International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Exploiting Ensemble Learning for Rainfall Prediction using Meta Regressors and Meta Classifiers

Kovvuri N Bhargavi, G.Jaya Suma

Abstract*: Intense rainfall produces flooding even on dry soil. As heavy rainfall is one of the causes for flooding it is necessary to predict the Rainfall to take necessary precautions for people who are living in risk zone areas. Prediction of Rainfall tomorrow is done accurately using Machine Learning regression and classification Techniques. For Rainfall prediction multiple attributes like Windspeed, Precipitation, Cloudcover, Humidity, Temperature and RainfallToday are considered to predict Rainfall Tomorrow. An ensemble approach is used where predictions from Regression models such as Linear Regression, Polynomial Regression, Ridge Regression and Lasso regression are stacked together and fed as new attributes to Meta Regressor along with Support Vector Regression for making final predictions. Also, predictions from classifications models such as Gaussian Naive Bayes, K-nearest neighbor, Support vector Machine and Random Forest are stacked together and fed as new attributes to Meta Classifier along with Logistic regression which is a binary classifier for higher predictive performance.

***Keywords* :Meta Regressor ,Meta Classifier, Support Vector Regressor, Random Forest, Ridge Regression, Lasso Regression.**

I. INTRODUCTION

Flooding is due to continuous water flow towards land that is commonly dry. In case of heavy rainfall for longer durations mostly urban areas are affected with floods due to lack of drainage System. Floods in Urban areas are a great trouble to the individuals in the city. People cannot go to their work as roads, streets get blocked. Flash floods are caused to Heavy Rainfall within a short period of time results with harmful and unpredictable destruction. Drinking water is contaminated due to floods results in spreading of water-borne and vector-borne communicable diseases. Floods are caused by different reasons, one of the chances for flooding is heavy rainfall. Finding Flood prone areas, identifying the cause of flooding and educating people in nearby areas helps from being affected. Various Machine learning algorithms are used to accurately predict the rainfall in advance based on the atmospheric attributes. If the rainfall is predicted in advance the flood intensity will be reduced by taking necessary precautions. Machine Learning Algorithms models are trained and tested to get accurate predictions. Accuracy of different Machine Learning Regression and classification are combined using Ensemble stacking Meta Classifier and Meta Regressor for accurate Rainfall Prediction.

Revised Manuscript Received on February14, 2020.

* Correspondence Author

Kovvuri N Bhargavi*, Sr,Asst Prof, Department of CSE, Aditya College of Engg& Tech, Surampalem, India. Email: bhargavinag@gmail.com

Dr.G.Jaya Suma, Prof&HOD, Department of IT, UCEV JNTUK Vizianagaram, India. Email: gjssuma@gmail.com

II. LITERATURE

JoRefonaa proposed a rainfall forecast model using linear regression. Rainfall Prediction is difficult to forecast as atmosphere changes are dynamic. By considering historical records of a geographical area rainfall prediction proves to be more reliable showing good accuracy.[1] Brett W Robertson proposed a model which analyses the Social media information gathered for real time disaster. urgency and time period of social media images are classified using deep learning Convolution Neural network and Multilayer Perceptron neural networks.[2]. Sagnik Proposed methodology by considering the mean day by day Gauge heights, mean day by day precipitation, and the mean day by day river discharge values to estimate4 days ahead of time. These attributes are fed as input to extreme learning machine regression model for mean gauge height Prediction.[3].The quantity of units in the ELM was improved to acquire the most extreme coefficient of assurance utilizing the molecule swarm advancement calculation (PSO) to develop ELM-PSO model .Dieu Tien proposed a flash flood prediction model with deep learning neural network approach. High frequency tropical storm area case study is discussed [4]. Rahul Proposed a Prediction model for Natural Calamities Detection by combining convolution neural network and SVM. Natural disasters like tsunami, cyclone, earthquake, volcano eruption, wild fire, landslide datasets are taken as input for classification.[5] Kishan Kumar proposed house hold analysis affected due to floods based on the flood data in several districts of Bangladesh. Machine learning techniques are used for prediction and influencing factors for conducting flood reduction programs. Analysis shows important factors for flood prediction and Principle component Analysis is applied to predictors and analyze their effect on flood damage.[6] Amir Mosavi introduces Hybridizing of existing novel machine learning models for finding more accurate and predictive models.[7] Jiansheng proposed a hybrid optimization strategy genetic algorithm for better accuracy of rainfall prediction.[8] Andrew Kusiak performed analysis using data mining predictive models support vectors, random forest, neural networks, K-nearest neighbors, regression tree and Random Forest for rainfall prediction.[9] Jeerna applied machine learning algorithms for pluvial flood forecasting which is a rare disaster with small duration resulting high impact in in urban areas.

