Automated Weather Event Analysis with Machine Learning

Nasimul Hasan[1], Md. Taufeeq Uddin[1], Nihad Karim Chowdhury[2]

[1] Department of Computer Science and Engineering, International Islamic University Chittagong, Bangladesh
[2] Department of Computer Science and Engineering, University of Chittagong, Bangladesh
nasimul_hsn@yahoo.com, mohamed.taufeeq.uddin@gmail.com, nihadcuet@gmail.com

***Abstract*—Weather forecasting has numerous impacts in our daily life from cultivation to event planning. Previous weather forecasting models used the complicated blend of mathematical instruments which was insufficient in order to get higher classification rate. In contrast, simple analytical models are well-suited for weather forecasting tasks. In this work, we focus on the weather forecasting by means of classifying different weather events such as normal, rain, and fog by applying comprehensible C4.5 learning algorithm on weather and climate features. The C4.5 classifier classifies weather events by building the decision tree using information entropy from the set of training samples. We conducted experiments on LA weather history dataset; from evaluation results, it is revealed that C4.5 classifier classifies weather events with f-score of around 96.1%. This model also indicates that climate features such as rainfall, visibility, temperature, humidity, and wind speed are highly discriminative toward events classification.**

***Keywords*—Weather Events, Forecasting, Machine Learning**

I. INTRODUCTION

The situation of weather plays a crucial role in almost every aspects of human life. Note that intelligent weather analysis techniques can help us to make efficient decisions that can lead us to save valuable lives, properties, and time. As a consequence, researchers focus on the automated analysis of weather and climate data such as forecasting rainfall, predicting air temperature to understand and to extract useful information. As modernization continued, prediction of weather events draws more attention. From the very beginning of civilization, people want to know the pattern of weather change. Discovering the weather pattern and forecasting weather has been a field of interest from the exploration of science and technology.

Weather forecasting involves foreseeing how the current situation with the air will change in which present climate conditions are taken by ground perceptions such as from boats, airplane, radiosondes, Doppler radar, and satellites. The collected data is then sent to meteorological focuses in which the information are gathered, analyzed, and made into an assortment of outlines, maps, and charts. Algorithms exchange a huge number of perceptions onto surface and upper air maps and draw the lines on the maps with assistance from meteorologists. Algorithms draw the maps as well as anticipate how the maps will look at some point later on. The determination of climate condition using algorithms is outlined as numerical or computational climate forecasting [1].

Numerical or computational models for weather forecasting are the dynamic representations of the systems is being used in present days. These models discretize regions or bodies in a few measurements by separately utilizing estimated capacities to portray the behavior of the climatic variables of interest [2]. Nowadays, numerical or computational models are irreplaceable for atmosphere estimation. For instance, Bayesian networks [3] with time differed scaling features can be used to review whether there are factually noteworthy patterns in the climate information. In addition, Tae-wong [4] demonstrated a space-time model that displays the short time and geographical conditions of the day by day rain event.

Although there are several techniques available for weather forecasting, weather forecasting is actually a challenging task due to the complicated physics behind weather which depends on numerous features, and which is also boisterous and deterministically confusing natural event. Moreover, people produce numerous disasters, and change of climate or characteristics of climate such as air temperature, rainfall, dew point temperature, visibility, and humidity displays a strong role on the weather. Notice that several automated techniques including Artificial Neural Network (ANN), Support Vector Regressor (SVR), Genetic Algorithm (GA) were applied to forecast or model weather [5] [6] [7] [8], where ANN was the most commonly used technique that can forecast weather with decent performance rate.

In this paper, we apply comprehensible tree based learning algorithm, namely C4.5 [9] to classify the weather events as normal, rain, and fog using a set of 19 features e.g. minimum, maximum, and mean air temperature, dew point temperature, humidity, visibility, wind speed, sea level pressure, and total and highest rainfall that capture numerous information about climate and weather. Hence, C4.5 is a statistical classifier used to build a decision tree for classification in which C4.5 evaluates the goodness of a test using information theory-based formula choosing the test with the maximum amount of information from the set of examples. Note that the proposed C4.5 algorithm comprises of several good qualities including comprehensibility, pruning tree after construction, competitive computational performance both in training and prediction steps, and capability to handle both continuous and discrete features, missing value features, and feature with different costs. In our experiments, we found that C4.5 classified three weather events e.g. normal, rain, and fog with f-score of **0.979, 0.84,** and **0.845**, respectively, on LA weather history dataset [10]. To the best of our knowledge, this is the first time we apply C4.5 classifier for weather events classification task.

