods. Some of these include low cost, easy-to-understand convenience, and compatibility with databases. Because of these advantages, it is widely used in many classification problems. Classification in decision trees is carried out in two steps: learning and classifying. In the learning step, the model is designed with an edited training data. The trained DT model is defined as classification rules or decision tree. In the classification step, test data is used to determine the accuracy of the classification rules or DT. If high accuracy is obtained in the test result, the rules are used to classify new data. It should be determined which fields in the training data should be used to form the tree in which order [6, 7].

The basic structure of a DT consists of three basic parts called nodes, branches and leaves. In this tree structure, each attribute is represented by a node. Branches and leaves are other elements of the DT structure. The last part of the DT is called a leaf and the top part is called a root. Portions between roots and leaves are expressed as branches [7, 8]. In other words, a tree structure consists of a root node containing data, internal nodes (branches), and end nodes (leaves). By using the attribute information of the training data, the basic principle in establishing a DT structure can be expressed as a series of questions about the data and concluding in the shortest time by acting on the answers obtained. In this way, the DT collects the answers to the questions and creates decision rules. The root node, which is the

first node of the tree, begins to be asked questions for the classification of the data and the formation of the tree structure. This process continues until leaves or nodes without branches are found. The test data is used to determine the generalization capability of the generated tree for a new data set. A test data that is new to the tree structure created by the training data enters the root of the tree. This new data tested in the root is sent to a lower node according to the test result. This process is continued until it reaches a specific leaf of the tree. There is only one way or a single decision rule from root to every leaf. Figure 1 shows a simple (coarse type) tree structure consisting of four-dimensional attribute values of four classes. The tree structure in Figure 1 is the DT of the coarse type algorithm designed in this study. In Figure 1, the "VarNameA" attribute values; VarNameA <a, b, c, d and e values are the threshold values for branching. 1, 2, 3 and 4 show the class labels. Depending on the number of variables used in each stage of tree formation, there are single variable or multivariate decision tree structures [9-11].

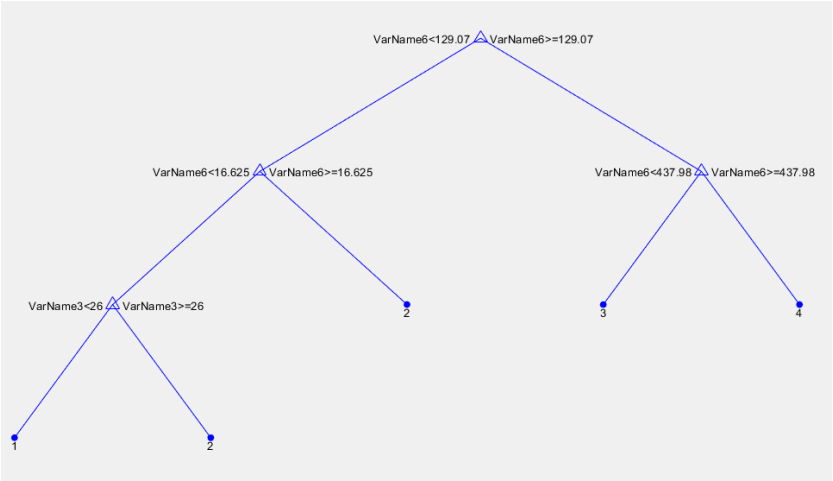

Figure 1. Coarse DT

By testing a single attribute in each internal node of the univariate decision tree, the data are divided into two or more subgroups. In this way, division continues until the decision tree repeatedly reaches a leaf node of the input data. The specific values of the decision limits in an univariate decision tree are experimentally estimated from the training data. In the case of persistent data, the $i_x > c$ shaped logical test is carried out, where x_i shows a characteristic in the data space of each internal node at the internal node, and that '*c'* is a threshold value in the observed range of x_i. The threshold value '*c*' can be determined using certain conditions, such as maximizing differences or minimizing similarity in descent nodes [12-14]. The most important step in the creation of decision trees is to determine the criteria for tree branching. There are various approaches developed to solve this problem. The most important of these are the 'information gain' and 'information gain ratio', the 'gini-index', the 'towing rule', and the 'square probability table statistic' approaches. ID3, C4.5, C5.0, CART, CHAID and QUEST algorithms are examples of algorithms using these approaches. According to the method of information gain, 'information theory' which includes 'entropy rules' is used in order to determine which according to characteristic to make branches in decision tree. Entropy is a measure of irregularity or uncertainty in a system. In single variable decision trees, ID3 algorithm uses informagion gain approach. The C4.5 algorithm, which is an improved version of this algorithm, uses the gain ratio approach calculated by taking advantage of the information gain. Gini Index approach was used in this study. All

variables are assumed to be continuous according to the Gini Index. It is assumed that there are many possible distinctions for each variable [14].

