Forecasting meteorological analysis using machine learning algorithms.

PAVULURI, B.L., VEJENDLA, R.S., JITHENDRA, P., DEEPIKA, T. and BANO, S.

2020

Forecasting Meteorological Analysis using Machine Learning Algorithms

Bhagya Lakshmi Pavuluri
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
bhagyapavuluri66@gmail.com

Ramya Sree Vejendla
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
ramyavejendla123@gmail.com

Pavuluri Jithendra
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
jithendrapavuluri43@gmail.com

Tinnavalli Deepika
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
deeputinnavalli1999@gmail.com

Shahana Bano
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
shahanabano@icloud.com

***Abstract*— Weather prediction is gaining up ubiquity quickly in the current period of Machine learning and Technologies. It is fundamental to foresee the temperature of the climate for quite a while. Decision trees, K-NN, Random Forest algorithms are an integral asset which has been utilized in several prediction works for instance, flood prediction, storm detection etc. In this paper, a simple approach for weather prediction of future years by utilizing the past data analysis is proposed by the decision tree, K-NN and random forest algorithm calculations and showing the best accuracy result of these three algorithms. Weather prediction plays a significant job in everyday applications and in this paper the prediction is done based on the temperature changes of the certain area. All these algorithms calculate the mean values, median, confidence values, probability and show the difference between plots of all the three algorithms etc. Finally, using these algorithms in this work we can predict whether the temperature increases or decreases, is it a rainy day or not. The dataset is completely based on the weather of certain area including few objects like year, month, and temperature, predicted values and so on.**

Keywords*— *Decision Tree Algorithm, Random Forest Algorithm, K-NN Algorithm, Classification, Weather Prediction, Hot, Cold, Rain.

I. INTRODUCTION

Utilizing the right algorithm for the future predictions is viral these days. That's the reason we done work on weather forecast too. Machine learning algorithms provide accurate results for predicting climate such as outlook, humidity, rainfall, temperature, floods and storms. This part is colossally reliant on past information and man-made consciousness. Foreseeing the future climate additionally causes us to settle on choices in crops, international sports as well as numerous parts of human lives. The main reason of taking three algorithms is to test which algorithm performs well regarding the weather forecast and after observing many reference papers we observed that they performed mostly using one or two algorithms with few weather objects that's reason we are motivated to analyze three machine learning algorithms on one purpose by adding more variable to weather dataset. So, we decided to anticipate the normal temperature of the day whether it is hot, cold or rainy by using three following algorithms:

A. Decision Tress Algorithm:

Decision trees models are regularly utilized to examine the dataset and actuate the tree and its process is used for predictions. To build the decision tree there are few various algorithm such as Classification and Regression Trees (CART), C5.0 and ID3 etc. But in this work we used classification using ID3 for weather prediction. Generally, the tree contains branch nodes and each node significates the decision between various other options, and final leaf node signifies the decision. According to the algorithm tree produces two or more branches for each node and when it produces two branches it is known as binary tree or multiway tree for more branches. In this paper the classification tree predicts the weather values for meteorological information of all months in 2016 and 2017.

B. Random Forest Algorithm:

Random Forest method is the most popularly used algorithm in many fields like medical, power industries, identifying climate changes and weather predictions etc. It is the best algorithm to control the high- dimensional data for both classification and regression. The combination of many individual decision trees as training data and testing sample values is the random forest. Which gives the average single weather prediction by the output of all individual trees. This bagging process helps in improving stability, solving overfitting problem of the large dataset, decreasing the variance and helps in giving best accuracy results.

C. K-Nearest Neighbours Using Neural Networks Algorithm:

This algorithm is used to identify the nearer values for weather prediction by utilizing the training dataset and it depends on feature similarity for the prediction values of new data. So, first the K value is initialized and then the distance is calculated for all the point through Euclidian distance formula (dist((x,y),(a,b)) = square root(x-a)^2 + (y-b)^2) in between the input and training tests. On repeating this process several times, we get the collection of nearest neighbors of the given weather dataset which will be utilized to predict the

weather. Finally, for the best representation of K-NN prediction we add the neural network to algorithm so that the initial point can be easily identified and remaining as hidden point in the hidden layers leading to the output points as prediction points.

II. LITERATURE SURVEY

The review of previous studies regarding weather prediction are explained with the motivation of the proposed work.

