

## Article

# A Unique Unified Wind Speed Approach to Decision-Making for Dispersed Locations

Ayman M. Mansour<sup>1</sup>, Abdulaziz Almutairi<sup>2</sup>, Saeed Alyami<sup>2,\*</sup> , Mohammad A. Obeidat<sup>3</sup> ,  
Dhafer Almkahles<sup>4</sup>  and Jagabar Sathik<sup>4</sup>

<sup>1</sup> Department of Communication, Electronics and Computer Engineering, Tafila Technical University, Tafila 66110, Jordan; Mansour@ttu.edu.jo

<sup>2</sup> Department of Electrical Engineering, College of Engineering, Majmaah University, Al-Majmaah 11952, Saudi Arabia; adalmutiri@mu.edu.sa

<sup>3</sup> Department of Electrical Power and Mechatronics Engineering, Tafila Technical University, Tafilah 66110, Jordan; maobaidat76@ttu.edu.jo

<sup>4</sup> Renewable Energy Lab, Prince Sultan University, Riyadh 12435, Saudi Arabia; dalmakhles@psu.edu.sa (D.A.); mjsathik@ieee.org (J.S.)

\* Correspondence: s.alayami@mu.edu.sa

**Abstract:** The repercussions of high levels of environmental pollution coupled with the low reserves and increased costs of traditional energy sources have led to the widespread adaptation of wind energy worldwide. However, the expanded use of wind energy is accompanied by major challenges for electric grid operators due to the difficulty of controlling and forecasting the production of wind energy. The development of methods for addressing these problems has therefore attracted the interest of numerous researchers. This paper presents an innovative method for assessing wind speed in different and widely spaced locations. The new method uses wind speed data from multiple sites as a single package that preserves the characteristics of the correlations among those sites. Powerful Waikato Environment for Knowledge Analysis (Weka) machine learning software has been employed for supporting data preprocessing, clustering, classification, visualization, and feature selection and for using a standard algorithm to construct decision trees according to a training set. The resultant arrangement of the sites according to likely wind energy productivity facilitates enhanced decisions related to the potential for the effective operation of wind energy farms at the sites. The proposed method is anticipated to provide network operators with an understanding of the possible productivity of each site, thus facilitating their optimal management of network operations. The results are also expected to benefit investors interested in establishing profitable projects at those locations.

**Keywords:** data mining; decision tree; wind speed; renewable energy; system modeling; machine learning



**Citation:** Mansour, A.M.; Almutairi, A.; Alyami, S.; Obeidat, M.A.; Almkahles, D.; Sathik, J. A Unique Unified Wind Speed Approach to Decision-Making for Dispersed Locations. *Sustainability* **2021**, *13*, 9340. <https://doi.org/10.3390/su13169340>

**Academic Editors:**  
Avelino Núñez-Delgado and  
Pablo García Triviño

Received: 6 June 2021

Accepted: 10 August 2021

Published: 20 August 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Growing global interest in reducing the environmental pollution created by heavy reliance on oil derivatives for the production of electric power has motivated governments to take significant steps toward the implementation of renewable energy. One of the most important renewable energy sources is wind, with the 2019 total world capacity of wind energy estimated to be 650 gigawatts [1] and the annual global increase in wind energy calculated at 20% [2]. This expansion has resulted in wind energy technology becoming a principal source of energy in terms of sales and technical development. In spite of these advances, this energy resource remains unreliable at high rates, and increasing dependence on this technology is associated with the emergence of numerous problems for electrical system operators. Examples of these challenges are the substantial changes in wind production arising from the random behavior of wind speeds, as well as the

difficulty of accurately forecasting wind production, which gives rise to many issues during the operation of the electric grid. Any decision to increase the use of wind energy hence requires careful planning along with highly reliable methods of making rational and informed decisions [3,4].

To analyze and evaluate the effects of inconsistent wind behavior on the reliability and stability of an electrical grid as well as on short-term operation and long-term planning, several researchers have applied and reported probability-based methods. For example, in [5,6] a sequential Monte Carlo simulation (SMCS) method was used for representing the probability distribution and time-series characteristics of wind speed. Another efficient method is a Monte Carlo Markov chain (MCMC) method [7,8], which is based on the dependence of the wind speed at a given point in time on the speed during the previous moment. This feature makes this method effective for preserving the chronological characteristics of wind speed. Some studies [9–11] have also dealt with correlations between the output levels of wind turbines installed in separate geographical areas or between those of multiple wind farms in adjacent areas. These studies led to the conclusion that a determination of the type of correlation (positive, negative, or zero) is related to several factors, including the way the turbines are arranged on the site and the method employed for connecting the turbines with one another as well as with the electrical network.

An examination of the correlation between the output levels of distant wind energy sites is not usually of interest to researchers because the relationship is often a zero or an inverse correlation. However, we believe that reconsidering this factor is very important, especially with respect to the correlations among multiple wind energy sites in different regions of the same country or in different countries, which might be interconnected in an electrical network. Conducting such studies would offer several advantages: (1) Knowing the diversity of and variations in wind energy production from different sites would be beneficial for grid operation in terms of power quantity and time of supply. (2) Prior knowledge of the amount of variation and the type of correlation, even if negative, would help network operators achieve effective management of grid operations, such as load flow and network stability. (3) Identification of the potential of wind energy in each region of a country is crucial information for investors or decision makers.

The results of such a study would be very important for countries that feature large areas and substantial regional diversity. With an area of 2.25 million square meters encompassing regions that exhibit varied environmental characteristics, the Kingdom of Saudi Arabia (KSA) is one such country. The KSA is also one of the largest countries in the Middle East, and most of the nearby countries rely on the KSA for resilient grid interconnections for ensuring power security and economic benefits. The Saudi government is taking rapid steps toward diversification of energy sources and is investing heavily in sustainable energy. This trend is one of the main priorities and objectives of the KSA's Vision 2030. One of the most important of these subsidized projects is wind energy, since it is expected that wind energy capacity will reach 9.5 gigawatts by 2030 [12]. In 2018, the Renewable Energy Development Office (REDO) nominated about 50 companies to begin implementing the planned renewable projects [12], which include solar power stations with a capacity of 300 megawatts and wind farms with a capacity of 400 megawatts. These stations are to be operating and connected to the electric grid before the end of 2021. The Saudi government recently announced new renewable energy projects estimated at \$50 billion, with implementation expected to be completed in 2023. Establishing such projects requires accurate technical and economic studies so that suitable construction locations can be determined.