II. RELATED WORK

Over the last few decades, researchers conducted a number of automated analyses on weather and climate data such as dew point temperature prediction using several Artificial Intelligence (AI) techniques from different sub-domains of AI including model output statistics, fuzzy logic, expert system, machine learning, and data mining. For instance, Chevalier et al. [11] trained SVR on small, and minimally pre-processed meteorological dataset to predict air temperature. Devi et al. [12] developed ANN based temperature forecasting model using real-time quantitative data about the current state of the atmosphere. Olaiya and Adeyemo [13] also investigated the performance of ANN and decision trees during the classification of maximum, minimum, and mean temperature, rainfall, evaporation, and wind speed on meteorological data gathered from Nigeria. Lin and Chen [8] designed typhoon rainfall forecasting model using ANN feeding eight typhoon characteristics and spatial rainfall information, where they found that excessive spatial rainfall information may not increase the generalization of the forecasting model. Mohammadi et al. [14] predicted the dew point temperature on the daily scale on different climate conditions applying extreme learning machine algorithm on five common climate-related features such as mean air temperature, relative humidity, atmospheric pressure, vapor pressure and horizontal global solar radiation.

As per our knowledge based on literature review, Awan and Awais's research [15] is the only similar study available in the literature that also attempted to predict weather events. In their research, they aimed to predict weather events based on fuzzy RBS method for Lahore, Pakistan. They used two different datasets of 365 examples with only 4 features, and 2500 examples with 17 features e.g. temperature, dew point, humidity, sea level, visibility, wind speed, respectively, for experimentation. They mentioned in their finding that fuzzy RBS method was sensitive to random sampling with replacement technique that was applied to produce training and test dataset. In contrary, we applied comprehensible tree-based machine learning algorithm for events classification for Los Angeles, California, the USA in which we used 5325 examples with 19 features extending the feature set to include rainfall information to build the weather events prediction model.

III. LA WEATHER HISTORY DATASET

In this work, we used weather and climate data (LA weather history) [10] of 14 years from 2001 to 2015 of Los Angeles, California, USA, which is available at `http://www.wunderground.com/history/airport/KCQT/2015/1/1/CustomHistory.html`. LA weather history dataset comprises of 19 weather and climate features, and weather events label that contains 3 different events or classes such as normal, rain, and fog. The comprised features of LA weather history dataset covers the following information [16] about weather and climate:

- **Temperature (°C)** has the impact in making the kind of precipitation that outcomes from the water vapor exchange in between the earth and air. LA weather history dataset contains three temperature features e.g. minimum, maximum, and mean temperature.
- **Dew point (°C)** determines the comfort level of feeling on a day that figures out if it will rain or snow. The saturation point is that temperature at that moisture (water vapor) within the air begins to condense. The hotter the air is, the more moisture it will hold. LA weather history dataset comprises three dew point features e.g. minimum, maximum, and mean dew points.
- **Humidity (%)** is just a consequence of the way that water exists at different rates in the water vapor in our climate. The surface temperature of the earth would be much cooler with no water vapor in the environment. LA weather history dataset includes three humidity features e.g. minimum, maximum, and mean humidities.
- **Sea level pressure (hPa)** is the air pressure at the sea level at the time of measuring at a given rise ashore. The station pressure adjusted to water level by considering that the temperature falls at a lapse rate of 6.5 thousand per kilometer within the fictive layer of air between the corresponding weather station and water level. LA weather history dataset contains three sea level pressure features e.g. minimum, maximum, and mean sea level pressure.
- **Visibility (m)** in meteorology is a measure of the distance at which an article or light can be obviously recognized. Visibility influences all types of movement including roads, sailing, and aviation. Meteorological visibility refers to the transparency of air in the dark, and it remains an equivalent as in daylight for a similar air. LA weather history dataset comprises three visibility features e.g. minimum, maximum, and mean visibilities.
- **Wind speed (km/h)** displays how quick the air is moving past a specific point. It gives a notion regarding the position of low and high-pressure territories. In the event that wind speed increases, it is a sign of fortification of pressure systems. It brings cool or hot air from the spot of root or from the spots through which they pass. LA weather history dataset includes three wind speed features e.g. minimum, maximum, and mean wind speed.
- **Rainfall (mm)** is an influential feature / predictor for weather forecasting that is a major component of climate that helps to continue ecosystem the way it should be. Cultivation depends mostly on rainfall and it is one of the richest sources of fresh water. LA weather history dataset contains two rainfall features e.g. maximum, and total rainfall.