Upstream and downstream of Pattani river are tested using Bayesian Linear model.[10] Ajay Kumar proposes Back propagation NN and Radial basis NN are widely used for rainfall prediction. Radial functions parameters are very complex and sensitive to over-fitting. These models work well on past training data but may not be good in predicting future events. Machine learning algorithms are trained by mappings from inputs to outputs, based on samples data.[11] Indian Government in collaboration with Google is validating its machine learning flood forecasting and also generating high resolution maps for flood prediction.

III. METHODOLOGY

A. Prediction Analysis using Regression Models

- **Prediction Analysis using Linear Regression**

Linear Regression establishes a linear relationship between independent attributes Windspeed, Precipitation, Cloudcover, Humidity, Temperature, rainfall and target attribute RainTomorrow.

The LinearModel is defined as

$$y=0.008285x_1-0.0189232x_2+0.00777338x3 +0.02723107x_4-0.00121553x_5-0.00560831x_6+18.82369$$

where x(x1,x2,x3,x4,x5,x6) are independent attributes and y is the target attribute RainTomorrow. Fig1 shows the relationship between actual and predicted values of Linear regression. Fig. 1 shows the relationship between actual and predicted values of Linear regression

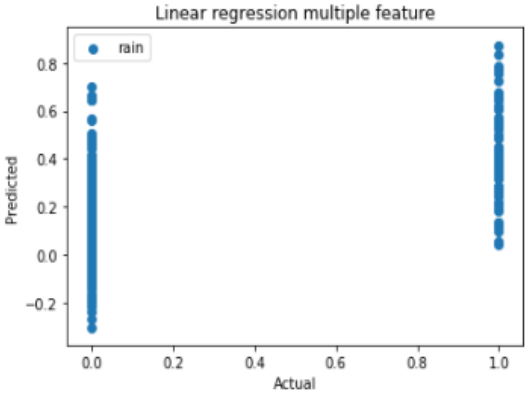

Fig. 1. Actual Vs Predicted using Linear Regression

- **Prediction Analysis using Ridge Regression**

Applying Prediction Analysis using Ridge regression is same as l2 regularization to linear regression. L2 regularization is performed by adding penalty to square of the magnitude of coefficients. Even regularization is applied to linear regression the error rate is same and coefficients are same as linear regression. Fig. 2 shows the relationship between actual and predicted values of Ridge Regression.

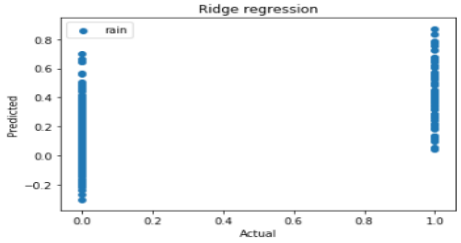

Fig. 2. Actual Vs predicted using Ridge Regression

- **Prediction Analysis using Lasso Regression**

Applying Prediction Analysis using Lasso Regression is same as applying L1 regularization to Linear regression. L1 Regularization is performed by adding penalty. Using Lasso regression model some coefficients are shrinked to zero behaves like sparse models. Ridge regression is same as lasso regression but does not result in sparse models. In Lasso regression simpler models are generated where large penalties result in coefficient values closer to zero. So, Lasso regression is easier to interpret than Ridge regression. Fig. 2 shows the relationship between actual and predicted values of Lasso regression. Coefficients of Lasso Regression generating simple models by elimination some of the coefficients. [0.00839667 -0.01660734 0.0116289 0.0059810 7 -0. -0.] Intercept:16.531212761493965