In the past few years, numerous studies dealing with wind energy in Saudi Arabia have appeared in the literature. As reported in [13–15], several studies involved the analysis of statistical parameters associated with different wind farm sites and the extraction of Weibull distribution parameters for each individual site. The limitation of these studies is that their findings with respect to site productivity were dependent on the overall assessment of the available wind speed data for each site. Based on this method, the evaluation might indicate that a site is currently unsuitable for a wind project, but that site

might in fact be considered a good choice for a specific period. These studies also relied on the assumption that an appropriate distribution for all sites is a Weibull distribution. Since such an assumption is neither accurate nor valid for all sites, the results could be over-approximations, according to [16–18]. To the best of our knowledge, no study has taken into account either wind speed data collected for different, distanced locations or the processing of those data as a single package to maintain the characteristics of the correlations among locations and thus to provide more accurate and detailed standard measures of wind speed productivity at those locations.

Addressing this point represents the core contribution of the work presented in this paper. Data mining techniques have recently been used in numerous applications because of the benefits these techniques offer with respect to developing models and making decisions.

Several studies have employed artificial intelligence techniques for renewable systems. For example, artificial neural networks are used in [19] to characterize PV modules. Application of data mining procedures that include support vector machines and fuzzy logic is also applied in several studies. In [20], a new methodology combining both Gaussian-kernel support vector machine and adaptive fuzzy inference system is developed. This methodology extracts the fuzzy rules directly from the training data to be used in the testing stage. In [21,22], EEG signals are analyzed using SVM, ANN, Naïve Bayes, and decision trees for epilepsy detection. In [23], authors have used the decision tree technique to detect adverse drug reactions and the system was optimized using a genetic algorithm. An efficient feature selection method was developed in [24] for enhancing Arabic text classification. In [25–27], texture classification techniques are developed based on independent component analysis and naïve base classifier.

In this study, a decision tree algorithm is used and the major contributions of this study in comparison to existing studies are as follows:

1. A unique and unified method for predicting wind speeds at diversified locations in the KSA is proposed. The proposed model enables the examination of deviations and correlations of wind speeds at different locations.
2. A model is developed that deals and examines an extensive range of data for a variety of sites. In addition, conclusions about the characteristics of these data using the least possible number of classifications can also be drawn to facilitate the understanding of the data and to expedite their use. The goal was to help decision makers arrive at quick, accurate, and informed decisions.
3. Finally, the capability of the assessed locations can be ranked to enable system operators to ascertain in advance the monthly productivity of each site so that they can implement appropriate planning and operating actions.

## 2. System Design and Methodology

This section provides details about the developed prediction system, which is based on a decision tree algorithm. Numerous decision tree algorithms are currently available, including random forests, random trees, the J48, and classification and regression trees (CART). A decision tree algorithm employs training data to build a tree model that is used for classification purposes. The developed classification algorithm involves three phases: data gathering, data preprocessing, and learning and classification. In the data-gathering phase, the training and test set is collected from wind station databases. The second phase involves the preprocessing of the data, including outlier detection and elimination, missing data treatment, and averaging. In the learning and classification phase, the goal is to develop an intelligent decision mechanism. A test set is then applied for determining the accuracy of the developed model.

### 2.1. Data-Gathering Phase

The five locations whose wind speed data were examined in this study were carefully selected to include all regions of the KSA [28]. Five sites were chosen to be representative

of each region: center, east, west, south, and north. The selection corresponds to the operational divisions of the Saudi Arabian electrical system. Figure 1 shows the sites where the data were collected.



**Figure 1.** Regional map of Saudi Arabia.

Table 1 provides a statistical summary of the data collected for each site. These statistics are a collection of indices that provide meaningful information regarding the location and variability of the data. To facilitate their interpretation, brief definitions of some of the statistics are given here [29]. The most common indicator of the central tendency of a random variable is the mean, which represents the average number of data points. For the selected sites, it can be noted that the means are about 3 m/s to 4 m/s, with the exception of the east region, where 1.9 m/s is the recorded mean. The standard error (SE) is the measure that indicates how close the mean of the sampled data is to the true population mean. An SE of 0.05 or less implies that the sample data are quite similar to those for the whole population, with a confidence level of 95%. As can be observed from a review of the results, the SE values for all sites are less than 5%, so the data sample for each site is thus large enough to represent the true population. The median is another measure of central tendency, and the mode refers to the most frequently or commonly occurring number in the data. Standard deviation and variance denote the spread of the data distribution. Kurtosis identifies whether the tails of a given distribution contain extreme values. Skewness is the measure of the symmetry of distribution, and it differentiates extreme values in one versus the other tail. The minimum is the smallest value in the data set while the maximum is the largest value in the data set. The sum shows the summation of the wind speeds of all data sets. The count shows how many items the data have. The results listed in Table 1 reveal noticeable differences among the statistical values associated with different sites. These discrepancies were expected due to the divergent distances between the sites and the diverse nature of the local weather.

**Table 1.** Data set statistics.

Statistics	Center	East	West	North	South
Mean	3.935	1.923	3.623	3.270	3.001
Standard Error	0.008	0.006	0.011	0.013	0.014
Median	3.800	1.800	3.300	3.000	2.600
Mode	3.300	1.700	2.900	2.700	0.000
Standard Deviation	1.572	0.999	1.939	1.688	1.891
Sample Variance	2.470	0.998	3.758	2.849	3.577
Kurtosis	0.548	0.711	0.563	1.257	−0.338
Skewness	0.541	0.713	0.813	0.901	0.516
Minimum	0.000	0.000	0.000	0.000	0.000
Maximum	12.20	7.600	13.70	15.20	11.10
Sum	97,878	39,455	77,928	57,344	57,031
Count	24,871	20,523	21,511	17,539	19,007

The data is a part of the Renewable Resource Monitoring and Mapping (RRMM) program prepared by King Abdullah City for Atomic and Renewable Energy (KACARE). KACARE monitored and recorded the wind speed data at different installed stations in the Kingdom of Saudi Arabia at 3 m height. Table 2 provides an example of data for one of the five sites. The size of the sample is associated with the amount of information provided and the determination of the precision or level of confidence about the desired estimate. Wind speed estimate always has an associated level of uncertainty, which depends upon the underlying variability of the data as well as the sample size: the smaller the sample size, the greater the uncertainty in the estimate. Similarly, a larger sample size can provide more information, thus the uncertainty is reduced. In this study, the sample size in all selected sites ranges from 19,000 to 25,000 data points. We tried to collect this large sample size to reduce the amount of uncertainty associated with the estimate and achieve reasonable results. The steps involved in the proposed model through the Weka tool consider different concepts of data mining, which are as follows. First, the Weka software allows preprocessing step for raw data to detect the outliers and irrelevant data by cleaning and clustering the data using the k- means technique. In addition, the data mining techniques cater to the uncertainty. This is noticed in the used decision tree methodology when applying the Gini impurity measure to decide the optimal split from a root node and subsequent splits. The Gini impurity measures the frequency at which any element of the dataset will be mislabeled when it is randomly labeled. The entropy is another way of measuring that is based on the selection of the optimum split for the features with less entropy.