If a '*T*' data set contains '*N*' samples from '*n*' different classes, the gini index, gini (*T*) is calculated as follows:

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2 \quad (1)$$

p_j denotes the relative frequency of the class j in T.

If the T data set is divided into two N_1 and N_2 magnitudes as T_1 and T_2 respectively, the gini index for the allocated data is:

$$gini_d(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2) \quad (2)$$

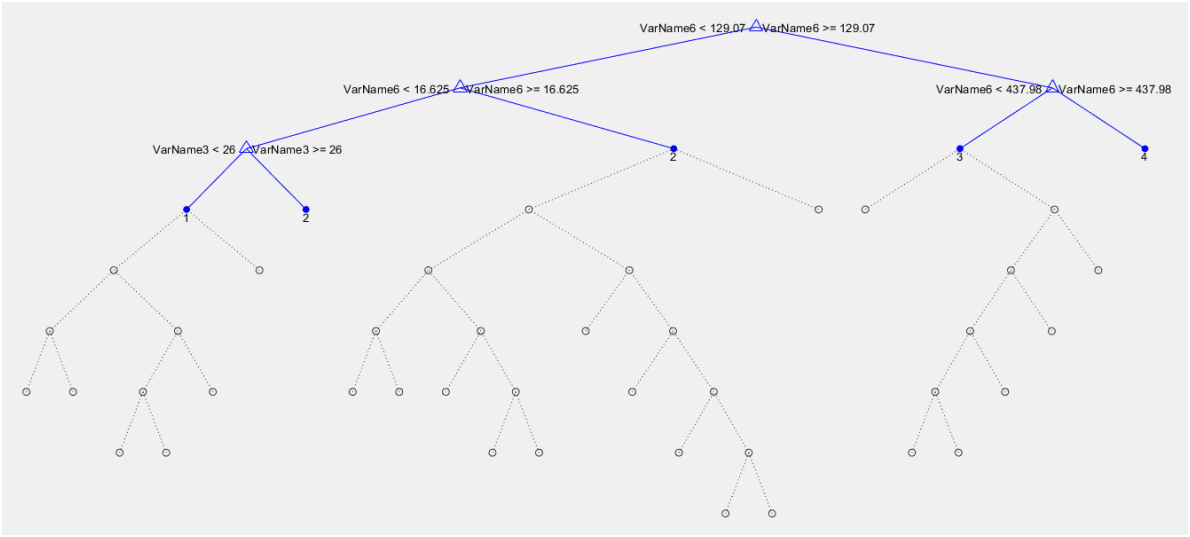

Figure 2. Fine decision tree model

The variable with the lowest gini value is selected. Another important point in the DT formation is the pruning of the tree structure. The DT classifier divides the training data into subsets that contain only one class, resulting in a very large and complex tree structure. Therefore, it may be possible to place a leaf in the decision tree instead of a sub-tree. In this way, the processing decision tree is called pruning [15-19]. With pruning, the parts of the decision tree which do not affect the accuracy of the classification are removed. Thus, a less complex and more understandable tree is obtained. Two methods are generally used to simplify the tree structure and reduce the process complexity by pruning. The first one is the pruning method which is decided to be reduced while the tree structure is formed, and the other is the last pruning method where the pruning is made after the tree structure is formed. If the pruning of the sub-tree with a single leaf or the most used branch of this tree will reduce the expected error rate, the tree is truncated. As the error rate in the sub-branches is reduced, the error rate for the whole tree will decrease. At the end of the pruning process, a tree with a minimum error rate is obtained. 2 shows a sample that can be prone to pruning in the *fine* type decision tree used in this study [16 – 19].