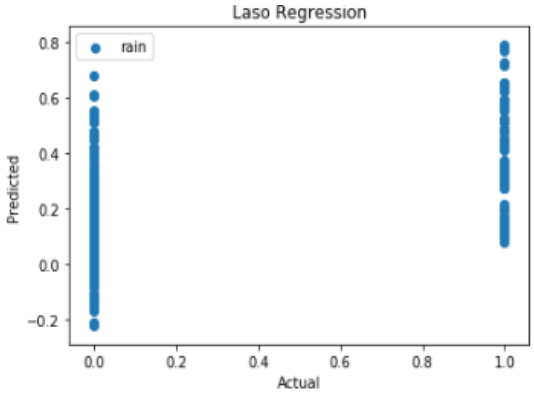

Fig. 3. Actual Vs Predicted using Lasso

- **Prediction Analysis using Polynomial Regression.**

Polynomial regression is same as linear regression but provides best approximate relationship between independent and dependent variables as nth degree polynomial. Prediction Analysis based on nth degree polynomial minimizes the error rate compared to Linear Regression. Fig. 4 shows Relationship between actual and predicted values for nth degree polynomial

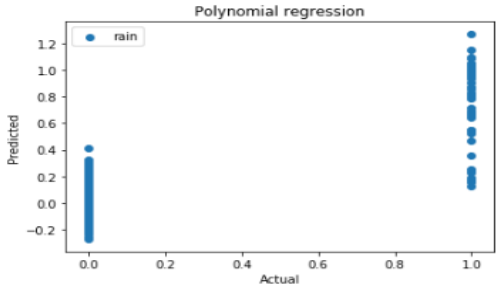

Fig. 4. Actual Vs Predicted using Polynomial

- **Prediction Analysis using SVR Regression**

Prediction Analysis of Target Attribute is done using radial basis function kernel. Support Vector regression works like linear regression in higher dimension by reducing the model complexity and minimizing the loss function. SVR works better compared to all the regression models by minimizing the mean squared error rate. In the Fig. 5 actual values Vs predicted values shows all the samples belong to class 1 i.e. RainTomorrow 1 are predicted as 1 accurately.