**Table 2.** Sample from the south site database.

Site	Latitude	Longitude	Date	Wind Speed (m/s)	Irradiance (Wh/m <sup>2</sup> )
Jazan University	16.96035	42.545865	14/01/2015 07:00:00	1.5	2.1
Jazan University	16.96035	42.545865	14/01/2015 08:00:00	2.2	83.4
Jazan University	16.96035	42.545865	14/01/2015 09:00:00	3.1	249.5
Jazan University	16.96035	42.545865	14/01/2015 10:00:00	3.9	452.7
Jazan University	16.96035	42.545865	14/01/2015 11:00:00	4	366.6
Jazan University	16.96035	42.545865	14/01/2015 12:00:00	4.9	519

A subset of the combined database is shown in Table 3. The data were collected from 9 January 2013, to 31 December 2016. The subset consists of 34,872 records. The information in Table 3 is only a small subset of the available database. Zero irradiances for the north region in this table were recorded at 4 and 5 am; this is normal at sunset time when the sun disappears.

**Table 3.** A subset from the combined database.

Region	Latitude	Longitude	Date	Wind Speed at 3 m (m/s)	Irradiance (Wh/m <sup>2</sup> )
West	21.49604	39.24492	29/05/2013 10:00:00	2.6	674.3
West	21.49604	39.24492	29/05/2013 11:00:00	4	840.5
North	27.39	41.42	1/1/2015 4:00	2.9	0
North	27.39	41.42	1/1/2015 5:00	2.5	0
East	25.34616	49.5956	29/05/2013 08:00:00	3	471.6
East	25.34616	49.5956	29/05/2013 09:00:00	3.5	671.7
Center	24.52958	46.43635	9/1/2013 11:00	4.9	611.3
Center	24.52958	46.43635	9/1/2013 12:00	3.8	697.9
South	16.96035	42.545865	5/11/2014 8:00	0.6	206.9
South	16.96035	42.545865	5/11/2014 9:00	1.5	411.5

## 2.2. Data Preprocessing Phase

Data preprocessing includes data cleaning and missing data treatment. In this phase, information not needed for the wind speed model, such as the irradiance and the latitude and longitude, are removed from the database. Wind speed data missing for a specific date are then replaced by the average value of the wind speeds for that day [30–33]. That date is eliminated and simply replaced by the corresponding month; i.e., 29/05/2013 10:00:00 is replaced by May, as shown in Table 4.

**Table 4.** Database following preprocessing.

Region.	Date	Wind Speed at 3 m (m/s)
West	May	2.6
West	May	4
West	May	2.5
North	January	2.9
North	January	2.5
North	January	2.8
East	May	3
East	May	3.4
East	May	3.7
Center	January	4.9
Center	January	3.8
Center	January	3.7
South	November	0.6
South	November	1.5
South	November	2.6

The combined database is then rearranged to add an output label to a new set of input attributes. The new set of input attributes are defined as indicated in Table 5: month, center wind speed, south wind speed, east wind speed, north wind speed, and west wind speed. The output attribute consists of multi-labeled data: case 1 to case 120. Since the number of locations is five, the resultant possible number of output cases is  $5! = 120$  possibilities.

**Table 5.** Attribute list sample for developing the decision tree model.

Month	Center	South	East	North	West	Output
	Wind Speed (m/s)					
Jan	4.00	3.36	1.79	3.30	2.77	case 1
Feb	3.96	3.04	0.91	3.83	2.98	case 2
Mar	4.64	3.21	2.25	3.67	3.51	case 2
Apr	4.17	3.18	2.09	4.08	3.45	case 3
May	3.77	3.15	2.03	3.62	3.97	case 4
Jun	3.95	3.44	2.36	3.04	4.84	case 5
Jul	3.64	3.28	1.99	2.61	3.55	case 6
Aug	3.70	3.38	2.11	2.96	4.08	case 5
Sep	3.73	2.84	1.75	2.87	4.60	case 4
Oct	3.95	2.17	1.76	3.17	3.64	case 7
Nov	3.89	2.81	1.79	3.85	3.01	case 3
Dec	3.48	2.77	1.66	3.34	2.89	case 3

With the use of an association rule algorithm [34–37], the number of possible cases can be decreased to eight. The association rule algorithm caters for the correlation between wind speeds in different areas.

The association algorithm can be summarized in the following steps:

*Step 1:* Generate all association rules in the form if {A,B,C,D,...} then {E,F,G,...}, where A, B, C, D, E, F, G,... are items.

*Step 2:* Calculate confidence of the generated rules, i.e., if A then B using:

$$\text{Confidence} = \frac{\text{number of records containing both A and B}}{\text{number of records containing A}}$$

*Step 3:* Calculate support of the generated rules, i.e., if A then B using:

$$\text{Support} = \frac{\text{number of records containing both A and B}}{\text{total number of records}}$$

*Step 4:* Check if support is less than a pre-defined threshold, i.e., minsup.

*Step 5:* Check if confidence is less than a pre-defined threshold, i.e., minconf

*Step 6:* Prune rules that fail the minsup and minconf thresholds.

The wind speed of each location is labeled using a rank-based system. The developed ranking system distributes wind speeds evenly, measuring them only relative to a given location, but not according to the real value of any given speed. The developed ranking-based system includes five labels that identify the level of the wind speed: very high (VH), high (H), medium (M), low (L), and very low (VL). The database resulting after the labels have been assigned based on the wind speed ranking is shown in Table 6.

