iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48026

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Prediction and Classification of Weather Using Machine Learning

Yash Mali[1], Aarya Kurlekar[2], Shambhavi Lalsinge[3]
[1, 2, 3]*Vishwakarma Institute of Technology, Pune, India*

Abstract: ***To predict something we need some background study to understand the pattern. On earth, every phase of human life is influenced by nature. As we cannot avoid the natural changes and conditions, we have chosen to minimize their effect on our lives. Therefore, to achieve this we need to know the weather conditions beforehand, to make things work according to the changes in the environment. Here comes the role of prediction. To carry out prediction, accurate classification of data is required. The main objective of this project is to design a weather prediction model. Agriculture is the field that is most influenced by the weather. So, we can extend our scope to provide regional guidelines to farmers depending upon the classification done. This paper explores the details of this project.***
Keywords: Weather prediction, temperature, pressure, humidity, windspeed, machine learning.

I. INTRODUCTION

In recent years, the world has witnessed rapidly changing environmental conditions. Weather conditions affect all the major areas. Environmental change is a great deal of consideration for a long time because of the sudden changes that happen [2] So, weather forecasts are essential. Weather forecasting is the task of predicting the state of the atmosphere at a future time and a specified location [5]. Weather forecasting plays a very vital role in many fields. Weather forecasting plays an important role in meteorology [2] Traditionally, weather prediction was done by physical models of the atmosphere. The present state of the atmosphere is sampled, and the future state is computed by numerically solving the equations of fluid dynamics and thermodynamics. However, the system of ordinary differential equations that govern this physical model is unstable underneath perturbations and uncertainties within the initial measurements of weather [5]. But to make it easier and reliable, it can be done by making use of available technology. Machine learning can come to aid when it comes to weather forecasting or prediction as it is more robust and doesn't need a clear understanding of the physical process of forecasting as rightly stated in [5]. Thus, in this project, we have implemented 2 machine learning algorithms to do the weather prediction. Along with the prediction we have also included the classification of the weather depending on the predicted value. As discussed, weather forecasting plays a vital role in different fields. Agriculture is one of the major fields where weather prediction would be very much beneficial. So, the scope of the following project is expanded for the application of this forecasting in agriculture. Depending on the result of the weather-predicting model, we will be generating some important guidelines for the farmers in that region, so that they will be aware of it and take necessary actions. This paper further discusses which are machine learning algorithms used, how are they implemented, and the results and accuracy of those algorithms.

II. LITERATURE SURVEY

In [1] the paper "Analysis of Weather Prediction using Machine Learning & Big Data" by Shubham Madan, the prediction of weather was done by using big data processing and machine learning. The attributes/feature used for the prediction is maximum temperatures, minimum temperature, mean humidity, mean atmospheric pressure. The algorithms used were linear regression and support vector machines. To check the accuracy for the given project the author has mentioned the 'root mean squared' method.
In [2], the paper "Weather Forecast Prediction: An Integrated Approach for Analyzing and Measuring Weather Data" by Munmun Biswas, Tanni, Sayantanu Barua in the year 2018, the author has proposed a methodology of weather prediction using machine learning algorithms like 'Chi_sqaure' for prediction and 'Naïve-Bayes' for classification. The attributes used in this implementation were, 'outlook, temperature, humidity, wind'. The weather was simply classified into 'Good' or 'Bad'. The author concluded that This methodology could decide the nonlinear relationship that exists between the historical data (temperature, wind speed, humidity, and so forth,) provided to the system during the training phase and on that premise, predict what the weather would be in future.
In [3], the paper "Weather Prediction Based on Fuzzy Logic Algorithm for Supporting General Farming Automation System" published by Aris Pujud Kurniawan, Agung Nugroho Jati, Fairuz Azmi in 5th International Conference on Instrumentation, Control,

©IJRASET: All Rights are Reserved

and Automation (ICA) in the year 2017. They built a weather prediction system using the fuzzy logic algorithm to support general automation in the farming sector. They took the data from the weather service provider and underground. Their system also collects data from a rain sensor and soil moisture sensor and using fuzzy logic it decides whether to water the plants or not. It is mentioned that the whole system has been tested by observing the plants every day and the model if there are any errors or not. They have tested the system 33 times in eighteen days with 100% accuracy results.