Paper [1]: Analysis of weather prediction using Machine Learning & Big Data.
Techniques: Linear Regression and Support Vector Machine.
Performance Analysis: This paper illustrated, how to predict the weather of next 5 days using linear regression and SVM machine learning algorithms. In the end results are measured and confusion matrix for accurate prediction is given using Big Data.
Paper [2]: Survey on Weather Forecasting Using Data Mining.
Technique: ANN, SVM, Naïve Bayes, Decision Tree classification algorithms.
Performance Analysis: The purpose of this paper is to do survey the various methods and algorithms used for weather prediction in data mining field.
Paper [3]: Weather Forecasting Using Machine Learning Algorithm.
Techniques: Random Forest Classification algorithm, Raspberry Pi 3 B model and Python language.
Performance Analysis: To forecast the weather a system is prepared using Raspberry Pi and python. This project is to develop a less cost and efficient weather prediction application using machine learning.
Paper [4]: Weather Forecasting Using Artificial Neural Network.
Techniques: ANN, LSTM, Recurrent Neural Network.
Performance Analysis: In this paper trained the neural network with weather parameters and utilized the LSTM algorithm to gather weather information. After testing the data the weather is predicted through the developed model.
Paper [5]: Rainfall Prediction based on Deep Neural Network: A Review.
Techniques: Deep Neural Network model with optimization.
Performance Analysis: After completion of data pre-processing and feature extraction the model parameter optimization was done to compare the machine learning algorithms performance. The Adam optimizer is used in optimization and showed result as deep neural network works well than machine learning algorithms for weather prediction.

III. PROPOSED WORK

We used weather dataset which is taken from the Kaggle website because this website provides many various datasets and is popular source for datasets. In the weather dataset there are 19 variables and 3655 objects for prediction. Some features of the dataset are:

1. Minimum Frequency
2. 1st Quadratic
3. Median
4. Mean
5. 3rd Quadratic
6. Maximum Frequency
7. Centroid
8. Events

The target value of this dataset is events which contain values like hot, cold and rain without missing values in between the data. So, the dataset consists of 1084 cold records, 1388 hot records and 1183 rain records for predicting the weather.

A. Decision tree algorithm process:

Decision tree algorithm is a tree which represents nodes as variables, branches as decision rule and leaf nodes as outcomes of the dataset. There are different algorithms used to produce decision tree from data are: Classification and regression tree CART, ID3, CHAID, ID 4.5.As we used CART algorithm for the dataset it uses Gini index to represent metric. According to the weather data set the target function is to predict the events (hot, cold or rain) based on the weather values. From the data, maxtemp, mintemp, maxcold etc are the variables of the data. First to build the tree Gini index is calculated for all the features of dataset.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Eq. (1)

P_i is the proportion of samples that related to class 'c' for a particular node.
Gini process value of the dataset is 0.681025.

We can also use the entropy formula for determining the nodes of decision tree.

$$Entropy = \sum_{i=1}^{c} - p_i * log_2(p_i)$$

Eq. (2)

This is entropy is for all non-empty classes where p≠0 and if the samples at a node belong to the same class the entropy values '0'.
Entropy process value of the dataset is 1.72352 and it utilizes the probability of a particular output to make decision on how the nodes should be branched. Furthermore, it is a bit dif

ferent from Gini index because it is having more mathematic al intensive as there is logarithmic function is used in its calc ulation. After calculating all the features of the data we get t he final outcome of the weather data as decision tree and we add the prediction function to predict the test values.

B. Random Forest Algorithm:

Random Forest algorithm is the group of different decision trees and it combines the decisions of all the decision trees to figure out the prediction value, which signifies the single average of all the decision trees. This algorithm solves the overfitting problem of the data and is fast to train the data with test data. The bootstrap samples were produced as individual decision tree. So, we apply the decision tree algorithm process and add the random forest function to the dataset as ("output.forest") for results. We took 500 decision trees which are processed individually to classify the prediction (single average prediction). The following are random forest confusion values of the weather dataset:

Table 1. Confusion matrix values

objects	cold	hot	rain	class.err -or
cold	434	354	73	0.4976852
hot	198	862	252	0.3439878
rain	57	204	1073	0.2139194

C. K-Nearest Neighbour Algorithm:

This algorithm comes under the supervised learni ng algorithm it contains the dataset with training label measu rements(x, y) where x represents the features of dataset and y represents the target of the dataset (target= events). Acc-o rding to the classification problem, the K-NN said that for the given values the algorithm will identify the K nearest ne-ighbor of not seen data points and add the class to those data points to know the K neighbors. So, in order to find the dis-tance metrics of the points we utilize the following formula: Euclidean metric