**Table 6.** Attribute list with assigned labels.

Record	Center	South	East	North	West	Output
1	VH	H	VL	M	L	case 1
2	VH	M	VL	H	L	case 2
3	VH	M	VL	H	L	case 2
4	VH	L	VL	H	M	case 3
5	H	L	VL	M	VH	case 4
6	H	M	VL	L	VH	case 5
7	VH	M	VL	L	H	case 6
8	H	M	VL	L	VH	case 5
9	H	L	VL	M	VH	case 4
10	VH	L	VL	M	H	case 7
11	VH	L	VL	H	M	case 3
12	VH	L	VL	H	M	case 3

To minimize the number of output attributes, an association rule algorithm is applied for analyzing all of the relations between the cases. Table 7 shows the resulting cases and the corresponding locations of the rules that produce support and confidence levels greater than a given minimal support threshold (minsup = 0.01) and a given minimal confidence threshold (minconf = 0.5).

**Table 7.** Cases ordered according to location preferences.

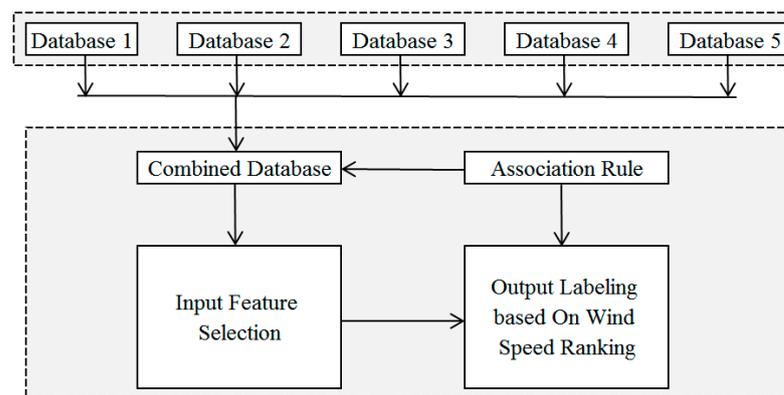
	Region 1	Region 2	Region 3	Region 4	Region 5
case 1	Center	South	North	West	East
case 2	Center	North	South	West	East
case 3	Center	North	West	South	East
case 4	West	Center	North	South	East
case 5	West	Center	South	North	East
case 6	Center	West	South	North	East
case 7	Center	West	North	South	East
case 8	West	Center	North	East	South

Table 8 provides a sample of association rules with their support and confidence levels. The table shows the minimum number of cases that can be achieved using the association algorithm with a unity confidence level.

**Table 8.** Support and confidence levels of sample rules.

Rule	Support	Confidence
Center = VH, South = H, East = VL, North = M, West = L → Case 1	1/12 = 0.083	1
Center = VH, South = H, East = VL, North = M, West = L → Case 2	2/12 = 0.16	1
Center = VH, South = H, East = VL, North = M, West = L → Case 3	3/12 = 0.25	1
Center = VH, South = H, East = VL, North = M, West = L → Case 4	2/12 = 0.16	1
Center = VH, South = H, East = VL, North = M, West = L → Case 5	2/12 = 0.16	1

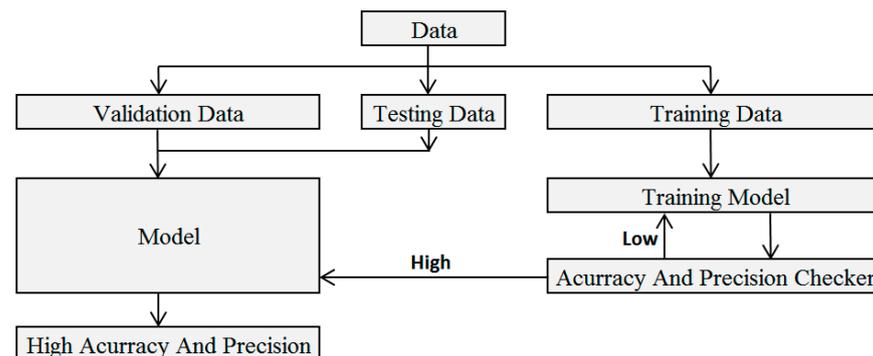
Figure 2 summarizes all of the steps described above for the data-gathering and preprocessing stages.



**Figure 2.** Data-gathering and preprocessing stages.

### 2.3. Learning and Classification Phase

Figure 3 displays a flowchart of the developed classification algorithm, which governs the processing of the data through three stages: training, testing, and validation. First, the training data are applied to the decision tree algorithm to obtain the initial model. For each iteration, the accuracy and precision are then calculated as a means of achieving the optimal model; the test data are applied so that the performance and efficiency of the model can be verified; and in the final step, the remaining verification data are employed to ensure that the results produced by the model have a high degree of accuracy and precision.



**Figure 3.** Developed classification algorithm.

A decision tree partitions the input space of the dataset into mutually exclusive regions by assigning each region a label. The decision tree begins with a root node and ends with a leaf node [23]. Multiple branches are formed between the root and the leaf nodes. The decision tree algorithm is performed based on splitting data into multiple regions and each region is divided into small parts. Furthermore, splitting continues until the terminal node reaches leaf nodes. The splitting is formed based on an impurity measure. Two common measures are used to obtain impurity values, Gini index, and entropy. In this paper, entropy is used as impurity measure that evaluates the homogeneity of the partition nodes too. The following steps summarize the decision tree algorithm.

*Step 1:* the entropy of the root node with  $n$  branches is calculated as

$$E(\text{root}) = - \sum_{i=1}^n P_i \log_2 P_i \quad (1)$$

where  $p$  is the fraction of records that belongs to class  $i$  at the node.

*Step 2:* the entropy of each partition with  $J$  sub classes is calculated as

$$E(\text{partition}) = - \sum_{i=1}^J P_i \log_2 P_i \quad (2)$$

Step 3: The branch entropy is calculated using the individual k partition entropies as

$$E(\text{branch}) = \sum_{i=1}^k \frac{n_i}{n} E(\text{partition } i) \tag{3}$$

where  $n_i$  is the number of records at partition  $i$ ,  
 $n$  is number of records at branch, and  
 $E$  is the entropy.

Step 4: The  $GAIN_{Split}$  which is used to decide the best partition is the best. The partition that produces the most reduction is chosen The  $GAIN_{Split}$  is shown below

$$GAIN_{Split} = E(\text{root}) - E(\text{branch}) \tag{4}$$

where  $E$  is the entropy.

If all input attributes are used, the algorithm for decision tree induction is as shown in Figure 4.