In the paper "Automated Weather Event Analysis with Machine Learning" by Nasimul Hasan, Md. Taufeeq Uddin, Nihad Karim Chowdhury [4], the authors have discussed the classification of weather using a machine learning algorithm CD4.5. 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 an information theory- based formula choosing the test with the maximum amount of information from the set of examples. In their experiments, they 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 set.

In the paper "Machine Learning Applied to Weather Forecasting" published by Mark Holmstrom, Dylan Liu, Christopher Vo in the year 2017 [5] they built a weather prediction model using 'linear regression' and 'functional regression' algorithms to predict the different factors affecting the weather. Later, they used '4-fold forward chaining time-series cross-validation' to check the accuracy of the model. After the completion of the model, they compared their model with the 'professional weather forecasting services' and as a result, both linear regression and functional regression were outperformed by 'professional weather forecasting services'. But after a certain period, the discrepancies or error rate in their model decreased significantly.

Above, the related work of our project is discussed. As mentioned above there are various machine learning algorithms implemented on different data sets for prediction of weather as well as for classification along with its accuracy measures. This literature survey helped us to learn which algorithms can be used and which would be most suitable for our project.

III. METHODOLOGY

The methodology used in this project for the prediction of weather and its classification comprises usage of two machine learning algorithms- Linear Regression and Support Vector Machine (SVM) respectively. The block diagram depicts the machine learning approach to the implementation of the project. Fig.1.

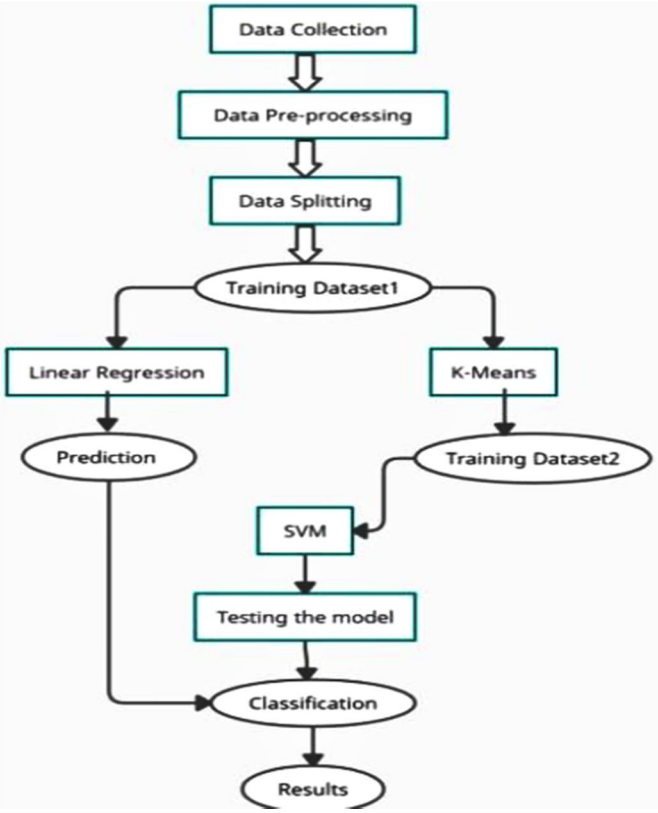


Fig.1. Block diagram of the methodology

Steps followed for implementation of this project:

1) *Data collection:* The data set used for this project was collected from the Kaggle site[ref]. It is the historic weather data of Delhi city in India. This data set has comprised a total of 5 columns i.e., 5 weather attributes and 1421 data tuples. This dataset was processed to get the desired dataset.
2) *Data Preprocessing:* The data pre-processing was done on the raw dataset starting with feature selection. This feature selection was done manually based on the literature survey and research. The selected feature that affects weather the most are mentioned in the table1 along with their unit.

