e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 7, July 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707039 |

A Machine Learning Rainfall Prediction System

Mrs.P.Vanitha, Ms.P.Mutharasi

Assistant Professor, Department of Computer Applications (UG), Hindhusthan College of Arts & Science, Coimbatore, Tamil Nadu, India

III BCA Student, Department of Computer Applications (UG), Hindhusthan College of Arts & Science, Coimbatore, Tamil Nadu, India

ABSTRACT: Predicting the amount of rain will assist prevent flooding, saving lives and property. Additionally, it aids in the management of water resources. Because heavy rainfall is strongly related to both the economy and human lives, forecasting it presents a significant challenge for the meteorological service. It is the cause of yearly natural disasters like drought and flood that affect people all over the world. Forecasting rainfall accurately is crucial for nations like India, whose economies primarily rely on agriculture. Because the atmosphere is dynamic, statistical methods are not very accurate in predicting rainfall. The primary goal of this project is to use machine learning to predict rainfall. We have utilized many data features for analysis in this project.

I. INTRODUCTION

In the current scenario, rainfall is a significant factor for most essential things happening throughout the world. The farming sector is regarded as one of the most critical factors determining the country's economy, and farming relies entirely on rainfall. This research uses machine learning techniques for rainfall prediction and conducts the comparative analysis of two machine learning techniques, respectively, depicting an efficient rainfall prediction method.

Rainfall prediction facilitates water resources management, flood alerts, flight operations management, limiting transportation, construction activities, and other factors that are most important to humankind. Rainfall data for forecasting is collected using weather satellites, wired and wireless instruments, and high-speed computers are used.

Rainfall prediction has been a fascinating and captivating sector since the dawn of civilization, and it remains one of the most complex and enticing domains. Scientists use various methods and techniques to predict rainfall, some of which are more precise than others. Weather forecasting gathers atmospheric conditions such as humidity, temperature, pressure, rainfall, wind direction &speed, evaporation, etc.

Presently, Rainfall prediction is the most crucial factor for most water storage schemes worldwide. The uncertainty of rainfall data is one of the most complex problems. Today, most rainfall forecasting methods are incapable of detecting hidden patterns or non-linear trends in rainfall data.

This research would help discover all hidden patterns and non-linear trends, which would be necessary for predicting accurate rainfall. Due to the presence of complex issues in existing methods that cannot find the hidden patterns and non-linear trends efficiently the majority of the time, the forecast predictions were incorrect, resulting in massive losses.

Thus, this research aims to find a rainfall prediction system that can solve all issues, find complexity and hidden patterns present, and provide proper and reliable predictions, therefore assisting the country in developing agriculture and the economy.

1.1 ABOUT THE PROJECT

Rainfall is considered the primary source of most of the economy of our country. Rainfall prediction is important as heavy rainfall can lead to many disasters. The prediction helps people to take preventive measures and moreover the prediction should be accurate. Prediction mostly short term prediction can gives us the accurate result. The main challenge is to build a model for long term rainfall prediction. The rainfall parameters in this study are collected, trained and tested to achieve the sustainable results through Random Forest regression. Rainfall prediction model mainly based Machine Learning.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707039 |

Henceforth, to find the best way to predict rainfall, study of both machine learning and neural networks is performed and the algorithm which gives more accuracy is further used in prediction. Proposed system presents a set of experiments which involve the use of prevalent machine learning techniques to build models to predict whether it is going to rain tomorrow or not based on weather data for that particular day. This comparative study is conducted concentrating on three aspects: modelling inputs, modelling methods, and pre-processing techniques.

II. SYSTEM ANALYSIS

2.1 Existing System:

When it comes to machine learning, LASSO regression and Neuro-Fuzzy is being used. One of the biggest challenges is the complexity present in rainfall data. Most of the rainfall prediction system, nowadays are unable to find the hidden layers or any non-linear patterns present in the system.

2.1.1 Disadvantages

- The system discussed in this particular study will operate with Matlab software
- The locations for the data processing used in this study are geographically different and distanced that could also impact the correlation efficient that will measure the performance
- Due to presence of the system which doesn't find the hidden layers and nonlinear patterns accurately, the prediction results to be wrong for most of the times and that may lead to huge losses

2.2 Proposed System

The point is to structure and actualize a programmed rainfall prediction determination framework utilizing Python Flask Framework. Every framework has two principle modules, specifically, preparing and testing, where 80% and 20% of the Dataset informational collection were arbitrarily chosen for preparing and testing purposes separately. Every framework likewise has an extra module known as case-based module, where the client needs to information esteems for required characteristics as indicated by the dataset informational index. To implement rainfall prediction we used machine learning algorithms Random Forest Regressor, and Flask API. Python pickling is used to save the model behaviour and python un pickling is used to load the pickle file whenever required.