IV. C4.5 FOR WEATHER EVENTS CLASSIFICATION

C4.5 [17] is a statistical classifier used to build a decision tree for classification. The key idea beneath C4.5 algorithm is that C4.5 creates decision trees using information entropy H from set of training samples e.g. $S = s_1, s_2, \ldots, s_n$ of pre-classified samples, where each sample s_i comprises of N-dimensional vector $x_{1,i}, \ldots, x_{N,i}$ in which x_j denotes feature of the sample and class in which s_i falls. At each node of the tree, C4.5 selects the feature that most effectively splits its set of samples into subsets using normalized information gain as splitting gain criteria. C4.5 makes a decision using the feature with highest information gain where information gain I_{Gain} is measured as follows.

$$I_{Gain}(Event, Feature) = H(Event) - H(Event|Feature) \tag{1}$$

where $Event$ denotes weather event class, and $Feature$ denotes available weather and climate features used in weather prediction model. Figure 1 shows a decision tree produced from C4.5 classifier on LA weather history dataset. The technical description of the C4.5 algorithm [18] is given in Algorithm 1.

Algorithm 1 C4.5 for weather events classification

Check for base cases
for each feature f **do**
    Compute normalized I_{Gain} ratio from splitting on f
end for
Let f_{best} be the feature with maximum normalized I_{Gain}
Build decision node based on the splitting on f_{best}
Recurse on the sublists achieved via splits on f_{best}, and finally, include those node as children of node
return decision trees

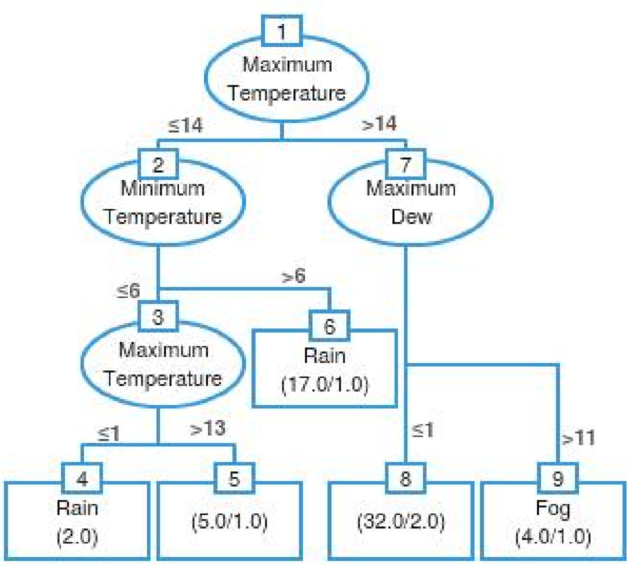

Fig. 1. An example decision tree generated by C4.5 classifier

V. EXPERIMENTS AND RESULTS

The proposed weather events classification model was evaluated on LA weather history dataset using J48, an open source implementation of C4.5 on Weka [19] data mining and machine learning tool, applying cross-fold validation e.g. $5-fold$, $10-fold$, $20-fold$, and random splitting e.g. $50\%-50\%$, $60\%-40\%$, $70\%-30\%$ strategies. We benchmarked C4.5 classifier against classic learning algorithm naive Bayes, and displayed the comparison in Table I and II. Each of the experiments was run for 20 times with different random seeds, and the results were obtained by averaging over 20 different experimental runs. We produced accuracy, precision, recall, and f-score for each weather event class to demonstrate the performance of the models. The larger values of the performance metric accuracy, precision, recall, and f-score indicate the higher weather events classification performance.