3. Methodology

The study was performed using Classification Toolbox in MATLAB environment. In this environment, decision trees are given as three separate structures. The first building is the most complex 'Fine' decision trees among the three. The second is 'Medium' and the third is the 'Coarse' decision tree structure. In this study, these three each of DT structures were designed to estimate wind speed. The Gini index was selected for all three decision trees. The data set was obtained and arranged before the decision tree models were created. Table 1 shows a summary of the data set used for each of the three models. Table 1 also shows the definitions of variables. The first 6 variables are "Predictor" and the last variable is "Respond". "Respond" in the data set shows the wind speed for 50 meter high. The data set, used for this study was obtained from "Turkish State Meteorological Service in 2014". The data set consists of meteorological data for Turkey Amasra district.

Table 1. Summary of the data set

Data	**Predictors**						**Respond**
Variable Label	VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7
Variable Definition	Daily average temperature (C°)	Daily average temperature (K°)	Daily average soil temperature 20 cm (C°)	Daily local relative humidity (%)	Ro, Air density (kg/m^3)	P/A, Wind power intensity (W/m^2)	Daily average wind speed, (m/s)
Min	-1.11	272.04	4.45	37	1.16	0.19	0.49
Max	29.40	302.55	27.6	96	1.29	10308.76	18.49

Since the DT method is a classification method, it is necessary to enumerate the wind speed data by classifying them in order to use this method for estimation purposes. Table 2 shows the numbering of the actual respond of the data set. In other words, the wind speed data representing the output of the model is divided into four parts and a new output is obtained.

Table 2 Numbering of respond

Win Speed 50 m (m/s)	Numbering
0_3	1
3_6	2
6_9	3
9 =>	4

In the study, three separate decision tree models were designed. The first model is the Coarse decision tree model with a very simple structure. Figure 1 shows the structure of this model. The second model is the Fine decision tree model. The Fine decision tree model shown in Fig.2 is 7-dimensional. Figure 3 shows the branches on the left side of the Fine decision tree created. At each node, the variables are given with their threshold values, whether they are large or small. The numbers in the last sections of each branch represent the wind speed.

Figure 4 shows the branches, nodes, and leaves in the middle of the Fine decision tree. In Figure 4, the thresholds at each node are compared with the corresponding variables. Figure 5 shows the right-hand branch and threshold values of the Fine DT. The third DT model designed in the study is less complicated than the Fine DT. Figure 6 shows the third model that is the Medium decision tree model.