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + \cdots + (x_n - x_n')^2}$$

Eq. (3)

'n' is the number of dimensions, x and x' are sample points, this is used to identify the k closet points.
Finally, the input value x is assigned to class including the largest probability.

$$P(y = j|X = x) = \frac{1}{k} \sum_{i \in A} I(y^{(i)} = j)$$

Eq. (4)

Conditional expectations of X and y points, 'k' represents th e samples.
Different variables will have various scaling unites so we do normalization for each variable of the data using **(x-min(x))/(min(x) - max(x))** to convert all into values between 0 and 1 and is easy to plot. This algorithm will perform well with the numeric variables because it is dealing with the distance.
Now, the dataset is splinted into training as well as testing sets and the k-nn function is applied for the target category to predict the weather. Finally, the accuracy of prediction is done to divide the right predictions of all the remaining predictions.
→ accuracy (tab)
[1] 61.20219

IV. FLOW CHART

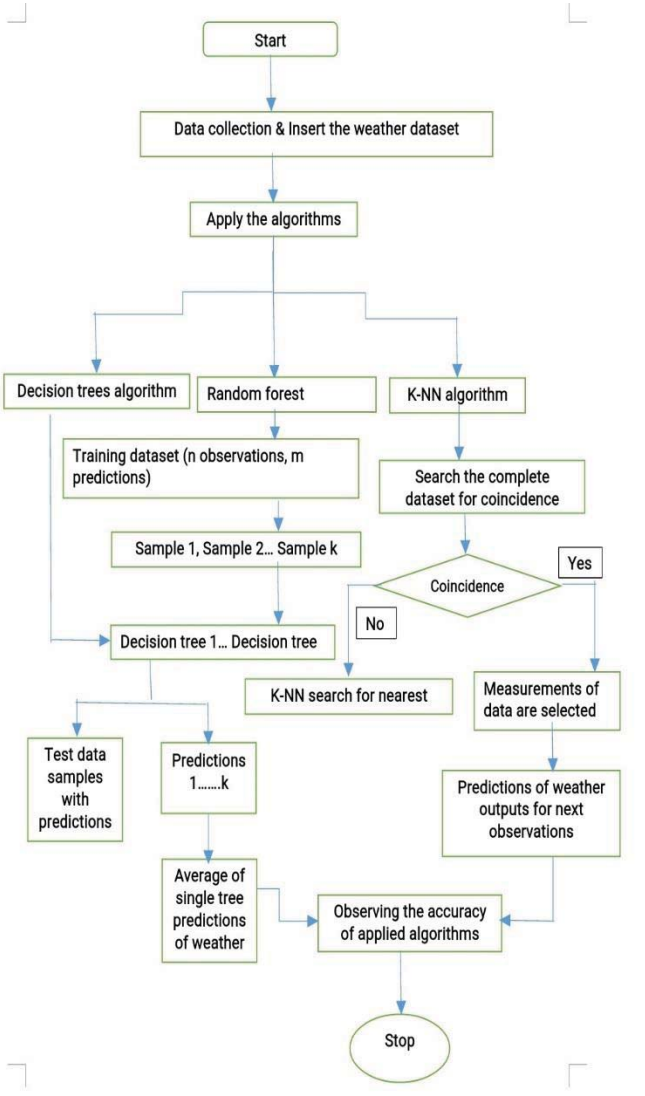

Fig.1. Overview of the process

This work starts by collecting the weather data of an area and then processing that data for data pre-processing which includes few steps such as required libraries are imported after importing the weather dataset thereafter have to clear the missing data, noisy data through few techniques and splitting the dataset as training and testing data for scaling the result. Now, apply the three algorithms to the same weather dataset individually to know the best accuracy of the weather prediction.

As the decision tree algorithm is applied for the data it divides the data in to various classes depending on the certain basis of the given data temperatures and produces the final weather prediction result as hot, cold, and rainy. Furthermore, the random forest algorithm is applied for the same weather dataset and it can handle the large dataset very easily because it is the combination of many decision trees. This algorithm utilizes bagging as well as feature randomness while constructing the individual trees for an uncorrelated forest whose prediction will be more accurate when compared to the individual trees and after observing all the produced bootstrap samples it calculates the average single tree prediction for providing the weather prediction. In addition, to observe the accuracy of other algorithm we used k- nearest neighbor with neural networks which calculates the distance of all the points of data from the initial k-point for finding any coincidence in between the points and after observing or calculating all the points from input layer to hidden layers it provides the weather prediction result as output layer.