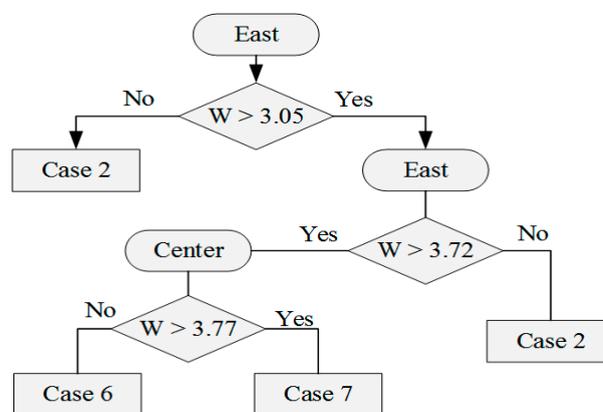


Figure 4. Decision tree model based on each location and its wind speeds.

If the prediction order is requested for a specific month and the wind speeds are unavailable at that moment, the decision tree induction model shown in Figure 5 is used. This model is based on a single input attribute: “month”.

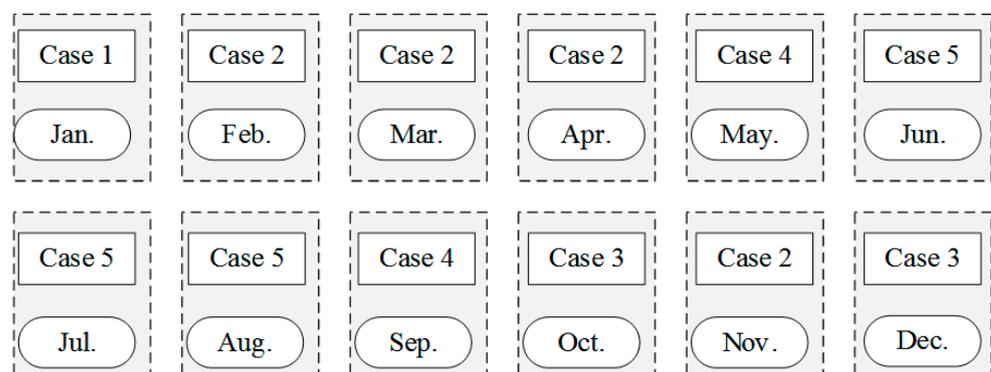


Figure 5. Decision tree model based on the month.

A new model, Model 2, is implemented based on the output of the previous model, Model 1, as shown in Figure 6. The implementation involves a comparison of the output for the five cases generated from the first model with that of the eight cases from the original training data. The output from these five cases along with the output from the original cases is then used as input to a similarity algorithm.

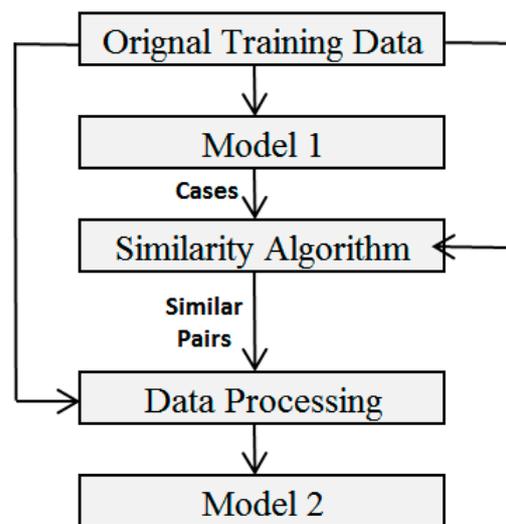


Figure 6. Development phases of Model 2.

Next, the similarity algorithm measures the similarity score between the five cases and each case from the original data, i.e., Case 8 is similar to Case 4, Case 7 is similar to Case 3, and Case 6 is similar to Case 5. The algorithm relies on edit distance, which is a technique for quantifying how dissimilar two strings (e.g., words) are to one another based on a count of the minimum number of operations required to transform the first string into the second. The edit distance between two cases for the five locations is the minimum number of operations required for transforming one case into another case. For example, the edit distance between “case 1 case 2 case 3 case 4 case 5” and “case 1 case 3 case 2 case 4 case 5” is two. A flowchart of the similarity algorithm is shown in Figure 7.

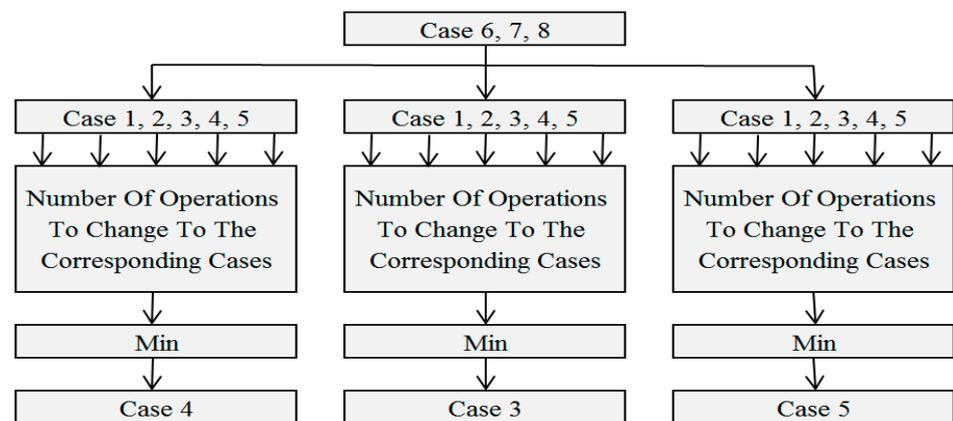


Figure 7. Similarity algorithm.

The resulting similarity pairs are employed for reprocessing the original training data through the replacement of the original cases with the similar cases, as shown in Table 9 compared with Table 6. The final step is that the resulting training data are applied for teaching Model 2, with the use of the decision tree as previously performed for developing Model 1. The degree of accuracy of Model 2 is then increased to 100%.

**Table 9.** The resulted database after similarity algorithm.

Record	Center	South	East	North	West	Output
1	VH	H	VL	M	L	case 1
2	VH	M	VL	H	L	case 2
3	VH	M	VL	H	L	case 2
4	VH	L	VL	H	M	case 3
5	H	L	VL	M	VH	case 4
6	H	M	VL	L	VH	case 5
7	VH	M	VL	L	H	case 5
8	H	M	VL	L	VH	case 5
9	H	L	VL	M	VH	case 4
10	VH	L	VL	M	H	case 3
11	VH	L	VL	H	M	case 3
12	VH	L	VL	H	M	case 3

### 3. Experiments and Results

For this study, Waikato Environment for Knowledge Analysis (Weka) software was employed [38] for constructing decision trees according to the training set, using the standard J48 algorithm [39–42]. This algorithm has been selected as one of the top 10 algorithms in data mining [43]. Java was used as the development language with J2SDK version 1.6.0\_22. Weka version 3.8.4 was employed for the experimental component of the model development.