Table 1

Sr, no	Attribute	Unit
1	Temperature	0_C
2	Wind Speed	Km/p
3	Mean humidity	%
4	Mean Pressure	hPa

After the feature selection, data cleaning was done by removing the null values followed by the removal of outliers by using the interquartile range (IQR) method. In the IQR method, the dataset was divided into 4 equal parts and the quartiles are Q1, Q2, Q3.
Q1 represents the 25th percentile of the data. Q2 represents the 50th percentile of the data. Q3 represents the 75th percentile of the data.
Then all the data points below 1.5*Q1 and 1.5*Q3 are the outliers and are dropped from the dataset and the dataset is cleaned.

Date	Temperature	Humidity	Wind Speed	Pressure
01/01/2013	10	84.5	0	1015.666667
02/01/2013	7.4	92	2.98	1017.8
03/01/2013	7.166666667	87	4.633333333	1018.666667
04/01/2013	8.666666667	71.33333333	1.233333333	1017.166667
05/01/2013	6	86.83333333	3.7	1016.5
06/01/2013	7	82.8	1.48	1018
07/01/2013	7	78.6	6.3	1020
08/01/2013	8.857142857	63.71428571	7.142857143	1018.714286
09/01/2013	14	51.25	12.5	1017
10/01/2013	11	62	7.4	1015.666667
11/01/2013	15.71428571	51.28571429	10.57142857	1016.142857
12/01/2013	14	74	13.22857143	1015.571429
13/01/2013	15.83333333	75.16666667	4.633333333	1013.333333
14/01/2013	12.83333333	88.16666667	0.616666667	1015.166667
15/01/2013	14.71428571	71.85714286	0.528571429	1015.857143
16/01/2013	13.83333333	86.66666667	0	1016.666667
17/01/2013	16.5	80.83333333	5.25	1015.833333
18/01/2013	13.83333333	92.16666667	8.95	1014.5
19/01/2013	12.5	76.66666667	5.883333333	1021.666667
20/01/2013	11.28571429	75.28571429	8.471428571	1020.285714
21/01/2013	11.2	77	2.22	1021
22/01/2013	9.5	79.66666667	3.083333333	1021.8

Fig.2. Data set after pre-processing

3) *Linear Regression:* It is a supervised machine learning algorithm, the most basic type of regression. Basically, it is the statistical model that analyzes the linear relationship between a dependent variable with a given set of the independent variable(s). In the project, the simple linear regression was used to predict the individual attribute of the dataset. For this 75% of the dataset was the training dataset i.e., used for training the model and the remaining 25% was used to test the dataset.

Equation of simple linear regression,

$$y = b0 + b1\ x \qquad (1)$$

Y - dependent variable X - independent variable
b1 - slope of the regression line b0 – Y-intercept

4) *K-Means Clustering:* K-means algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group**.** It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. The dataset that we used for this project is unlabeled. That K-means algorithm was used to make the clusters in our dataset. To decide the number of clusters the elbow method was used. In the elbow method, we choose the number from which a linear descend is seen in the elbow-shaped graph plotted. In our case, it is 4i.e., 4 number of clusters is the optimum number of clusters. The graph for the elbow method is shown in Fig3