V. RESULTS

A. DECISION TREE ALGORITHM:

Initially, here we considered a weather dataset and the dataset is read and pre-processed. Now select a sample of data and this part of dataset is trained then the data apart from the sample is to test the algorithm here number of dimensions of both test data and train data are taken into account. Now apply r-part to terminate the growth of a decision tree and limit its depth. Now here plot the decision tree constraints in tree structure with various nodes and leaf nodes classifying various conditions with asset of queries to differentiate the data and classify it. Also to evaluate certain constraints such as temperature, humidity or precipitation values. Here to obtain and analyse the mechanism the temperature values are predicted and the decision tree is obtained for all the subset of temperature values such as hot, cold and rain. And the dataset discriminates the data based on predicted temperature variables and construct the tree.

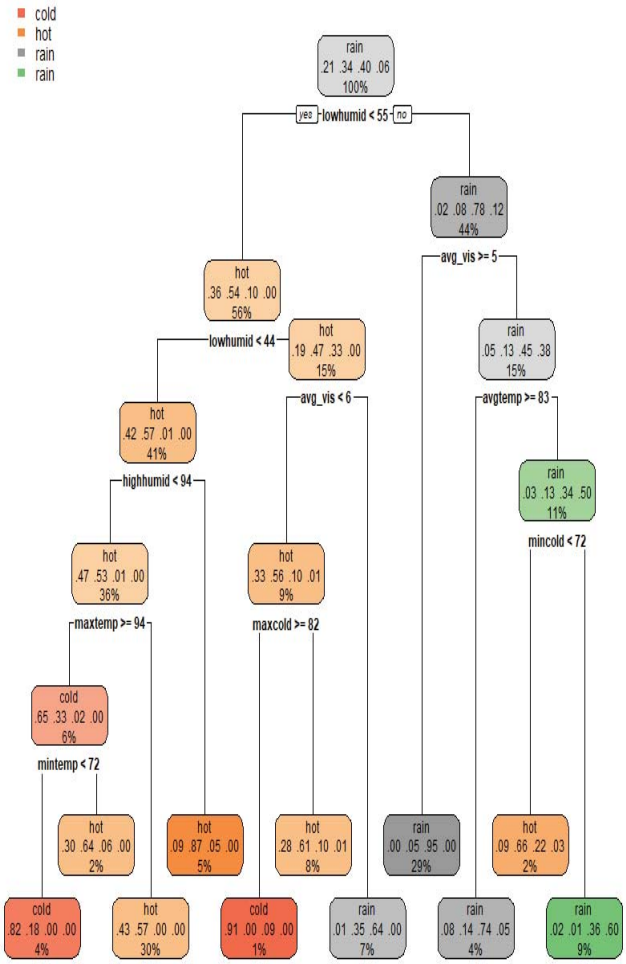

Fig.2. Decision Tree Result

B. RANDOM FOREST ALGORITHM:

The above result explains the random forest output which is showing the prediction results individually depending on the given weather dataset. The y-axis represents the error values and x-axis represents the n-tree values where the n-tree value is given 500 trees as default. Prediction values are provided as three categories where hot is signified with green colour in the graph, cold as black and rain as red colour depending on the attributes used in the dataset. These results are given with the error values are compared with the accuracy because if the error value decreases the accuracy value will increase and vice versa. According to the graph we can say that error value is high for hot result and almost same for both the rain and cold values but as the bootstrap samples of trees increases to 500 we can observe that the error value is reduced closer to zero. Coming to the process to produce this result first the required libraries are installed such as “random forest”, party, mice, VIM, lattice, caret etc. After completion of data reading and pre-processing, the data modelling is done by applying “output.forest” function as “rf” which indicates the random

forest function for the attributes Maxtemp, Mintemp, Maxcold etc., in the dataset.