The first use of Weka software is to do data pre-processing before applying machine learning algorithms on it. The wind speed data for selected sites are recalled from CSV files. This can be done by clicking the “Open file” button and loading the data file. The loaded dataset is then processed to Cross-validation to randomly partition the data into  $k$  subsamples for training and testing. The number entered in the Fold section is used to divide the dataset into the number of Folds specified. Then classifier J48 is used as a decision tree to create a pruned tree. The Classifier Model part illustrates the model as a tree and gives some information about the tree, like the number of leaves, size of the tree, etc. Next is the stratified cross-validation part and it shows the error rates. It shows how successful the model is. By right-clicking “Visualize tree”, the developed model’s tree can be visualized.

The performance measurements for this work were recall, precision, the classifier F1-score, and accuracy. Examining the data for accuracy and precision establishes the credibility of the results. Accuracy refers to how closely the measurements match the desired “true” value. Precision indicates how well repeated measurements agree with and are approximate to one another. As with the order of decisions about wind speed location, it is important that the values be close, i.e., a high level of precision, and at the same time, that the decisions be correct, i.e., a high degree of accuracy. The accuracy and the precision is defined in (5) and (6)

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (5)$$

where  $T_P$  is true positive,  $T_N$  is true negative,  $F_P$  is false positive and  $F_N$  is false negative. The true positive and true negative is the outcomes where the developed model correctly predicts the cases. By contrast, a false positive and a false negative are the outcomes for which the model incorrectly predicts the cases.

$$\text{Precision}(P) = \frac{T_P}{T_P + F_P} \quad (6)$$

Recall (R) is the ratio of the accurate data to the total relevant data. Its formula is shown in (7).

$$R = \frac{T_P}{T_P + F_N} \quad (7)$$

where  $T_P$  is true positive and  $F_N$  is false negative.

The classifier F1-score is calculated based on the harmonic mean. It is given as

$$F_1 = \frac{2 * P * R}{P + R} \quad (8)$$

where P is the precision and R is the recall.

The performance measurement results are listed in Table 10.

**Table 10.** Overall performance results (training and validation set).

Model	Model 1	Model 2
Total number of instances	11.43	11.43
Correctly classified instances	95.26%	100%
Kappa statistic	0.93	1
Mean absolute error	0.027	0.07
Root mean squared error	0.11	0.12
Relative absolute error	5.64%	25.03%
Root relative squared error	24.36%	31.60%

Measurements from another performance indicator established with the use of a confusion matrix are presented in Table 11. The confusion matrix was built based on the data testing, and a confusion matrix was constructed for each class in the form shown in Table 12.

**Table 11.** Recall, precision, and F1-score measurements for each class.

Recall		Precision		F1-Score		Class
Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	
96.89%	100%	98.19%	100%	0.97541	1	case1
92.84%	100%	90.96%	100%	0.91895	1	case2
92.36%	100%	88.80%	100%	0.90546	1	case3
93.80%	100%	95.58%	100%	0.94681	1	case4
97.43%	100%	98.45%	100%	0.97941	1	case5

Because of the limited number of training cases, exercising care when minimizing and reserving the number of training samples for testing purposes is extremely important. Cross-validation was employed for testing, checking, and verifying the generalizability of the model. In training any model, a frequent tendency is to overfit, and cross-validation was applied as a means of avoiding this effect. The best way to improve the performance of a system is to reserve a small portion of the training data itself for use in validating the model since this approach provides an idea of the ability of the model to predict the previously unseen reserved data. K-fold cross-validation is a technique commonly used for this purpose. In a 10-fold version of k-fold cross-validation, the training set is randomly split into groups of 10 that have approximately the same size. The classifier is then trained using eight subsets. One of the two remaining subsets is used for validation and the last, for testing. This process is repeated until all folds, one by one, have an opportunity to be the assigned test version. This technique establishes the generalizability of the model, especially when limited data makes it difficult to break the data down into test data and

training data. Table 13 shows the average degree of accuracy for 2-fold, 4-fold, 6-fold, and 8-fold cross-validation and for the 10-fold cross-validation used in this paper.

**Table 12.** Confusion matrix (training and validation set).

		Real System				
		case 1	case 2	case 3	case 4	case 5
Model 1	case 1	2500	28	13	0	5
	case 2	18	1803	74	13	74
	case 3	2	91	1475	89	4
	case 4	16	20	33	1665	8
	case 5	44	0	2	8	3450
Model 2	case 1	2578	0	0	0	0
	case 2	0	1942	0	0	0
	case 3	0	0	1597	0	0
	case 4	0	0	0	1775	0
	case 5	0	0	0	0	3541

**Table 13.** Degrees of accuracy for 2-fold, 4-fold, 6-fold, 8-fold, and 10-fold cross-validation.

K-fold	Accuracy (%)	
	Model 1	Model 2
2-fold	68.654	69.38
4-fold	65.145	88.62
6-fold	78.224	95.64
8-fold	87.325	98.32
10-fold	95.26	100.00

In this research, a unique system was developed to arrange places according to wind speed. The process was carried out through three stages, i.e., the data collection stage, the processing stage, and the design stage. In the first stage, data are collected from different places, for example in the center, north, south, east, and west of the region. These data contain wind speed and other additional information such as location data from longitude and latitude and the date of collected samples. The data are collected in a central database and this database contains all the information deduced from the databases spread in different places. In the second stage (data processing stage), the information that is not useful in this research, such as longitude and latitude, is discarded and the date is replaced by the month. Then the central database is rearranged and the number of cases is reduced by using the association rules (a famous method of finding relationships) and this is done by studying all cases and their relationship to each other. This developed theory can be used for other places and other databases, and the developed method does not exist before in the literature. Machine learning methods depend on a set of algorithms, and these algorithms are applied to a set of data to build models that help in making decisions. This model is not limited to these data. This model can be used as a solid foundation to address similar problems in different areas. Other factors such as the direction of the wind, the maximum and minimum wind speed per day are important and might serve different applications. In this paper, however, the focus was on the wind speed to achieve a specific goal of providing the network operators with an understanding of the possible productivity of each wind site location, thus facilitating the optimal management and installation of wind plants and network operations. Such other factors open the door for great future

work. The wind direction especially will play an important role in determining the place of the wind plants and the layout of wind turbines.