Note that, we modeled the weather events prediction task as classification problem as we aimed to estimate the probable weather event using weather and climate features. Table I displays the performance of C4.5 and naive Bayes classifiers for multiple cross-fold validations and random splitting strategies, where Table II shows the performance of C4.5 and naive Bayes classifiers during the classification of three weather events e.g. normal, rain, and fog. From Table II, it can be indicated that C4.5 classifier was better than naive Bayes since C4.5 classified all three events with higher f-score than naive Bayes. More precisely, naive Bayes classified fog event with a low precision rate of 34.5%, and f-score rate of 49.9% that was extremely worst than the performance of C4.5 classifier. Another important point to note is that C4.5 basically confused fog and rain events with normal, and normal event with rain event, while naive Bayes widely confused fog event with both normal and rain events, and rain event with the normal event. According to the experimental results from Table II, we can outline that the proposed C4.5 classifier can efficiently classify each of the three weather events. Hence, C4.5 can be extensively utilized for weather event prediction or forecasting. In addition, C4.5 is extremely viable weather event classifier as the C4.5 classifier is comprehensible and interpretable, can deal with the over-fitting issue and may take care of persistent features.

The relative importance scores of weather and climate features in Table III were measured using C4.5 classifier with I_{Gain} ranking filter which investigate the relevance of a feature by measuring the I_{Gain} with respect to the weather event class. To measure the related importance score or rank of the available features, we trained the C4.5 classifier using LA weather history dataset with $5-fold$ cross validation method. It can be pointed out from feature importance measurements that total rainfall was more significant than maximum rainfall features. On the other hand, minimum visibility, and the high temperature had more influence than maximum or mean visibility, and minimum temperature, respectively. Furthermore, mean and minimum dew point features had an almost same

TABLE I
WEATHER EVENTS CLASSIFICATION BENCHMARK. THIS TABLE DISPLAYS THE WEATHER EVENT CLASSIFICATION ACCURACY, GENERATED FROM NAIVE BAYES AND C4.5 ON LA WEATHER HISTORY DATASET

Models	**Cross-fold validation**					**Random splitting**				
	10-fold	20-fold	30-fold	40-fold	50-fold	50%-50%	60%-40%	70%-30%	80%-20%	90%-10%
Naive Bayes	0.9075	0.9069	0.9063	0.9067	0.9062	0.9161	0.9122	0.9075	0.9047	0.9188
C4.5	0.962	0.9626	0.9633	0.963	**0.9637**	0.946	0.949	0.9452	**0.9528**	0.9452

TABLE II
WEATHER EVENTS CLASSIFICATION BENCHMARK. THIS TABLE DISPLAYS THE PRECISION, RECALL, F-SCORE, AND THEIR STANDARD DEVIATON (SD), FOR EACH WEATHER EVENT, GENERATED FROM NAIVE BAYES AND C4.5 ON LA WEATHER HISTORY DATASET. THE DISPLAYED RESULTS POINTS OUT THAT C4.5 CLASSIFIER IS COMPARATIVELY BETTER THAN NAIVE BAYES

Events	**Naive Bayes**						**C4.5**					
	Precision	Precision SD	Recall	Recall SD	F-score	F-score SD	Precision	Precision SD	Recall	Recall SD	F-score	F-score SD
Normal	**0.978**	0.00071	0.918	0.00141	0.947	0.00005	0.971	0.00003	**0.987**	0.00071	**0.979**	0.00002
Rain	0.733	0.00283	**0.824**	0.00283	0.776	0.00283	**0.874**	0.00212	0.809	0.00636	**0.84**	0.000283
Fog	0.345	0.00495	**0.906**	0.00495	0.499	0.00424	**0.98**	0.00002	0.742	0.00012	**0.845**	0.00005

level of importance, while high dew point feature comprised less importance than other two dew point features. Sea level features had less discriminative power towards classifying weather events in which minimum sea level feature contained more relevance than the maximum and mean sea level features.

TABLE III
THIS TABLE DISPLAYS AVERAGE MERIT AND RANK OF EACH FEATURE OF LA WEATHER HISTORY DATASET, MEASURED VIA I_{GAIN} USING 5-FOLD CROSS VALIDATION STRATEGY