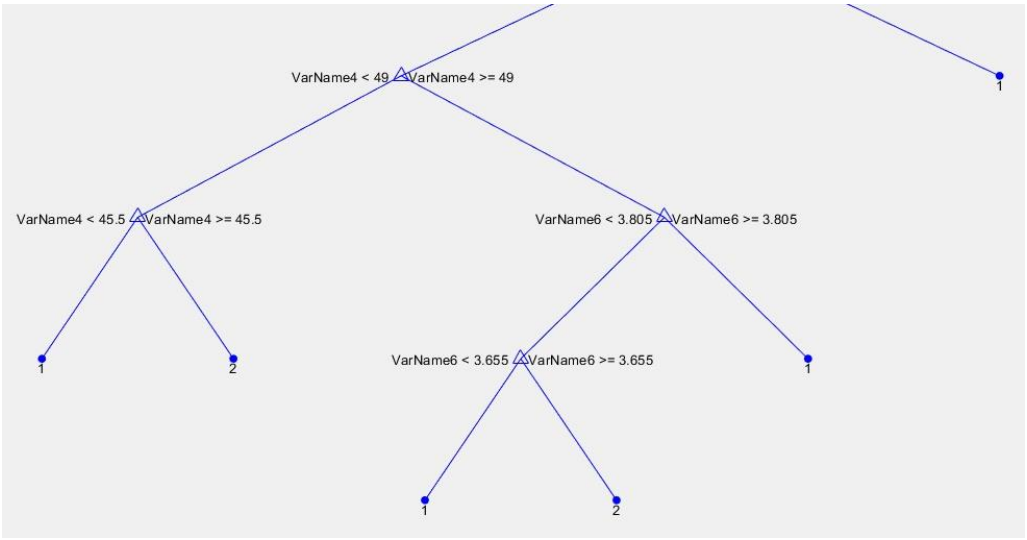

Figure 3. Left side of fine decision tree

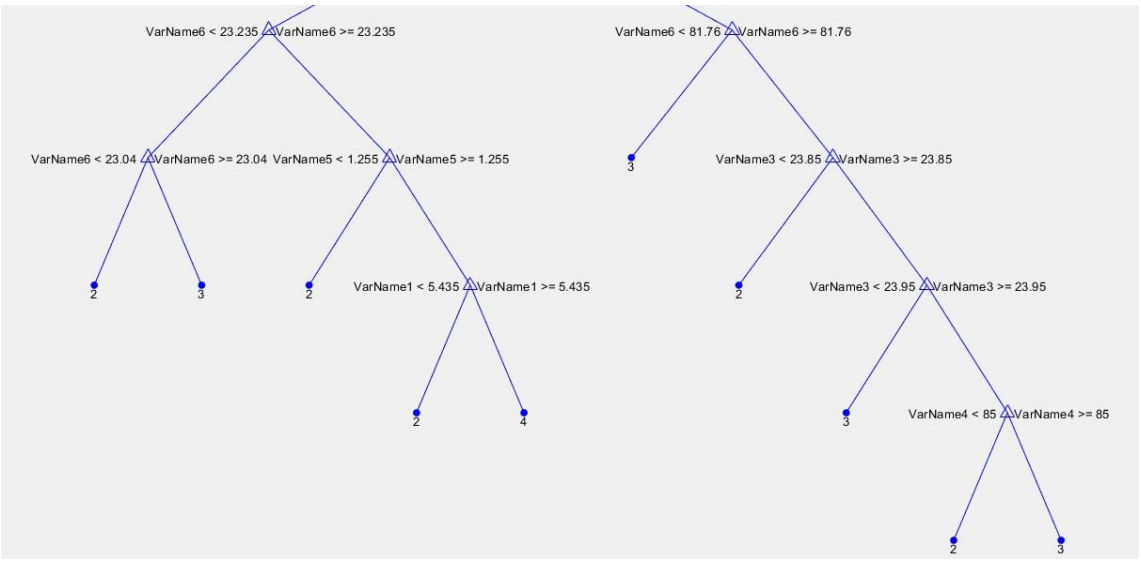

Figure 4. Middle part of decision tree

As can be seen from Figure 6, although Medium DT consists of fewer branches than Fine DT, it has the same numer of exit points and answers. Also in Figure 7, middle part of Medium DT is illustrated.

