
ISSN: 2321-2152
IJMECE
International Journal of modern
electronics and communication engineering
E-Mail
editor.ijmece@gmail.com
editor@ijmece.com
www.ijmece.com

ISSN2321-2152www.ijmece .com

Vol 08, Issuse.2 April 2021

Predicting Rainfall with Machine Learning: An Evaluation

AMITAV SARAN

Abstract

Predicting heavy precipitation is a major challenge for the weather service. Lasso regression, ridge regression, elastic net regression, random forest, gradient boosting, and the decision tree regress or are only few of the Machine Learning (ML) models that are analyzed in this work. Evaluation criteria including R2 score, Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error have been used to determine how well each model performs (RMSE). This research aims to examine and contrast several machine learning regression techniques using the rainfall dataset. Six different ML models were evaluated, and after careful consideration, we found that Lasso regression of the linear model provided the best results. At 80% training data set and 20% test dataset, the Lasso model achieved a 99.21% R2 score, a 13.68 MAE, a 6432.41 MSE, and an 80.20 RMSE.

Introduction

In the hydrological study, the main problem is accurately predicting the rainfall. Due to natural hazards and storm, farmers will lose and destroy their crops. To avoid these problems, accurately and timely predict the rainfall prediction earlier and give caution more first to farmers. Rainfall is said to be an environmental aspect which affects the human activities such as farming production, construction, energy generation, forestry and tourism, etc. The rainfall prediction is more essential as it is concerned with the maximum association with other factors like landslides, flooding, avalanches, earthquakes, and so

ASSISTANT PROFESSOR, Mtech,Ph.D
Department of CSE
Gandhi Institute for Technology,Bhubaneswar.

On [1]. The rainfall prediction is more required as it is concerned with the maximum association with other factors like landslides, flooding, avalanches, earthquakes, and so on. Such disasters affect the public severely for many decades [2]. Hence, developing effective model to predict the rainfall helps to prevent the natural disaster to the limited extent [3]. We applied different regression techniques of machine learning algorithms to build the ML models to make accurate and timely predictions. Machine learning is used to study and develop the system behavior model.Machine learning modeling techniques used to design models which can be further predicted vital system parameters with regards to Indian panther ecosystem [4]. This article aims to deliver end to end machine learning life cycle right from Data acquisition to evaluating the models. For evaluation metrics of regress or is R2, Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square (RMSE). The article has been organized as follows: Segment II elaborates literature review for rainfall prediction using different regression algorithms. Segment III explains the various Regression models of machine learning algorithms that applied in this research article for regression purposes. Segment IV describes the novel experiments conducted towards the deployment of in-depth learning solutions in rainfall prediction. Segment V explains the conclusion and future work.

Literature Review

Accurately weather prediction is a big challenge for us. Rainfall forecasting methods involve a grouping of computer models and patterns. Accurate and timely weather predicting is a challenging issue for the scientific community. Rainfall forecasting modeling consists of a cluster of computer models and observation. Regression is a statistical and empirical method used in business and climate forecasting. El-shafie [5] et al. used Artificial Neural Network to forecast rainfall-runoff association in a catchment zone of Japan. They suggested a model with the practice of feed-forward backpropagation with hyperbolic tangent neurons in the processing layer and linear neuron in the target layer. Model performance is evaluated by other statistical indexes like correlation coefficients and mean square error. The proposed model was more accurate. Nikhil Sethi, [6] et al. It has proposed a method for rainfall prediction in the future by knowing climate factors, which is very helpful for farmers for their agriculture purpose. In this article, the author proposes only one model that is multiple linear regressions of machine learning algorithms. Ashwani [7] et al. Data mining techniques like ANN and Decision tree Algorithms had been applied in estimating whether by using in meteorological data and which is gathered at a particular period. Standard implementation metrics of algorithms given the accurate scores and which were used to compare the model's performance and choose the beater model to predict the weather. Liu et al 2001 [8] developed an alternate model. It is used to find the employment of Genetic Algorithms (GA) which can be applied as Feature Selection (FS) model, Naive Bayes (NB) as prediction technique. These modules are divided into two predictive methods: rainfall event which is referred to be a binary prediction module and a classification of rainfall which may be light, gradual as well as severe rainfall. The application of GA is to select the inputs, which exhibit a viable option to minimize the difficulty of dataset achieving identical or optimal function.

Research Methodology

The whole process of the Framework for the machine learning method is explained in figure1. The entire process is classified into four stages had been a, namely data acquisition, data pre-processing, Build the machine learning model, and predict the target variable with the trained model. From figure 1, the framework for the machine learning model has been explained below.

Data Acquisition

In the first stage, Raw Data which is collected from the India Meteorological Department (IMD) Govt. of India [9]. Raw data is observed in figure 2.