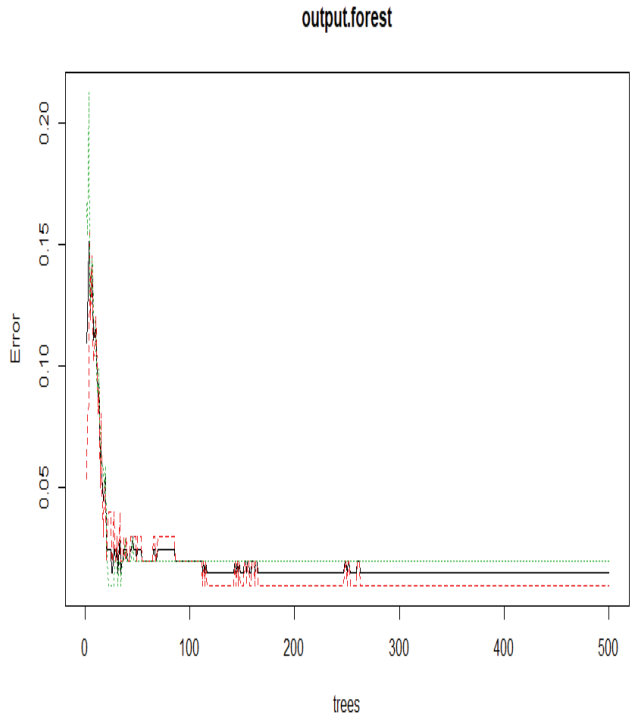

Fig.3. Random Forest Result.

C. K-NN USING NEURAL NETWORKS ALGORITHM:

This neural network plot shows the prediction result produced by the k-nn algorithm for 150 objects and few attributes like Maxtemp, Mintemp and so on including the one hidden layer and output layer. The k-nn calculates the distance in between all the points through Euclidian distance formula and show it as neural network so while installing the libraries neural net is the important one among other libraries and to give more detailed data few functions are used such as head, tail, length, str and its summary. Thereafter, the sample size is set for n-rows of the data and divided the data as training, testing for applying "nn" function which indicates the neural net function. The hidden layers can be mentioned in the neural net function and to show the clear result, here we used only single hidden layer and produced the predicted result as well as we can see that the error is 0.03 on using this algorithm. The fig.4 represents the overall prediction value of the output while the fig.5 shows all the variable of the dataset and distance calculated values with the hidden layer to provide the output layer.

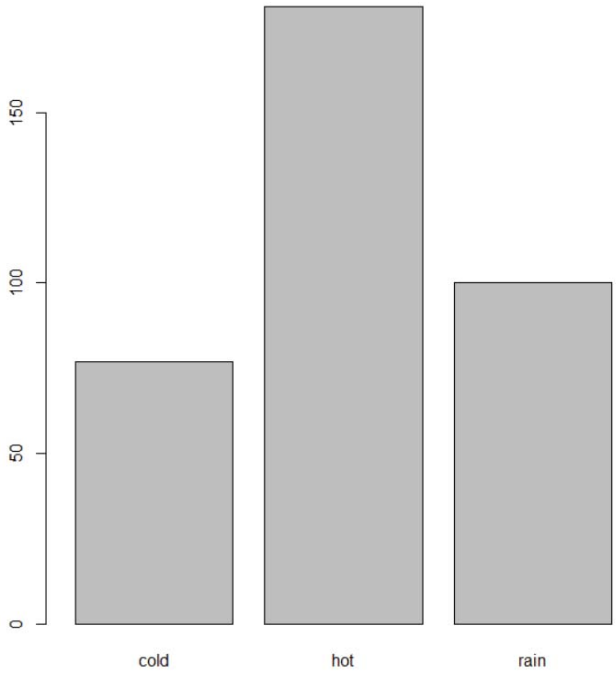

Fig.4. K-NN Plot.