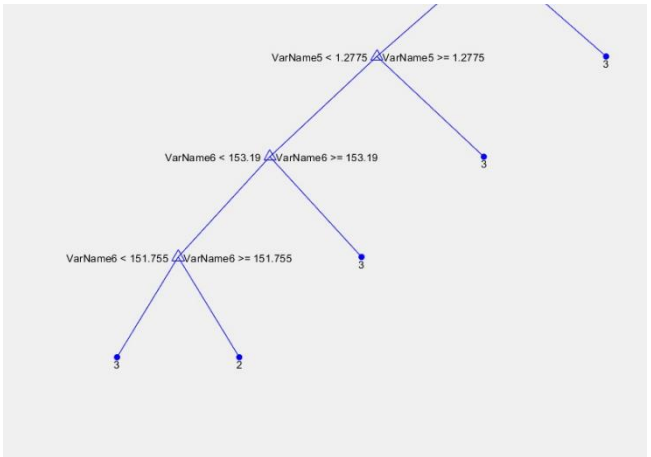

Figure 5. Right side of fine decision tree

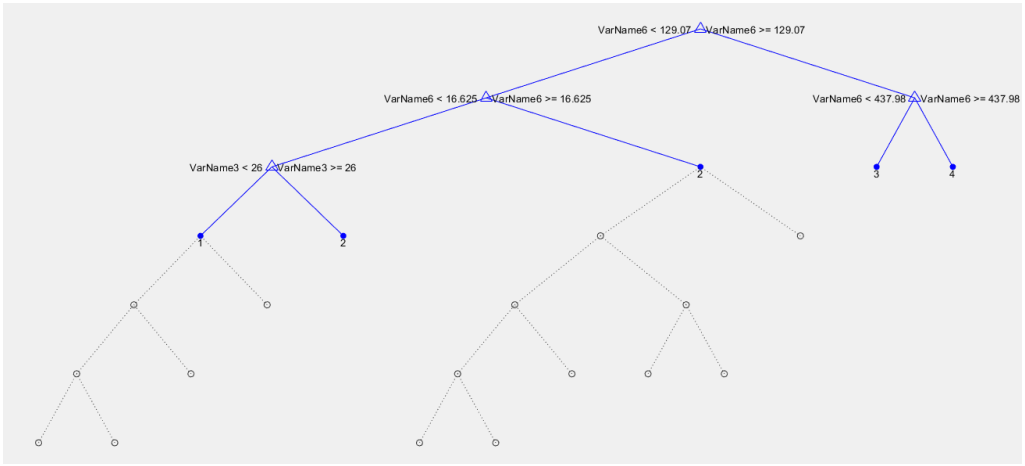

Fig.6 Medium decision tree

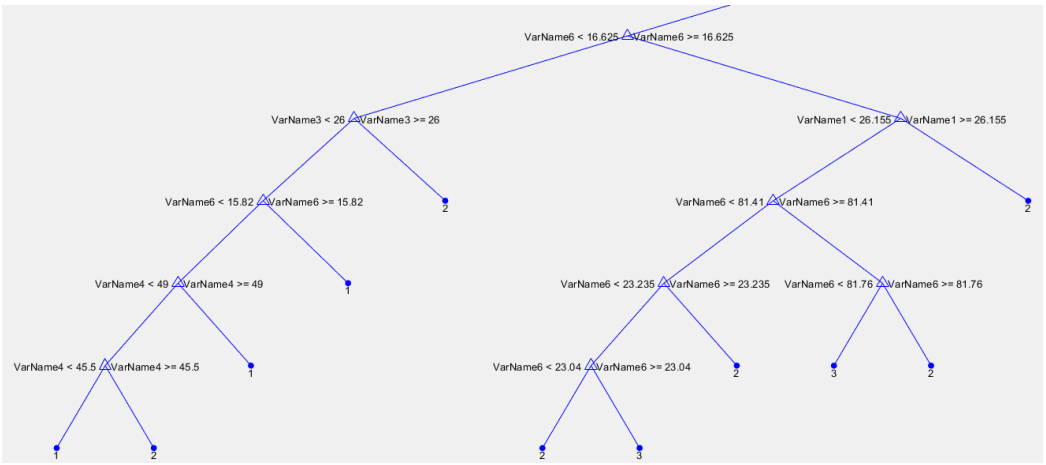

Figure 7. Middle part of medium decision tree

4. Results

Table 3 shows the results obtained from the study. Estimates were obtained at high accuracy rates in all three decision tree methods.

Table 3 The results of applications

	Accuracy (%)	Prediction speed (obs/sec)	Training time (sec)	Maximum number of split
Fine DT	95.6	1400	0.93103	100
Medium DT	95.9	8200	2.0152	20
Coarse DT	96.7	39000	0.52248	4

Figure 8 shows how two variables affect output. The Scatter plot is only available between VarName6 and VarName1. Figure 9 shows the ROC curves obtained for all three decision tree models