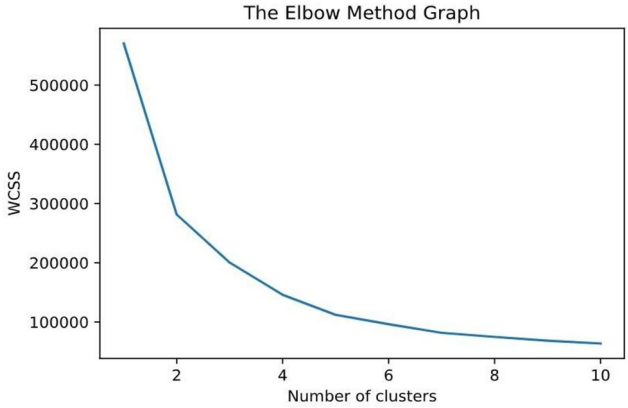


Fig 3. Elbow method for cluster selection

5) *Support Vector Machine:* Support Vector Machine is an algorithm widely used in classification objectives. In SVM hyperplanes are decision boundaries that help classify the data points and we are looking to maximize the margin between the data points and the hyperplane. Once the clustering is done. And in the output, we get the labeled dataset as shown in fig: This data is passed to the SVM algorithm. SVM algorithms use a set of mathematical functions that are defined as the kernel. The kernel function used for the classification of this weather data is 'Polynomial kernel'. The function of a polynomial function is:

$$k(xi, xj) = (xi.\ xj + c)^{d} \qquad (2)$$

It is suitable for this project because of its high accuracy and its efficient application in multiclass classification. multiclass or multinomial classification is the problem of classifying instances into one of three or more classes. The results of this implementation and its analysis are discussed further.

IV. RESULT AND ANALYSIS

After the implementation of the methodology proposed above the results is discussed below along with its analysis and output. The linear regression model successfully predicted the real values of all the attributes with some error. This was tested against the test data which was formed at train-test split. As the regression model used in this project was simple, the method used for checking the accuracy was the R-Squared method. The formula, for R- Squared:

R2=

R-squared (R^2) is a measure that represents the proportion of the variance for a variable that is explained by associate variable quantity or simply variables used in a regression model. In Table III the r-squared values of all the attributes are shown.

R-squared values vary from 0 to 1 and are normally stated as percentages from 0% to 100%. If R-squared is 100% it means that all movements of the dependent variable are completely explained by movements in the independent variable.

Generally, an R2 value above 0.7 is considered as good R2 reading whereas a value above 0.9 shows an excellent accuracy.

The implementation of K-means was done to get different clusters in the dataset. As a result, 4 different clusters were obtained-labeled as 'Clear', 'cold', 'cloudy', 'partly cloudy'. Fig 4, Shows 4 different clusters.

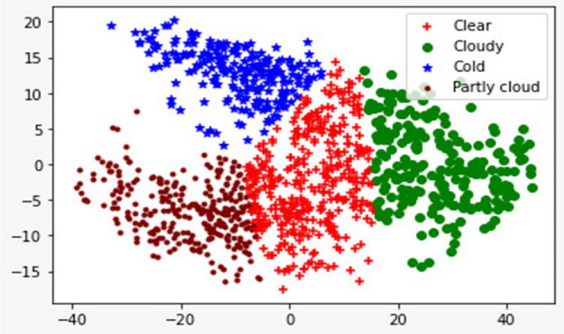


Fig 4. - Different clusters obtained from the dataset

TABLE 2

Attributes	R-squared value
Temperature	0.9488
Humidity	0.7157
Pressure	0.9527

The implementation of SVM was done on the dataset obtained from clustering and again split into Train and test datasets with proportions 80% and 20% respectively. Fig 5 shows the hyperplane between the classes.

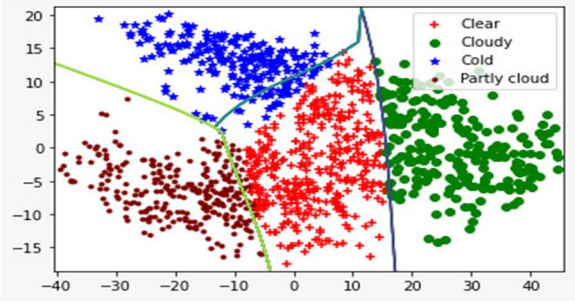


Fig 5. – Classification of classes and hyperplane between them

As all the kernels of SVM were implemented, the results and accuracy for each of them were calculated using the confusion matrix method.