The proposed model shows great promise, so that two locations are sufficient for obtaining the order of preference of the locations. For example, if it is known only that the wind speed in the east region is below 3.05 m/s, then this scenario follows Case 2. Once the cause is known, the order of the wind speed values at all locations can be determined. If the wind speed in the east region is greater than 3.5 m/s but less than 3.72 m/s, the status of the wind speed at the other locations can be extracted from the Case 4 scenario. If the wind speed in the east region is greater than 3.77 m/s, the status of the wind speed at the center location and whether it follows Case 6 or Case 7 can be determined. Indeed, this feature of the proposed model saves the time and effort that would otherwise be required for predicting the wind speed at multiple locations. This model can thus be very helpful to system operators who desire an easy, quick, and accurate method of determining the status of the wind speeds at different locations.

#### 4. Conclusions

This paper has presented a machine learning-based decision-making method for the assessment of potential wind speed productivity in different locations. To preserve the characteristics of the correlations among these sites, the new method employs wind speed data from multiple sites as a single package. Machine learning using Weka software is then employed to test the correlations among the sites to rank the sites into different cases. Wind speed becomes the primary classification factor for prioritizing the sites in order. The implementation of training tests for big data sets improves the prediction of appropriate locations for wind farms. Using real data, the decision model has been constructed, tested, and verified. The data is a part of the Renewable Resource Monitoring and Mapping (RRMM) Program prepared by King Abdullah City for Atomic and Renewable Energy (KACARE). KACARE monitored and recorded the wind speed data at different installed stations in the Kingdom of Saudi Arabia at 3 m height. 10-fold cross-validation was used in the experimental part. The proposed model shows great results, so that the information about two locations is sufficient for obtaining the order of the remaining locations. The developed model shows high accuracy (up to 95.26%) in the test data. The final performance of Model 1 has been improved by developing Model 2, where the accuracy has increased to 100%. Electric network planners could use the proposed model as a means of enhancing their ability to conduct feasibility studies for any plans for establishing wind farm projects at different distanced locations. A system operator could also use this method for assessing likely wind power productivity at each site so that network operational activities can be managed effectively. The results of this study also offer electricity market investors helpful input for making appropriate investment decisions.

**Author Contributions:** The authors' contributions are as follows: data mining, data analysis and software, A.M.M.; Conceptualization and methodology, A.A. and S.A.; investigation, M.A.O.; writing—original draft preparation, D.A. and J.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received funding from the Deputyship for Research and Innovation, Ministry of Education, Saudi Arabia under project number (IFP-2020-02).

**Institutional Review Board Statement:** Not Applicable.

**Informed Consent Statement:** Not Applicable.

**Data Availability Statement:** The data that support the findings of this study are available from the corresponding author.

**Acknowledgments:** The authors extend their appreciations to Deputyship for Research and Innovation, Ministry of Education, Saudi Arabia, for funding this research work through project number (IFP-2020-02).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Statista. Installed Wind Power Capacity—Worldwide, 2001–2019. Available online: <https://www.statista.com/statistics/268363/installed-wind-power-capacity-worldwide> (accessed on 28 April 2020).
2. Renewables Information 2019—Analysis. Comprehensive Historical Review and Current Market Trends in Renewable Energy, IEA. Available online: <https://www.iea.org/reports/renewables-information-overview> (accessed on 1 July 2020).
3. Georgilakis, P.S. Technical challenges associated with the integration of wind power into power systems. *Renew. Sustain. Energy Rev.* **2008**, *12*, 852–863. [\[CrossRef\]](#)
4. Albadi, M.H.; El-Saadany, E.F. Overview of wind power intermittency impacts on power systems. *Electr. Power Syst. Res.* **2010**, *80*, 627–632. [\[CrossRef\]](#)
5. Billinton, R.; Wangdee, W. Reliability-based transmission reinforcement planning associated with large-scale wind farms. *IEEE Trans. Power Syst.* **2007**, *22*, 34–41. [\[CrossRef\]](#)
6. Billinton, R.; Karki, R.; Gao, Y.; Huang, D.; Hu, P.; Wangdee, W. Adequacy assessment considerations in wind integrated power systems. *IEEE Trans. Power Syst.* **2012**, *27*, 2297–2305.
7. Chao, H.; Hu, B.; Xie, K.; Tai, H.-M.; Yan, J.; Li, Y. A Sequential MCMC Model for Reliability Evaluation of Offshore Wind Farms Considering Severe Weather Conditions. *IEEE Access* **2019**, *7*, 132552–132562. [\[CrossRef\]](#)
8. Almutairi, A.; Ahmed, M.H.; Salama MM, A. Use of MCMC to incorporate a wind power model for the evaluation of generating capacity adequacy. *Electr. Power Syst. Res.* **2016**, *133*, 63–70. [\[CrossRef\]](#)
9. Gao, Y.; Billinton, R. Adequacy assessment of generating systems containing wind power considering wind speed correlation. *IET Renew. Power Gener.* **2009**, *3*, 217–226. [\[CrossRef\]](#)
10. Chen, F.; Li, F.; Wei, Z.; Sun, G.; Li, J. Reliability models of wind farms considering wind speed correlation and WTG outage. *Electr. Power Syst. Res.* **2015**, *119*, 385–392. [\[CrossRef\]](#)
11. Sun, M.; Feng, C.; Zhang, J. Conditional aggregated probabilistic wind power forecasting based on spatio-temporal correlation. *Appl. Energy* **2019**, *256*, 113842. [\[CrossRef\]](#)
12. Hasan, S.; Al-Aqeel, T.; Peerbocus, N. Saudi Arabia’s Unfolding Power Sector Reform: Features, Challenges and Opportunities for Market Integration. *ResearchGate* **2020**. [\[CrossRef\]](#)
13. Baseer, M.; Meyer, J.; Rehman, S.; Alam, M. Wind power characteristics of seven data collection sites in Jubail, Saudi Arabia using Weibull parameters. *Renew. Energy* **2017**, *102*, 35–49. [\[CrossRef\]](#)
14. Rehman, S.; Halawani, T.; Mohandes, M. Wind power cost assessment at twenty locations in the Kingdom of Saudi Arabia. *Renew. Energy* **2003**, *28*, 573–583. [\[CrossRef\]](#)
15. Bassyouni, M.; Gutub, S.A.; Javaid, U.; Awais, M.; Rehman, S.; Hamid, S.M.-S.A.; Abdel-Aziz, M.H.; Abouel-Kasem, A.; Shafeek, H. Assessment and analysis of wind power resource using weibull parameters. *Energy Explor. Exploit.* **2015**, *33*, 105–122. [\[CrossRef\]](#)
16. Qin, Z.; Li, W.; Xiong, X. Generation system reliability evaluation incorporating correlations of wind speeds with different distributions. *IEEE Trans. Power Syst.* **2013**, *28*, 551–558. [\[CrossRef\]](#)
17. Almutairi, A.; Nassar, M.E.; Salama, M.M.A. Statistical evaluation study for different wind speed distribution functions using goodness of fit tests. In Proceedings of the IEEE Electrical Power and Energy Conference (EPEC) 2016, Ottawa, ON, Canada, 12–14 October 2016.
18. Ouarda, T.; Charron, C.; Shin, J.-Y.; Marpu, P.R.; Al-Mandoos, A.; Al-Tamimi, M.; Ghedira, H.; Al Hosary, T. Probability distributions of wind speed in the UAE. *Energy Convers. Manag.* **2015**, *93*, 414–434. [\[CrossRef\]](#)
19. Almonacid, F.J.M.F.; Rus, C.; Hontoria, L.; Munoz, F.J. Characterisation of PV CIS module by artificial neural networks. A comparative study with other methods. *Renew. Energy* **2010**, *35*, 973–980.
20. Khait, J.A.; Mansour, A.M.; Obeidat, M. Classification based on Gaussian-kernel Support Vector Machine with Adaptive Fuzzy Inference System. *MargIn* **2018**, *5*, 16–24.
21. Mansour, A.M.; Alaqtash, M.M.; Obeidat, M. Intelligent Classifiers of EEG Signals for Epilepsy Detection. *WSEAS Trans. Signal Process.* **2019**, *15*, 2224–3488.
22. Obeidat, M.A.; Mansour, A.M. EEG Based Epilepsy Diagnosis System using Reconstruction Phase Space and Naïve Bayes Classifier. *WSEAS Trans. Circuits Syst.* **2018**, *17*, 2224–2266.
23. Mansour, A.M. Decision Tree-Based Expert System for Adverse Drug Reaction Detection using Fuzzy Logic and Genetic Algorithm. *Int. J. Adv. Comput. Res.* **2018**, *8*, 110–128. [\[CrossRef\]](#)
24. Hawashin, B.; Mansour, A.; Aljawarneh, S. An Efficient Feature Selection Method for Arabic Text Classification. *Int. J. Comput. Appl.* **2013**, *83*, 17. [\[CrossRef\]](#)
25. Ayman, M.M. Texture Classification using Naïve Bayes Classifier. *Int. J. Comput. Sci. Netw. Secur.* **2018**, *18*, 112–120.
26. Al Nadi, D.A.; Mansour, A.M. Independent Component Analysis (ICA) for texture classification. In Proceedings of the 5th International Multi-Conference on Signals and Devices, IEEE SSD, Amman, Jordan, 20–23 July 2008.
27. Hawashin, B.; Mansour, A.; Abukhait, J.; Khazalah, F.; Alzubi, S.; Kanan, T.; Obaidat, M.; Elbes, M. Efficient Texture Classification Using Independent Component Analysis. In Proceedings of the IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 9–11 April 2019; pp. 544–547.
28. Renewable Resource Atlas. King Abdullah City for Atomic and Renewable Energy. Available online: <http://rratlas.kacare.gov.sa> (accessed on 10 June 2020).