Advantages

- It is a powerful technique for testing relationship between one dependent variable and many independent variables
- system predicts rainfall for the approach which is more accurate
- One of the biggest advantages of random forest is its versatility.
- It can be used for both regression and classification tasks, and it's also easy to view the relative importance it assigns to the input features.

2.3 Machine Learning Algorithm

Flask

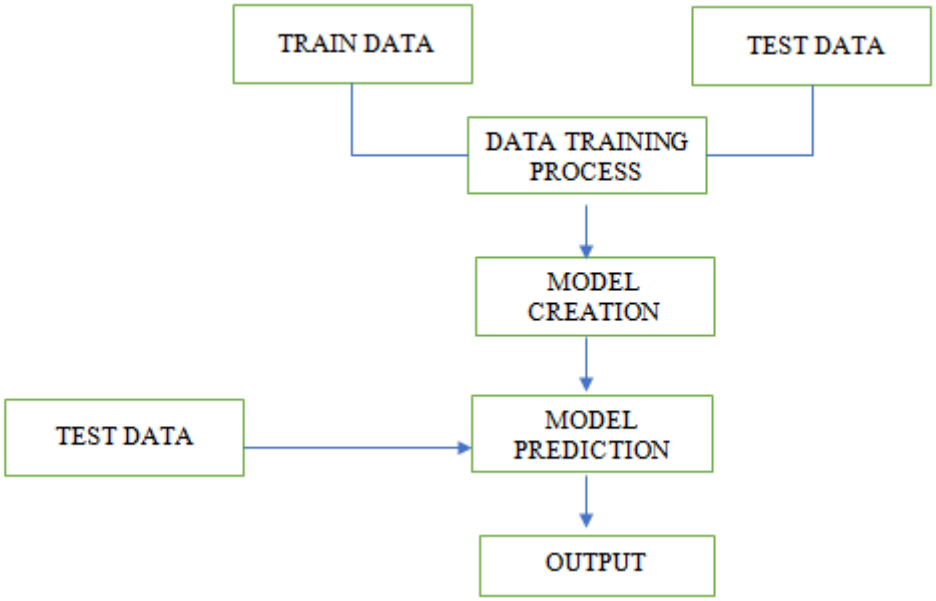
Flask is a Python-based micro framework used for developing small scale websites. Flask is very easy to make Restful API's using python. As of now, we have develop a model i.e model. pkl which can predict a class of the data based on a various attribute of the data.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707039 |

III. SYSTEM IMPLEMENTATION



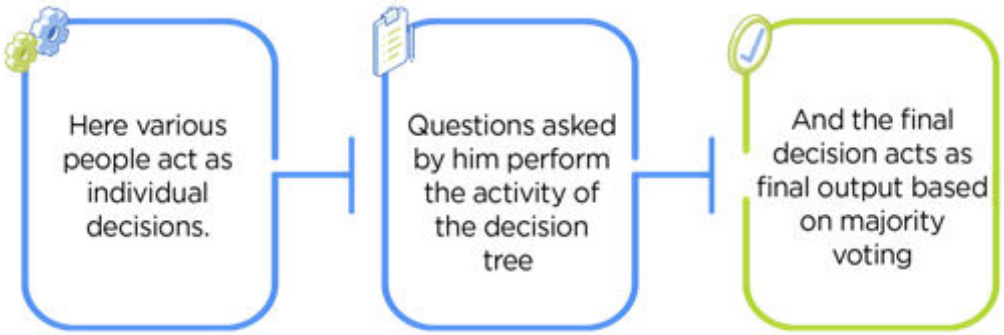
3.1 Random Forest Regressor

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

3.2 Random Forest Algorithm:-

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification task.



3.3 Working of Random Forest Algorithm

Before understanding the working of the random forest algorithm in machine learning, we must look into the ensemble learning technique. Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