Confusion matrix calculations:
Accuracy = (TP + TN)/ (TP + TN + FP + FN).
Precision is a proportion of the accuracy given that an explicit class has been predicted.
Precision = TP/ (TP + FP)
Recall = Sensitivity = TP / (TP + FN) where,
TP- numbers of true positive FP-numbers of false positive.
TN-numbers of true negatives. FN- numbers of a false negative.
The accuracy of all the kernels is mentioned in TABLE 3:

TABLE 3

Kernel	Accuracy
Polynomial	96%
RBF	18%
Sigmoid	18%

From the table, the polynomial Kernel for SVM is most accurate; the in-detail accuracy metrics and confusion are mentioned in Fig 6.

```
[[49  0  0  6]
 [ 0 53  2  2]
 [ 0  0 75  2]
 [ 0  2  0 94]]
              precision    recall  f1-score   support

       Clear       1.00      0.89      0.94        55
      Cloudy       0.96      0.93      0.95        57
        Cold       0.97      0.97      0.97        77
Partly Cloudy      0.90      0.98      0.94        96

    accuracy                           0.95       285
   macro avg       0.96      0.94      0.95       285
weighted avg       0.95      0.95      0.95       285
```

Fig 6.

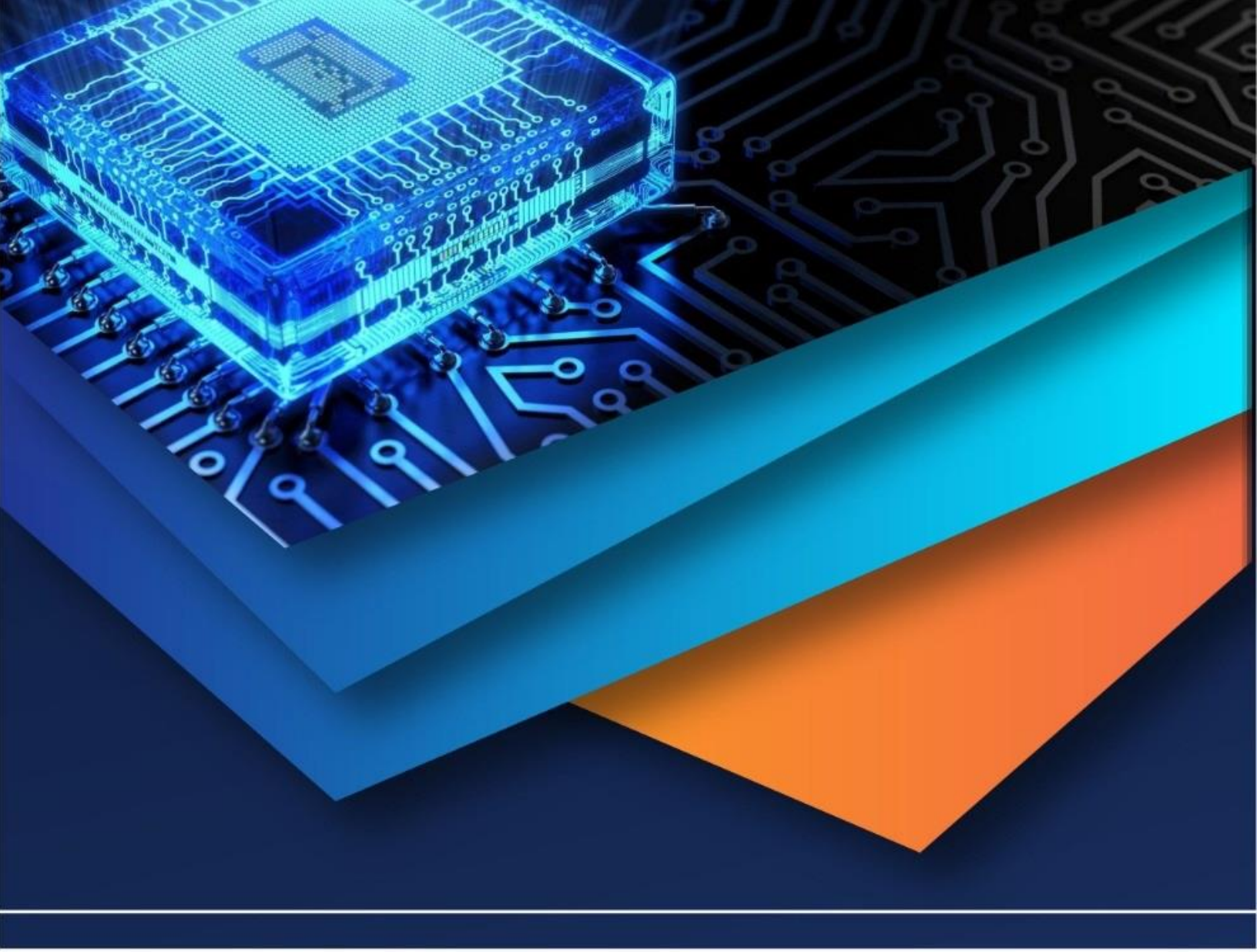
From the above result, the confusion matrix explains the number of true positive, true negative, false positive, and false negatives.