Feature	**Average merit**	**Average rank**
Total rainfall	0.39 ± 0.005	1 ± 0
Minimum visibility (Min vis)	0.226 ± 0.002	2 ± 0
Maximum temperature (Max temp)	0.139 ± 0.005	3 ± 0
Maximum humidity (Max hum)	0.089 ± 0.005	4.6 ± 0.8
Average visibility (Avg vis)	0.089 ± 0.001	4.6 ± 0.49
Average temperature (Avg temp)	0.084 ± 0.003	5.8 ± 0.4
Average humidity (Avg hum)	0.066 ± 0.003	7.4 ± 0.49
Average wind speed (Avg w speed)	0.064 ± 0.001	7.8 ± 0.75
Maximum rainfall (Max rainfall)	0.062 ± 0.003	8.8 ± 0.4
Minimum humidity (Min hum)	0.052 ± 0.005	10.8 ± 0.98
Average dew point (Avg dew)	0.05 ± 0.002	10.8 ± 0.75
Minimum dew point (Min dew)	0.048 ± 0.001	11.4 ± 0.49
Maximum wind speed (Max w speed)	0.044 ± 0.002	13.2 ± 0.4
Minimum temperature (Min temp)	0.044 ± 0.002	13.8 ± 0.4
Maximum dew point (Max dew)	0.034 ± 0.002	15 ± 0
Maximum visibility (Max vis)	0.028 ± 0.002	16 ± 0
Minimum sea pressure (Min sea pr)	0.019 ± 0.001	17.2 ± 0.4
Maximum sea pressure (Max sea pr)	0.018 ± 0.002	18.4 ± 0.49
Average sea pressure (Avg sea pr)	0.018 ± 0.001	18.4 ± 0.8

Figure 2 highlights the correlation among the weather and climate features analyzed in this study. The legend or the scale of figure 2 denotes the correlation scale, while X-axis and Y-axis represent all available features. Recall that when the value of correlation between two features is 1, it indicates that there is a strong correlation between the two features, when the value of correlation between two features is -1, it indicates that there is a very weak correlation between the two features, and finally, when the correlation value between two features is 0 that indicates that there is no correlation between the two features. In our experimented feature set, there was strong correlation between air temperature and dew point, dew point and humidity, humidity, and temperature, temperature, and visibility, humidity and rainfall features. There is a significant correlation between wind speed and rainfall, visibility and humidity, humidity and wind speed, air temperature and rainfall features.

VI. CONCLUSION AND FUTURE WORK

This paper we introduce C4.5 classifier in order to tackle the challenge of weather events classification such as normal, rain, and fog using weather and climate information. Our findings suggest that C4.5 classifier is able to predict different weather events with a very small error rate using a minimal number of weather and climate features. We also outline relevant and influential weather event features e.g. rainfall, visibility, temperature, wind speed, dew point computing their relative importance scores. In addition, feature correlation plot demonstrates that air temperature and dew point, dew point and humidity, humidity, and temperature, temperature, and visibility, humidity and rainfall are highly correlated. In our future work, we will include an extension of the weather event class to more complex events such as rain-fog, thunderstorm, rain-thunderstorm, tornado, rain-tornado, rain-thunderstorm-tornado, and fog-rain-thunderstorm. Next, the comparison of the proposed model with many other state-of-the-art techniques, and the validation of the proposed event

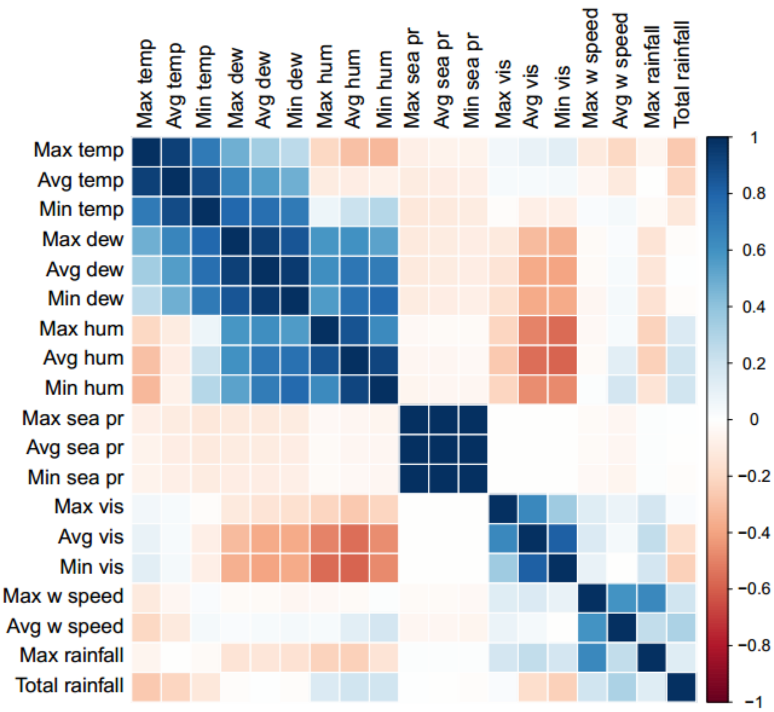

Fig. 2. Feature correlation plot of the investigated features on the LA weather history dataset

prediction model on more complex and unbalanced weather dataset with different climate conditions.