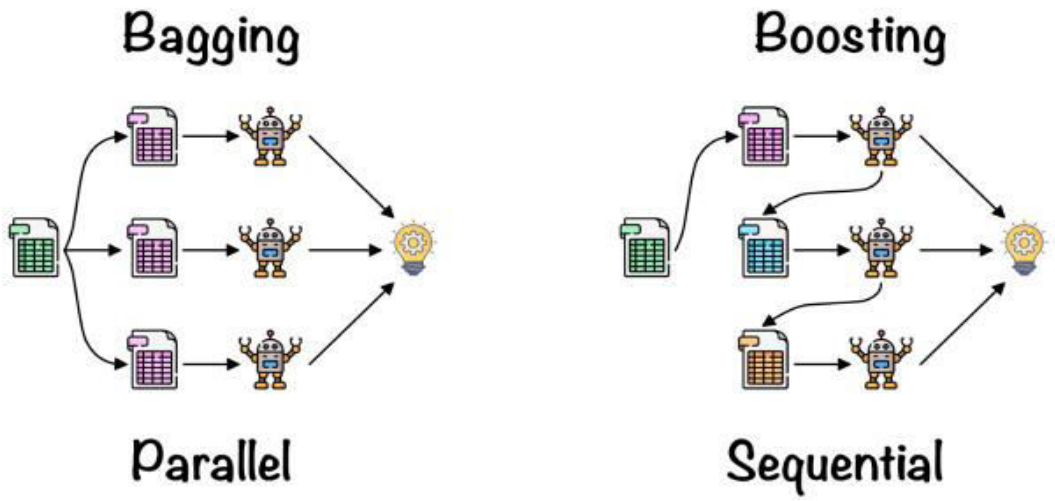
Ensemble uses two types of methods:

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707039 |

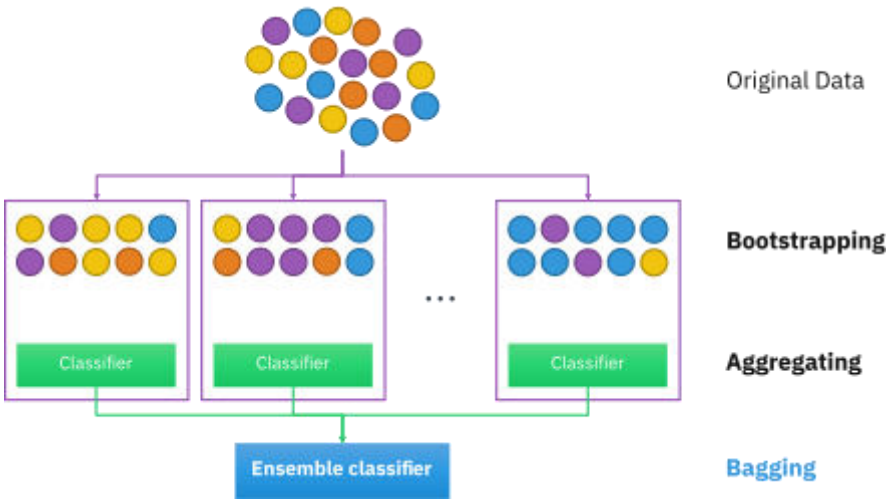
1. Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.
2. Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.



As mentioned earlier, Random forest works on the Bagging principle. Now let's dive in and understand bagging in detail.

3.4 Bagging

Bagging, also known as Bootstrap Aggregation, is the ensemble technique used by random forest. Bagging chooses a random sample/random subset from the entire data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently, which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting, is known as aggregation.



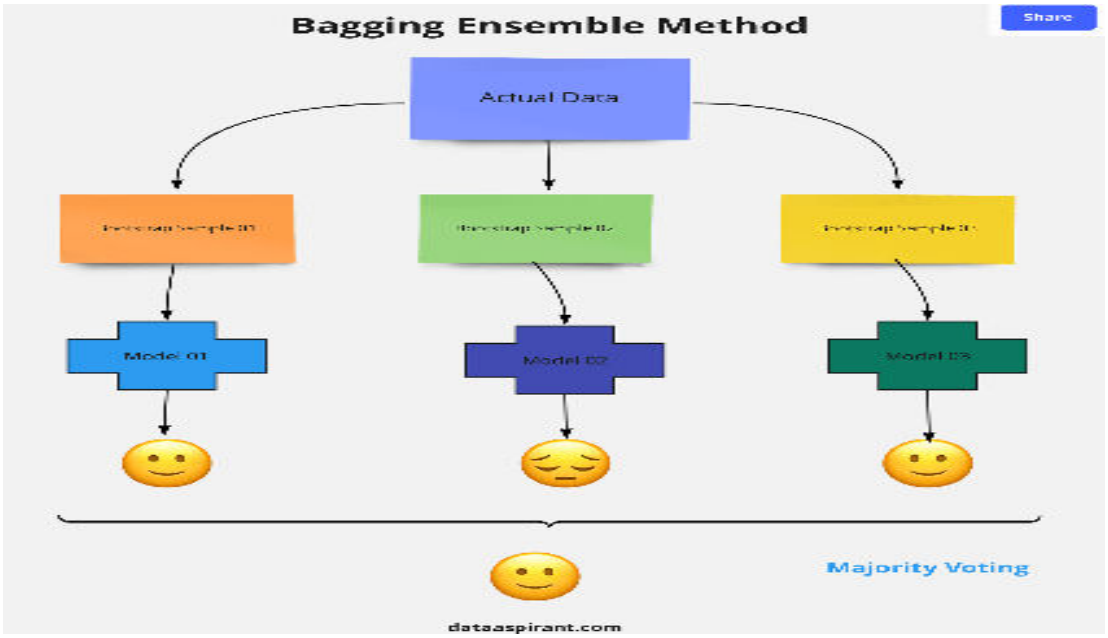
Now let's look at an example by breaking it down with the help of the following figure. Here the bootstrap sample is taken from actual data (Bootstrap sample 01, Bootstrap sample 02, and Bootstrap sample 03) with a replacement which means there is a high possibility that each sample won't contain unique data. The model (Model 01, Model 02, and Model 03) obtained from this bootstrap sample is trained independently. Each model generates results as shown. Now the Happy emoji has a majority when compared to the Sad emoji. Thus based on majority voting final output is