V . CONCLUSION

According to all the results discussed in the paper, we conclude that prediction and classification of weather can be done using machine learning algorithms mentioned in this methodology i.e., Linear regression and SVM. The prediction was done based on weather attributes like temperature, pressure, humidity, and wind speed, and the weather was classified into 4 classes: Cloudy, Partly Cloudy, Sunny, Cold. The linear regression model could perform the prediction with an accuracy of 94%, 95%, 71% for temperature, pressure, and humidity respectively. Also, the classification done based on SVM performs with an accuracy of 96% for the polynomial kernel while for other kernels it performs poorly. Further, this prediction and classification data can be used for the generation of some regional instructions for the farmers for agricultural benefits.

REFERENCES

[1] Shubham Madan, Praveen Kumar, Seema Rawat, Tanupriya Choudhury, “Analysis of Weather Prediction using Machine Learning & Big Data,” International Conference on Advances in Computing and Communication Engineering (ICACCE-2018) Paris, France 22-23 June 2018.
[2] Munmun Biswas, Tanni Dhoom, Sayantanu Barua “Weather Forecast Prediction: An Integrated Approach for Analyzing and Measuring Weather Data” International Journal of Computer Applications (0975– 8887) Volume 182 – No. 34, December 2018.
[3] Aris Pujud Kurniawan, Agung Nugroho Jati, Fairuz Azmi “Weather Prediction Based on Fuzzy Logic Algorithm for Supporting General Farming Automation System,” International Conference on Instrumentation, Control, and Automation (ICA) Yogyakarta, Indonesia, August 9-11, 2017.
[4] Nasimul Hasan, Md. Taufeeq Uddin, Nihad Karim Chowdhury “Automated Weather Event Analysis with Machine Learning,”.
[5] Mark Holmstrom, Dylan Liu, Christopher Vo, “Machine Learning Applied to Weather Forecasting” Stanford University (Dated: December 15, 2016).
[6] 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.
[7] Mr. Sunil Navadia, Mr. Jobin Thomas, Mr. Pintukumar Yadav, Ms. Shakila Shaikh, "Weather Prediction: A novel approach for measuring and analyzing weather data", International conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud), (I-SMAC 2017), IEEE, pp 414-417
[8] Imran Maqsood, Muhammad Riaz Khan, and Ajith Abraham, “An ensemble of neural network for weather forecasting”, Neural Comput & Applic (2004) 13: 112–122
[9] Youguo Li, Haiyan WU “A Clustering Method Based on K-Means”



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)