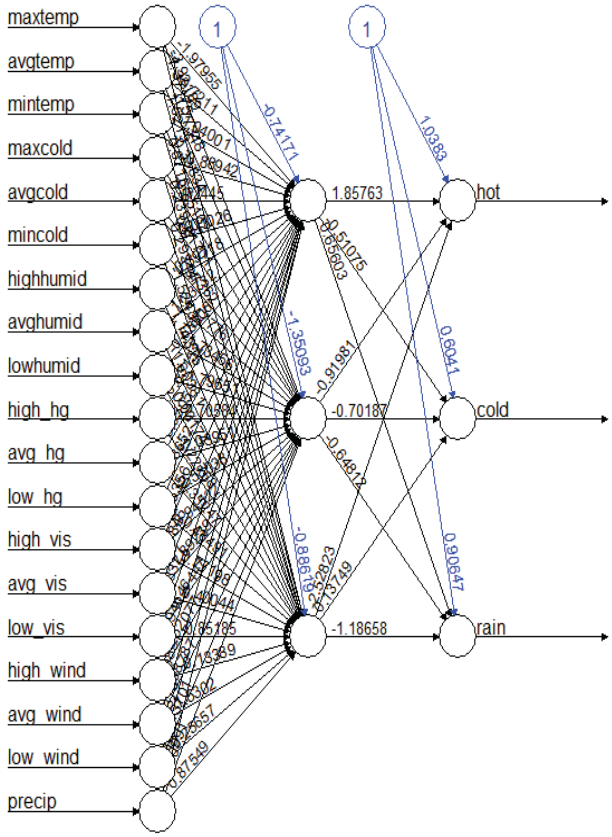

Fig.5. K-NN using Neural Networks Plot.

Table 2. Comparison of results.

SI.NO	Results	Decision Tree	Random Forest	K-NN
1.	Error values	0.48437	0.025	0.035511
2.	Accura-cy	55%	80%	61.20219%
3.	Accura-cy score	0.55	0.80	0.61
4.	Data handlin-g	Handles minimu-m data	Handles large data	Handles minimum data
5.	Benefit-s	Classifie-s the decision easily	Solves overfittin g problem.	Reads all data and finds the best probabili ty results out of it

Table 3. Decision Tree Values of data-test for split [, 5].

Predict- p	hot	cold	rain
hot	0	0	3
cold	7	7	1
rain	0	0	17

VI. CONCLUSION

To conclude, our paper has performed the weather prediction using machine learning algorithms to classify whether it is a hot, cold or rainy day. We utilized various machine learning algorithms such as Decision Tree, Random Forest and K-NN with Neural Networks for the prediction and compared the prediction accuracy of those algorithms individually for the same weather dataset with few attributes like maximum temperature, minimum temperature, and maximum cold and so on. The main reason of taking three algorithms is to test which algorithm performs well regarding the weather forecast. From the above results we can say that Random Forest algorithm provides the best accuracy out of them.

All these algorithms comes under supervised learning because of classification process and work well with the training data than the testing data but may not predict the other categories like temperature percentage, outlook and rainfall amount. So, we cannot suggest this for upcoming years or for any weather changes of various areas and that is why we are planning to forecast other different weather condition by adding more attributes with other algorithms like SVM, Naïve Bayes and ANN.

VII. REFERENCES

[1] Analysis of Weather Prediction Using Machine Learning & Big Data 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE- 2018) Paris, France 22-23 June 2018, 978-1-5386-4485-0/18.

[2] Survey on Weather Forecasting Using Data Mining, Proc. IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2018) 2-3 March 2018, 978-1-5386-3479-0/18.

[3] Weather Forecasting Using Machine Learning Algorithm, 978-1-5386-9436-7/19, 7-9 March 2019.

[4] Weather Forecasting Using Artificial Neural Network, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE, 978-1-5386-1974-2/18.

[5] Rainfall Prediction based on Deep Neural Network: A Review, Proceedings of the Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020) IEEE.

[6] Rainfall Forecasting in Bandung Regency using C4.5 Algorithm, 2018 6th International Conference on Information and Communication Technology.

[7] Haze weather recognition based on multiple features and Random forest, 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC).

[8] Dynamic Line Rating Using Numerical Weather Prediction and machine learning, 2017, IEEE.

[9] Comparative Analysis of Temperature Prediction using Regression Methods and Back Propagation Neural Networks Survey on Weather Forecasting Using Data Mining, Proc. IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2018) 2-3 March 2018, 978-1-5386-3479-0/18.

[10] Weather Analysis to predict rice cultivation time using multiple linear regression to escalate farmers exchange rate

[11] 2017 International Conference on Advanced Informatics, Concepts, Theory and Applications. Weather prediction based on fuzzy logic algorithm for supporting general farming automation system, 2017 5th International Conference on Instrumentation, Control, and Automation (ICA).

[12] A Hadoop based weather prediction model for classification of weather data, 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT).

[13] A Quick Review of Machine Learning Algorithms, 2019, International Conference on Machine Learning, Big Data, Cloud and Parallel Computing.

[14] Long-time Prediction of Climate-weather Change Influence on Critical Infrastructure Safety and Resilience, 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM).

[15] Study of prediction algorithms for selecting appropriate classifier in machine learning.(Journal of Advanced Research in Dynamical and Control Systems, 9(Special Issue 18), 257-268. 2017)