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707039 |

obtained as Happy emoji.



3.5 Boosting

Boosting is one of the techniques that use the concept of ensemble learning. A boosting algorithm combines multiple simple models (also known as weak learners or base estimators) to generate the final output. It is done by building a model by using weak models in series.

There are several boosting algorithms; AdaBoost was the first really successful boosting algorithm that was developed for the purpose of binary classification. AdaBoost is an abbreviation for Adaptive Boosting and is a prevalent boosting technique that combines multiple “weak classifiers” into a single “strong classifier.” There are Other Boosting techniques. Steps Involved in Random Forest Algorithm

Step 1: In the Random forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

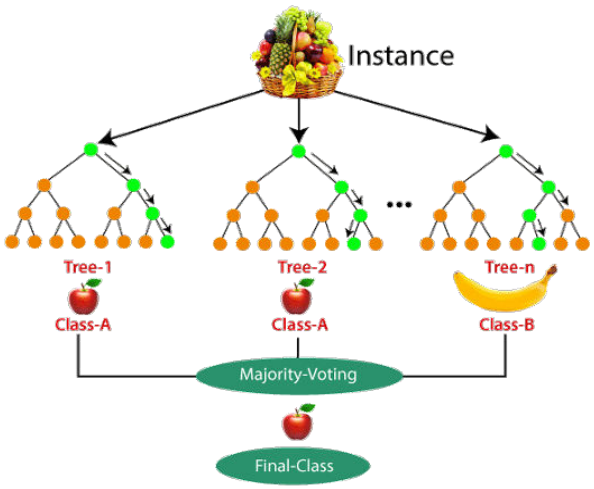
Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression, respectively.

For example: consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket, and an individual decision tree is constructed for each sample. Each decision tree will generate an output, as shown in the figure. The final output is considered based on majority voting. In the below figure, you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707039 |



Decision trees	Random Forest
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data, and the final output is based on average or majority ranking; hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation.	2. It is comparatively slower.
3. When a data set with features is taken as input by a decision tree, it will formulate some rules to make predictions.	3. Random forest randomly selects observations, builds a decision tree, and takes the average result. It doesn't use any set of formulas.

Thus random forests are much more successful than decision trees only if the trees are diverse and acceptable. Important Hyper parameters in Random Forest. Hyper parameters are used in random forests to either enhance the performance and predictive power of models or to make the model faster. Hyper parameters to Increase the Predictive Power

n_estimators: Number of trees the algorithm builds before averaging the predictions.

max_features: Maximum number of features random forest considers splitting a node.

mini_sample_leaf: Determines the minimum number of leaves required to split an internal node.

criterion: How to split the node in each tree? (Entropy/Gini impurity/Log Loss)

max_leaf_nodes: Maximum leaf nodes in each tree

Hyperparameters to Increase the Speed

n_jobs: it tells the engine how many processors it is allowed to use. If the value is 1, it can use only one processor, but if the value is -1, there is no limit.

random_state:controls randomness of the sample. The model will always produce the same results if it has a definite value of random state and has been given the same hyperparameters and training data.

oob_score: OOB means out of the bag. It is a random forest cross-validation method. In this, one-third of the sample is not used to train the data; instead used to evaluate its performance. These samples are called out-of-bag samples.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707039 |

Random forest is a great choice if anyone wants to build the model fast and efficiently, as one of the best things about the random forest is it can handle missing values. It is one of the best techniques with high performance, widely used in various industries for its efficiency. It can handle binary, continuous, and categorical data. Overall, random forest is a fast, simple, flexible, and robust model with some limitations.