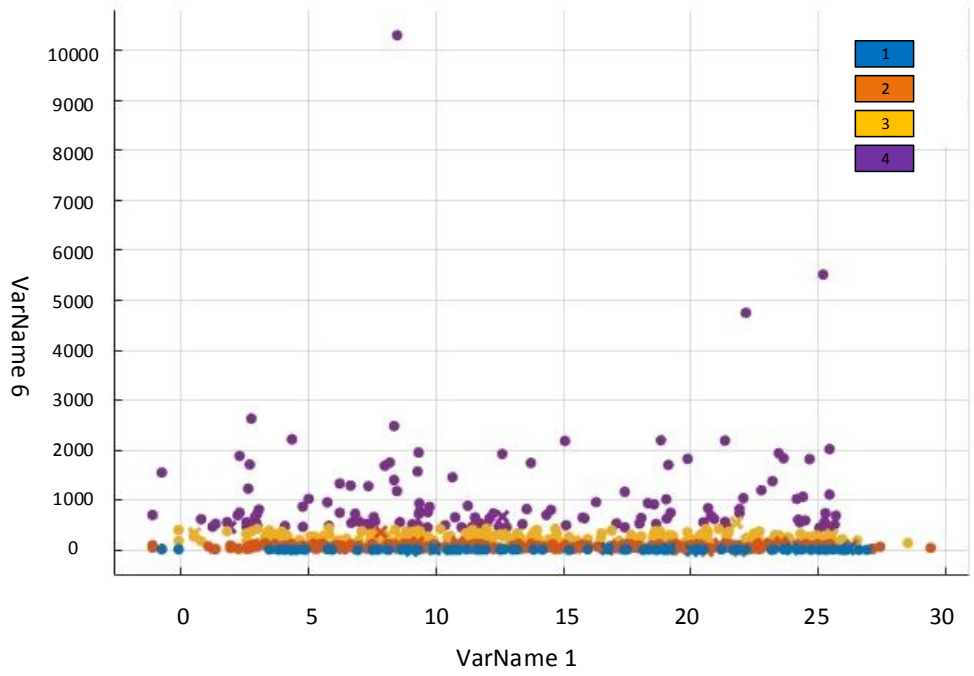

Figure 8. Scatter plot for fine tree

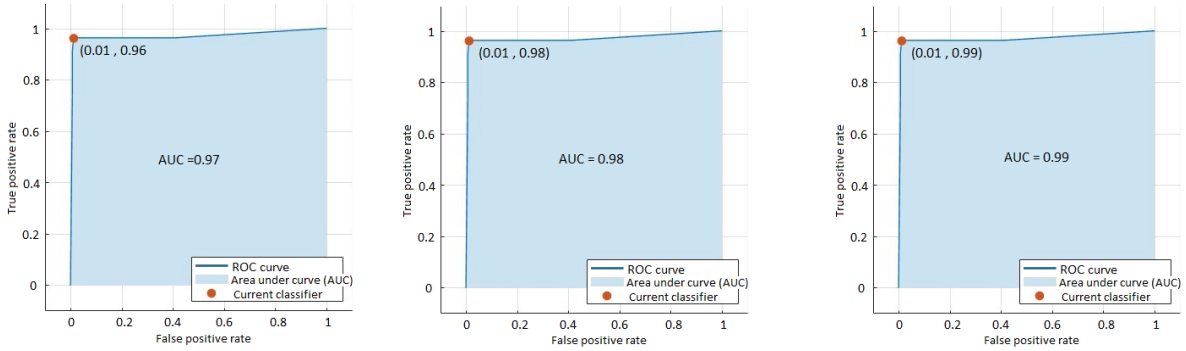

Figure 9. ROC curves of a) Fine DT, b) Medium DT, c) Coarse DT

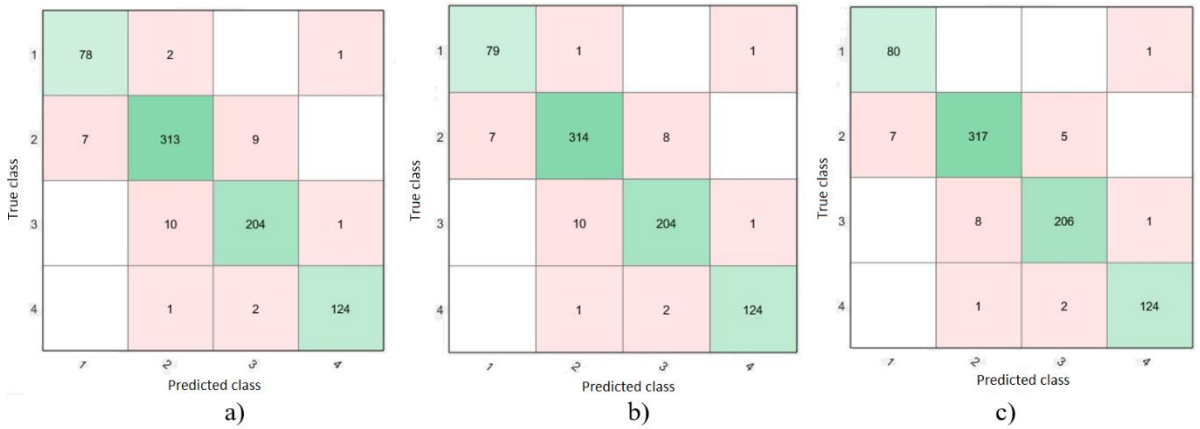

Figure 10. Confusion matrixes of a) Fine DT, b) Medium DT, c) Coarse DT

Figure 10 shows the confusion matrices obtained according to the estimation results of all three models. With this confusion matrix, it is also possible to understand the reason for each prediction and which output is correct or incorrect for which output.

5. Conclusions

The following results can be obtained from the study.