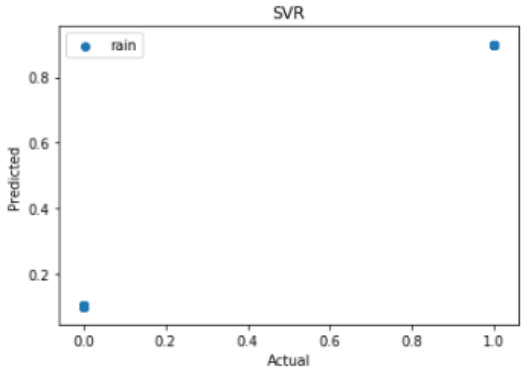

Fig. 5. Actual Vs Predicted using SVR

B. Combining the predictions of Regression models using Meta Regressor:

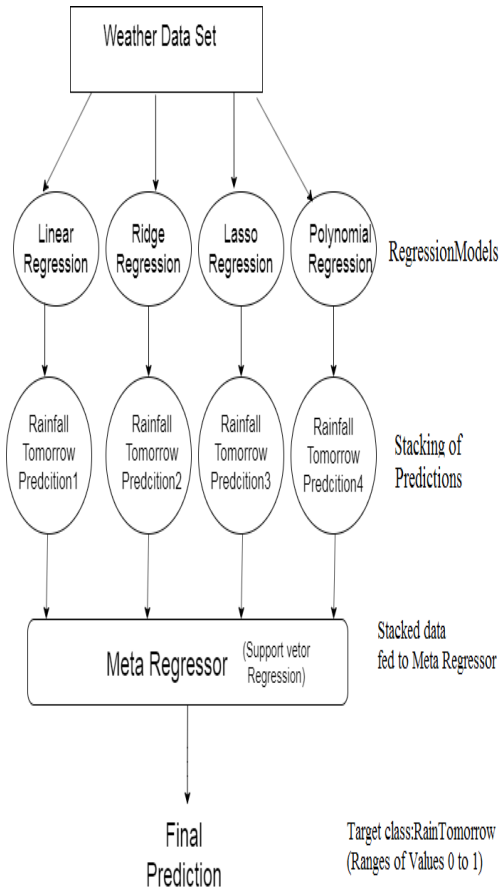

Fig.6. Meta Regressor

Meta regressor is an ensemble approach used to combine the predictions of set of regression models. For prediction analysis of target attribute RainTomorrow linear regression, polynomial regression, Ridge regression, lasso regression and support vector Regression are stacked together and given as input to Meta Regressor for Prediction as shown in Fig 6. Relationship between actual and predicted values are shown in Fig7.

Algorithm:

Input: Rainfall Data set with attributes X= {Windspeed, Precipitation, Cloudcover ,Humidity, Temperature, RainToday}

Output: Target Attribute y=RainTomorrow

Step1: In first phase Regression models Linear regression, Ridge regression, Lasso regression and Polynomial Regression are trained on X.

Step2: Each trained Model performs prediction on target attribute y.

Step3: In second level all the predictions of Regression models are stacked together and given as input to the Meta Regressor. Also, another regression model chosen is given directly to the Meta Regressor.

Step4: Meta Regressor is trained on this data and produces the final prediction.

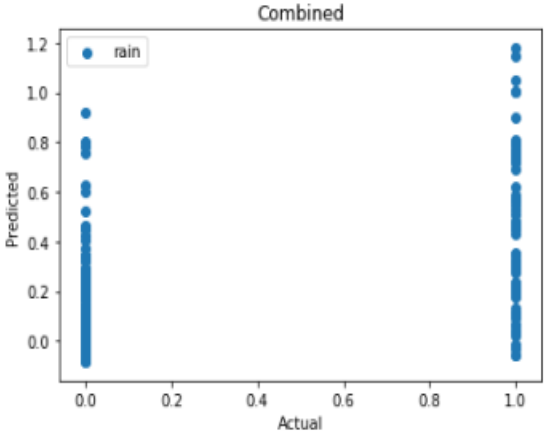

Fig.7. Actual Vs Predicted using Meta Regressor

C. Prediction Analysis using Classification Models

- **Prediction Analysis using Gaussian Naive Bayes Classification model**

Gaussian Naive Bayes classification technique is used for predicting the Rainfall Tomorrow target attribute using continuous attributes Windspeed, Precipitation, Cloudcover, Humidity, Temperature are distributed according to Gaussian Distribution.

Performance Analysis for evaluation of target variable RainTomorrow using Confusion Matrix and Accuracy Metrics as in Table-1.

Class 0 indicates RainTomorrow='yes' and Class 1 indicates RainTomorrow='no'

Confusion Matrix:

[[94 5]
[6 5]]

Accuracy Score: 0.9

Table -I: Accuracy Report

	precision	recall	f1-score	support
Class 0	0.94	0.95	0.94	99
Class 1	0.5	0.45	0.48	11
accuracy			0.9	110
macro avg	0.72	0.7	0.71	110
weighted avg	0.9	0.9	0.9	110