IV. CONCLUSION

In this analysis, we used Exploratory Data Analysis (EDA) and a Random Forest Regressor to predict rainfall in millimeters (mm) based on various meteorological features such as temperature, humidity, wind speed, and pressure. First, we performed EDA to understand the relationships between the predictor variables and the target variable. We found that temperature and humidity had the strongest positive correlation with rainfall, while wind speed and pressure had weaker correlations. Next, we trained a Random Forest Regressor on the dataset and evaluated its performance using various metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The model was able to accurately predict rainfall with low error rates, indicating that it was a good fit for the data. Overall, the results of this analysis suggest that Random Forest Regression can be a powerful tool for predicting rainfall based on meteorological features, and that EDA can help identify which features are most strongly correlated with the target variable. These insights can be useful for a range of applications, from predicting agricultural yields to managing water resources in regions prone to drought.

REFERENCES

1. Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. Water (Switzerland), 10(11). https://doi.org/10.3390/w10111536
2. Janani, B; Sebastian, P. (2014). Analysis on the weather forecasting and techniques. International Journal of Advanced Research in Computer Engineering & Technology, 3(1), 59–61. http://ijarcet.org/wp-content/uploads/IJARCET-VOL-3-ISSUE-1-59- 61.pdf
3. Chaudhari, M. S., &Choudhari, N. K. (2017). Open Access Study of Various Rainfall Estimation & Prediction Techniques Using Data Mining. American Journal of Engineering Research (AJER), 7, 137– 139. http://www.ajer.org/papers/v6(07)/Q0607137139.pdf
4. AakashParmar, KinjalMistree, M. S. (2017). Machine Learning Techniques for rainfall prediction: A Review. International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS). https://www.researchgate.net/profile/Aakash_Parmar4/publication/319503839_Machine_Learning_Techniques_For_Rainfall_Prediction_A_Review/links/59afb922458515150e4cc2e4/Machine-LearningTechniques-For-Rainfall-Prediction-A-Review.pdf
5. Aftab, S., Ahmad, M., Hameed, N., Bashir, M. S., Ali, I., & Nawaz, Z. (2018). Rainfall prediction using data mining techniques: A systematic literature review. International Journal of Advanced Computer Science and Applications, 9(5), 143–150. https://doi.org/10.14569/IJACSA.2018.090518
6. Meng – Hua Yen , Ding – Wei Liu , Yi – Chia Hsin, C. – E. L. and C. – C. C. (2019). Application of the deep learning for the prediction of rainfall in Southern Taiwan. Scientific Reports, 9(1), 1–9. https://doi.org/10.1038/s41598-019-49242-6
7. Shi, X., Chen, Z., & Wang, H. (2015). Convolutional LSTM Network. Nips, 2–3. https://doi.org/[]
8. Wahyuni, E. G.Fauzan, L. M.F.Abriyani, F.Muchlis, N. F., &Ulfa, M. (2018). Rainfall prediction with backpropagation method. Journal of Physics: Conference Series, 983(1). https://doi.org/10.1088/1742- 6596/983/1/012059
9. Zeyi Chao, FanglingPu, YukeYinLing, B. and X. (2018). Research on real-time local rainfall prediction based on MEMS sensors. Journal of Sensors, 2018. https://doi.org/10.1155/2018/6184713
10. Etuk, E. H., & Mohamed, T. M. (2014). Time Series Analysis of Monthly Rainfall data for the Gadaref rainfall station, Sudan, by Sarima Methods. International Journal of Scientific Research in Knowledge, July, 320–327. https://doi.org/10.12983/ijsrk-2014- p0320-0327
11. Kar, K., Thakur, N., & Sanghvi, P. (2019). Prediction of Rainfall Using Fuzzy Dataset. International Journal of Computer Science and Mobile Computing, 8(4), 182–186. https://ijcsmc.com/docs/papers/April2019/V8I4201937.pdf
12. KavithaRani, B., & Govardhan, A. (2014). Effective Features and Hybrid Classifier for Rainfall Prediction. International Journal of Computational Intelligence Systems, 7(5), 937–951. https://doi.org/10.1080/18756891.2014.960234
13. Prabakaran, S., Naveen Kumar, P., &Sai Mani Tarun, P. (2017). Rainfall prediction using modified linear regression. ARPN Journal of Engineering and Applied Sciences, 12(12), 3715–3718. http://www.arpnjournals.org/jeas/research_papers/rp_2017/jeas_0617 _6115.pdf

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com