1. From Table 3, it can be observed that the Coarse Tree has reached the highest accuracy rate. Coarse tree is the simplest decision tree method. Therefore, as can be seen from Table 3, it has a higher estimation rate than other methods. At the same time, the duration of training is shorter than the duration of the other methods. In the Coarse decision tree method, the maximum number of splits is considerably less than the others, depending on the fact that the tree is not complex.
2. According to the ROC curves in Figure 9, the AUC (Area Under the Curve) values of medium and coarse decision trees are equal to each other and are higher than the Fine decision trees. As can be seen from the ROC curves, all three decision trees are successful in the forecastings. In addition, the simplest and most understandable Coarse DT method can be said to be more successful and acceptable.
3. According to the confusion matrices in Figure 10, if the diagonal figures of the 4x4 matrix (green ones) are examined and the values of the confusion matrix belonging to each decision tree method are compared to each other, the superiority of the coars can be noticed. Diagonal numbers indicate the number of accurately predicted data. It is obvious that the maximum number is in Coars. Figures in other regions indicate incorrect estimates.
4. As a result, the decision trees method produces satisfactory estimates by replacing the artificial neural networks when the correct numbering and correct calculation option and the correct decision tree method are used. These estimates, however, are predetermined rounded output values, as indicated above. It is understood from this study that it is possible to find wind speed estimations needed for any region with decision tree close to reality.

References

[1] Gupta, B.,, Rawat, A., Jain, J., Arora, A., Dhami, N. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications*, 163(8), 15-19.

[2] Ray, P.K., Kishor, N. (2014). Optimal Feature and Decision Tree-Based Classification of Power Quality Disturbances in Distributed Generation Systems. *IEEE Transactions on Sustainable on Energy*, 5(1), 200-208.

[3] Sangita B.P., Deshmukh, S.R. (2011). Use of Support Vector Machine, decision tree and Naive Bayesian techniques for wind speed classification. *International Conference on Power and Energy Systems Power and Energy Systems (ICPS)*.

[4] Retscreen Engineering & Cases Textbook, Clean Energy Project Analysis, Clean Energy Decision Support Centre ISBN: 0-662-35670-5 Catalogue no.: M39-97/2003E-PDF, © Minister of Natural

Resources Canada 2001 - 2004. http://unfccc.int/resource/cd_roms/na1/mitigation/Module_5/Module_5_1/b_tools/RETScreen/Manuals/Wind.pdf, last acces date: Feb 27th, 2019

[5] Chien, C. F., Chen, L. F. 2008. Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-Technology Industry. *Expert Systems with Applications*, 34, 280-290.

[6] Tsang, S., Kao, B., Yip, K.Y., Ho, W.S., Lee, S.D. (2011). Decision Trees for Uncertain Data, *IEEE Transaction on Knowledge and Data Engineering*, 23(1), 67-78.

[7] Rokach, L., Maimon, O. (2014). *Data Mining with Decision Trees Theory and Applications*. 2nd edition, 81, World Scientific Publishing Co. Pte. Ltd.

[8] Quinlan J.R, (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 302.

[9] Dougherty, G., *Pattern Recognition and Classification*. Springer New York Heidelberg Dordrecht London.

[10] Loh, W.Y., Shih, Y.S. (1997). Split Selection Methods for Classification Trees, Statistica Sinica, 7(4), 815-840.

[11] Friedl, M.A., Brodley C.E., (1997). *Decision tree classification of land cover from remotely sensed data*. Remote Sensing of Environment, 61, 399–409

[12] Safavian S.R., Landgrebe D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems Man and Cybernetics*, 21, 660-674

[13] Tan, P.N., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining*. Pearson Addison-Wesley.

[14] Maimon, O., Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*, Springer 2010.

[15] Suneetha, N., Hari, Ch.V.M.V., Kumar, S. (2010). *Modified Gini Index Classification: A Case Study of Heart Disease Dataset*. International Journal on Computer Science and Engineering, 2(6),1959-1965

[16] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth, 358 s.

[17] Raileanu L.E., Stoffel, K., *Theoretical Comparison between the Gini Index and Information Gain Criteria*. Annals of Mathematics and Artificial Intelligence, 41(1):77-93.

[18] Teknomo, K. (2012). *Decision Tree Tutorial*. www. revoledu.com Online edition.

[19] Mingers J., (1989). *An empirical comparison of pruning methods for decision tree induction*. Machine Learning, 4, 227–243.