- **Prediction Analysis using K-nearest neighbor classification Model**

K nearest neighbor Classifier predicts the target attribute RainTomorrow based on the 'feature similarity' of k neares t neighbors. To find the nearest k neighbors' distance betwee n the data points is measured using distance metrics like Eucl idean Distance, Hamming, Manhattan and Minkowski.
Performance Analysis for evaluation of target variable RainTomorrow using Confusion Matrix and Accuracy metri cs
as in Table-II:
Confusion Matrix:
[[92 7]
[8 3]]
Accuracy Score: 0.8636363636363636

Table-II: Accuracy Report

	precision	recall	f1-score	Support
Class 0	0.92	0.93	0.92	99
Class 1	0.3	0.27	0.29	11
accuracy			0.86	110
macro avg	0.61	0.6	0.61	110
weighted avg	0.86	0.86	0.86	110

- **Prediction analysis using SVM classification model**

Support Vector Machine Classifier is a one of the supervise d classification techniques that defines a hyperplane which s eparates the binary target attribute into two classes RainTom orrow=0 or 1. With the help of this hyperplane new records can be classified to either RainTomorrowis equal 0 or 1.

Performance Analysis for evaluation of target variable RainTomorrow using Confusion Matrix and Accuracy metri cs as in Table-III:
Confusion Matrix:
[[99 0]
[11 0]
Accuracy Score: 0.9

Table -III: Accuracy Report

	precision	recall	f1-score	Support
Class 0	0.9	1	0.95	99
Class 1	0	0	0	11
Accuracy			0.9	110
macro avg	0.45	0.5	0.47	110
weighted avg	0.81	0.9	0.85	110

- **Prediction analysis using Random Forest Classification Model**

Using Random Forest Classifier multiple decision trees are constructed on sub samples of the Weather data set. To classify new instances based on the target attribute RainTomorrow classifications of multiple decision trees are considered and the result of best voted prediction is chosen as the final prediction.
Performance Analysis for evaluation of target variable RainT omorrow using Confusion Matrix and Accuracy metrics as
in Table-IV:
Confusion Matrix:
[[95 4]
[5 6]]
Accuracy Score: 0.9181818181818182

Table-IV: Accuracy Report

	precision	Recall	f1-score	support
Class 0	0.95	0.96	0.95	99
Class 1	0.6	0.55	0.57	11
accuracy			0.92	110
macro avg	0.77	0.75	0.76	110
weighted avg	0.91	0.92	0.92	110

- **Prediction Analysis using Logistic Regression**

Logistic Regression is a binary classification technique for predicting binary target variable Rainfall Tomorrow as 1 (Rainfall) or 0 as (no Rainfall). When compared to linear regression in which predicted values of RainTomorrow range between 0 to 1, logistic regression directly classifies positive class (RainTomorrow=1) or negative class (RainTomorrow=0).
Logistic regression model is represented as
$y = 1/(1 + e^{-x})$
where
$x = (a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6)$,
y is target variable Rainfall Tomorrow and $x_1, x_2, x_3, x_4, x_5, x_6$ are independent attributes Windspeed, Precipitation, Cloudcover, Humidity, Temperature, RainToday and $b_1, b_2, b_3, b_4, b_5, b_6$ are the coefficients.
Performance Analysis for evaluation of target variable RainTomorrow using Confusion Matrix and Accuracy metrics as in Table-V.
Confusion Matrix:
[[92 7]
[4 7]]
Accuracy Score: 0.9

Table-V: Accuracy Report

	precision	recall	f1-score	Support
Class 0	0.96	0.93	0.94	99
Class 1	0.5	0.64	0.56	11
accuracy			0.9	110
macro avg	0.73	0.78	0.75	110
weighted avg	0.91	0.9	0.91	110