REFERENCES

[1] F. Olaiya and A. B. Adeyemo, "Application of data mining techniques in weather prediction and climate change studies," *International Journal of Information Engineering and Electronic Business (IJIEEB)*, vol. 4.1, p. 51, 2012.

[2] G. M. S.-Z. R. G.-G. De la Torre-Gea, Guillermo and E. Rico-Garca., "Bayesian networks for defining relationships among climate factors," *Int. J. Phys. Sci.*, vol. 6, no. 1, pp. 4412–4418.

[3] C. H. Lima and U. Lall, "Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow," *Journal of Hydrology*, vol. 383, no. 3, pp. 307–318, 2010.

[4] T.-w. Kim, H. Ahn, G. Chung, and C. Yoo, "Stochastic multi-site generation of daily rainfall occurrence in south florida," *Stochastic Environmental Research and Risk Assessment*, vol. 22, no. 6, pp. 705–717, 2008.

[5] J. Wu, L. Huang, and X. Pan, "A novel bayesian additive regression trees ensemble model based on linear regression and nonlinear regression for torrential rain forecasting," in *Computational Science and Optimization (CSO), 2010 Third International Joint Conference on*, vol. 2. IEEE, 2010, pp. 466–470.

[6] A. Bautu and E. Bautu, "Meteorological data analysis and prediction by means of genetic programming," in *Proceedings of the 5th Workshop on Mathematical Modeling of Environmental and Life Sciences Problems Constanta, Romania*, 2006, pp. 35–42.

[7] J. Wu and E. Chen, "A novel nonparametric regression ensemble for rainfall forecasting using particle swarm optimization technique coupled with artificial neural network," in *Advances in Neural Networks–ISNN 2009*. Springer, 2009, pp. 49–58.

[8] G.-F. Lin and L.-H. Chen, "Application of an artificial neural network to typhoon rainfall forecasting," *Hydrological Processes*, vol. 19, no. 9, pp. 1825–1837, 2005.

[9] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[10] "La weather history dataset," Last Accessed: 23 January 2016. [Online]. Available: http://www.wunderground.com/history/airport/KCQT/2015/1/1/CustomHistory.html

[11] R. F. Chevalier, G. Hoogenboom, R. W. McClendon, and J. A. Paz, "Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks," *Neural Computing and Applications*, vol. 20, no. 1, pp. 151–159, 2011.

[12] C. J. Devi, B. S. P. Reddy, K. V. Kumar, B. M. Reddy, and N. R. Nayak, "Ann approach for weather prediction using back propagation," *International Journal of Engineering Trends and Technology*, vol. 3, no. 1, 2012.

[13] F. Olaiya and A. B. Adeyemo, "Application of data mining techniques in weather prediction and climate change studies," *International Journal of Information Engineering and Electronic Business (IJIEEB)*, vol. 4, no. 1, p. 51, 2012.

[14] K. Mohammadi, S. Shamshirband, S. Motamedi, D. Petković, R. Hashim, and M. Gocic, "Extreme learning machine based prediction of daily dew point temperature," *Computers and Electronics in Agriculture*, vol. 117, pp. 214–225, 2015.

[15] M. S. K. Awan and M. M. Awais, "Predicting weather events using fuzzy rule based system," *Applied Soft Computing*, vol. 11, no. 1, pp. 56–63, 2011.

[16] G. De la Torre-Gea, G. M. Soto-Zarazúa, R. Guevara-González, and E. Rico-García, "Bayesian networks for defining relationships among climate factors," *Int. J. Phys. Sci*, vol. 6, pp. 4412–4418, 2011.

[17] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Dec. 2007.

[18] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2007, pp. 3–24.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278