29. Navida, W. *Statistics for Engineers and Scientists*, 3rd ed.; McGraw-Hill: New York, NY, USA, 2011; ISBN 978-0-07-337633-2.
30. Zhang, Z. Missing data imputation: Focusing on single imputation. *Ann. Transl. Med.* **2016**, *4*, 1.
31. Curley, C.; Krause, R.M.; Feiock, R.; Hawkins, C.V. Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database. *Urban Aff. Rev.* **2019**, *55*, 591–615. [[CrossRef](#)]
32. Ordiano, J.; Ángel, G.; Waczowicz, S.; Reischl, M.; Mikut, R.; Hagenmeyer, V. Photovoltaic power forecasting using simple data-driven models without weather data. *Comput. Sci. Res. Dev.* **2017**, *32*, 237–246.
33. Khan, S.I.; Hoque, A.S.M.L. SICE: An improved missing data imputation technique. *J. Big Data* **2020**, *7*, 3. [[CrossRef](#)]
34. Angulakshmi, M.; Deepa, M.; Sudha, S.; Brindha, K. Association Rule Modeling using UML and Apriori Algorithm. In Proceedings of the International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore Institute of Technology, Vellore, India, 24–25 February 2020; pp. 1–5.
35. Agapito, G.; Milano, M.; Guzzi, P.H.; Cannataro, M. Mining Association Rules from Disease Ontology. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 2239–2243.
36. Kharya, S.; Soni, S.; Swarnkar, T. Weighted Bayesian Association Rule Mining Algorithm to Construct Bayesian Belief Network. In Proceedings of the 2019 International Conference on Applied Machine Learning (ICAML), Bhubaneswar, India, 25–26 May 2019; pp. 27–33.
37. Chen, C.; Chou, H.; Hong, T.; Nojima, Y. Cluster-Based Membership Function Acquisition Approaches for Mining Fuzzy Temporal Association Rules. *IEEE Access* **2020**, *8*, 123996–124006. [[CrossRef](#)]
38. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [[CrossRef](#)]
39. Bhargava, N.; Sharma, G.; Bhargava, R.; Mathuria, M. Decision tree analysis on j48 algorithm for data mining. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2013**, *3*, 1114–1119.
40. Kaur, G.; Chhabra, A. Improved J48 classification algorithm for the prediction of diabetes. *Int. J. Comput. Appl.* **2014**, *98*, 13–17. [[CrossRef](#)]
41. Ruggieri, S. Efficient C4.5 [classification algorithm]. *IEEE Trans. Knowl. Data Eng.* **2002**, *14*, 438–444. [[CrossRef](#)]
42. Hssina, B.; Merbouha, A.; Ezzikouri, H.; Erritali, M. A comparative study of decision tree ID3 and C4.5. *Int. J. Adv. Comput. Sci. Appl.* **2014**, *4*, 13–19. [[CrossRef](#)]
43. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37. [[CrossRef](#)]