- **Combining the predictions of Regression models using Meta Classifier:**

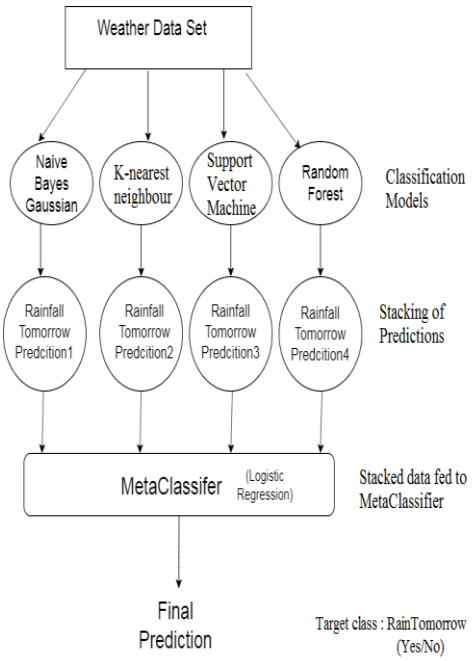

Fig. 8. Meta Classifier

Meta Classifier is an ensemble approach used to combine the predictions of set of classification models. For prediction analysis of target attribute RainTomorrow Gaussian Naïve Bayes, K nearest neighbor, Support Vector and Random Forest Classification models are stacked together and given as input to Meta Classifier for Prediction as in Fig 8.

Algorithm:

Input: Rainfall Data set with attributes X= {Windspeed, Precipitation, Cloudcover, Humidity, Temperature, rainfall today}

Output: Target Attribute y=RainTomorrow

Step1: In first phase classification models Gaussian Naive Bayes, K-nearest neighbor, Support Vector Machine, Random Forest Classification models are trained on X.

Step2: Each trained Model performs prediction on target attribute y.

Step3: In second phase all the predictions of classification models are stacked together and given as input to the Meta Classifier. Also, another Classification model Logistic Regression is given directly to the Meta Classifier.

Step4: Meta Classifier is trained on this new data and produces the final prediction.

Performance Analysis of stacked classifier for evaluation of target variable RainTomorrow using Confusion Matrix and Accuracy metrics as in Table-VI:

Confusion Matrix:

[[95 4]
[5 6]]

Accuracy Score: 0.9181818181818182

Table-VI: Accuracy Report

	Precision	recall	f1-score	support
Class 0	0.95	0.96	0.95	99
Class 1	0.6	0.55	0.57	11
Accuracy			0.92	110
macro avg	0.77	0.75	0.76	110
weighted avg	0.91	0.92	0.92	110

IV. RESULT AND DISCUSSION

- **Results Showing Performance of Regression Models**

Performance of regression models are measured and tabulated as shown in Table-VII

Table-VII: Measured Error Rate

Regression Models	Error
Linear Regression	MSE: 0.10110282104248075 RMSE: 0.31796669800858196 MAE: 0.23586298511086864
Polynomial Regression	MSE: 0.029923878392277913 RMSE: 0.1729851970322256 MAE: 0.10527167148435093
Lasso Regression	MSE: 0.10355612530683823 RMSE 0.32180137555150107 MAE: 0.2343476017344406
Support Vector Regressor	MSE: 0.00999762136771921 RMSE: 0.09998810613127548 MAE: 0.09998777347918802

Meta Regressor	MSE: 0.10156297891910297 RMSE: 0.31868947098877143 MAE: 0.1810369308154861

- **Results Showing Performance of Classification Models**

  Performance of classification models are measured and tabulated as shown in Table-VIII

Table-VIII: Accuracy values of Classifiers

Classifiers	Accuracy
Gaussian Naïve Bayes	0.9
K-nearest neighbor	0.86363
SupportVector Machine	0.9
Random Forest	0.91818
Logistic Regression	0.9
Meta Classifier	0.91818

V. CONCLUSION

Different Regression Models are trained on Weather data set to predict the target attribute RainTomorrow in the range of 0 to 1 value. Predictions of all the regression models are stacked together. Results shows that Support Vector Regression model shows least Mean squared error as 0.009 as best regression model. Similarly, different Classification models are trained on Weather dataset to predict RainfallTomorrow=Yes/No. Prediction of Classification models are stacked together. Results shows that the Meta Classifier accuracy and Random Forest accuracy are 98.18% as best classification models for Rainfall Prediction.

REFERENCES

1. J.Refonaa,M.Lakshmi,RazaAbbas,MohammadRaziullha"Rainfall Prediction using Regression Model" International Journal of Recent Technology and Engineering ISSN:2277-3878,Volume-8 July 2019
2. Brett W. Robertson,MJhonson,DhirajMuthy,W Routh Smith,K,K,Stephens "Using a combination of human insights and deep learning for real-time disaster communication" Progress in Disaster Science2 May 2019
3. Sagnik, Padmini "Flood forecasting using a hybrid extreme learning machine-particle swarm optimization algorithm model" Modelling Earth Systems and environment November 2019
4. Dieu Tien, Nhat-Duc Hoangb "A novel Deep Learning Neural Network approach for predicting flash flood susceptibility" Science of Total Environment 2018
5. N Rahul, Rishi Megha, Tiwari Amit, Dua Rajat " A Novel Deep Learning Framework Approach for Natural Calamities Detection" Springer Conference information and communication Technology for Competitive Strategies pages 561-569 2018
6. Kishan Kumar G,NaiduNahar,B M Mainul Hossain " : A case study of Floods in Bangladesh."International journal of disaster and risk reduction.Dec2018
7. Amir Mosavi, Pinar Ozturk and Kwok-wing Chau "Flood Prediction using Machine Learning Models:Literature Review" Water MDPI Oct 2018
8. jeeranaNoymanee, Nikolay O,Nikitin,Anna V Kalyuzhnaya "Urban Pluvial Flood Forecasting using open data with Machine learning Techniques in pattani basin" ScienceDirect Procedia Computer Science 119(2017)288-297
9. Jiansheng Wu, Jin Long and Mingzhe Liu" Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm" Elsevier Neuro computing 2015
10. Ajay Kumar and Nikita Tyagi "Comparative analysis of backpropagation and RBF neural network on monthly
11. rainfall Prediction" Inventive computational technologies 2016 International Conference.
12. Andrew K, Xiupeng Wei, Anoop prakashverma and Evan Roz "Modeling and Prediction of Rainfall using Radar Reflectivity Data: A Data-Mining Approach" IEEE Transactions on GeoScience and Remote Sensing, Vol 51 No.4,April 2013.
13. M.Ghasem, S.Adel, S Milad, D Morad "An integrated data-mining and multi-criteria decision-making approach for hazard-based object ranking with a focus on landslides and floods" Environmental Earth Sciences 2018
14. Prasad Pangali, j Zhang, U. Ashish, Sha Zhang and Krishna Suwal "Review of flood disaster studies in Nepal" International Journal of Disaster Risk Reduction 2019.
15. B Kumar, Sowmya, Kumar Nadh and Ranjan S "An Application of Data Mining Techniques for Flood Forecasting: Application in Rivers Daya and Bhargavi, India" The institute of Engineers 2018

AUTHORS PROFILE

Kovvuri N Bhargavi, working as Sr.Asst Prof in the Department of CSE, Aditya College of Engineering & Technology. She has 12 years of experience in teaching field. She has completed Mtech in the year 2009. She is a member of CSI.She is doing PartTimePh,D in JNUTK Kakinada under the esteemed Guidance of Dr.G.Jaya Suma. Her Areas of Interest are Data Mining, Machine Learning and IOT.

Dr.G.JayaSuma, is working as HOD in Department of IT JNTUK UCEV. She has been awarded Ph.D in Data Mining from Andhra University in the year 2011 and currently guiding research scholars. She has total 16 years of experience in teaching field and two years of experience in industry.She has published several research papers and Book Chapters in reputed International Journals. She has attended National and International conferences. She is a member of CSI, ISTE and IEEE.Her areas of Interest are Data Mining, Machine Learning, Softcomputing and Mobile Computing.