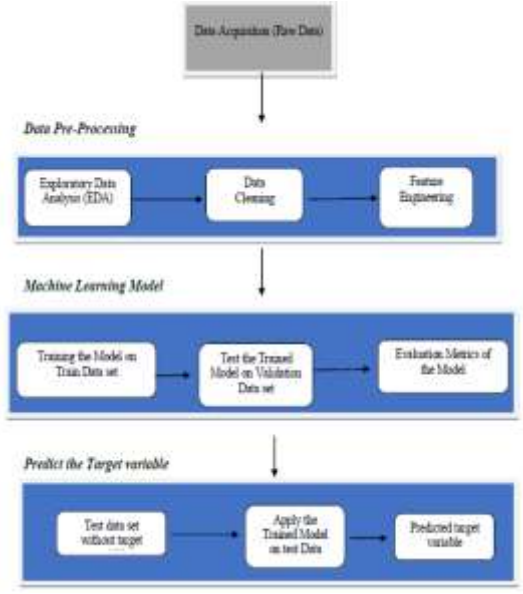

Figure 2.Framework for Machine learning Model

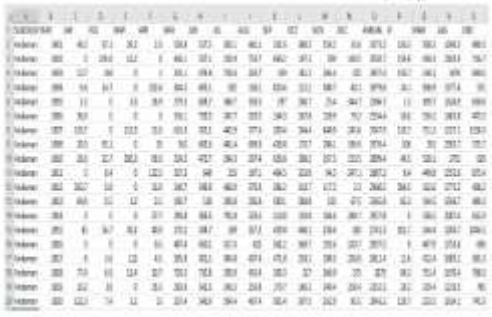

Figure 2. Raw Dataset of rainfall Collected data set includes monthly rainfall feature of 36 meteorological sub-divisions of India during the period of 1901-2017 is observed in figure 2. Given data set contains 4188 instances and 19 features are observed in figure 3. Is Among19 variables „Annual‟ variable is a target variable and remaining 18 variables are 'SUBDIVISION', 'YEAR', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL','AUG', 'SEP', 'OCT', 'NOV', 'DEC', 'JF', 'MAM', 'JJAS', 'OND' considered as input variables (predictor variables).

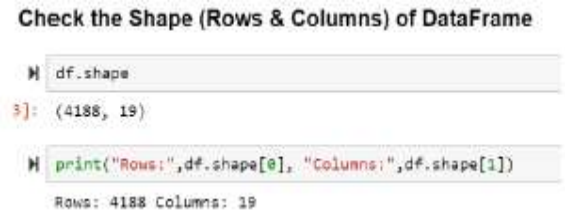

Figure 3. Shape of the data set *3.2 Data Pre-Processing* For achieving better results, the applied ML models, Data Pre-processing, is the second stage of the research methodology, which has three sub-parts, which are Exploratory Data Analyses (EDA), data cleaning, and feature engineering. *3.2.1 Exploratory Data Analysis (EDA)* Exploratory Data Analysis is about applying methods on data to gain observation before applying machine learning techniques. EDA explains data by means of statistical and visualization techniques. It brings out the essential aspects of the data. EDA also plays a crucial role in helping choose the right ML model to solve a specific problem. *3.2.1.1 Average Annual Rainfall in Each Subdivision* Here, we are finding the subdivisions with the highest and lowest rainfall, from figure 4 we notice that Subdivisions with highest annual rainfall are "COASTAL KARNATAKA," "ARUNACHAL PRADESH" and "KONKAN & GOA" with an approximate yearly rainfall of 3403 mm, 3397 mm and

2987 mm respectively. Subdivisions with the lowest annual rainfall are "HARYANA DELHI & CHANDIGARH," "SAURASHTRA & KUTCH," and "WEST RAJASTHAN" with an approximate annual rainfall of 528 mm, 496 mm and 294 mm respectively.

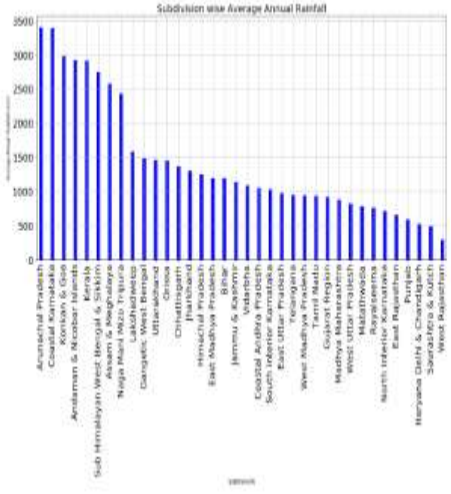

Figure 4. Subdivision wise highest and lowest rainfall

3.2.1.2 Rainfall in Subdivisions From figure 5, we noticed that, majority of rainfall is received in the months of JUNE, JULY, AUGUST, SEPTEMBER (JJAS) from Coastal Karnataka, Arunachal Pradesh, Konkani Goa, and Kerala and which are receiving the highest rainfall.

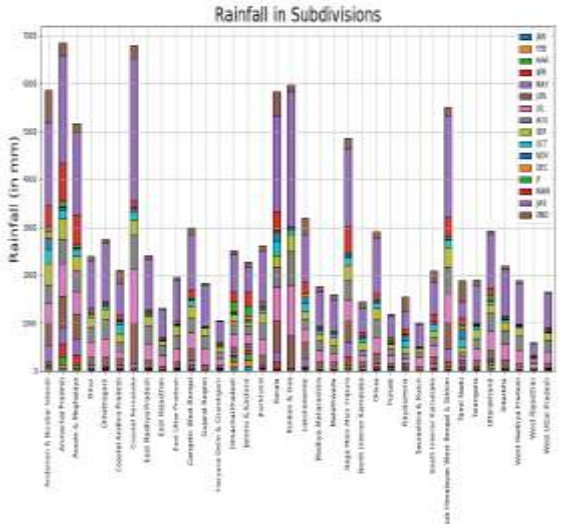

Figure 5. Rainfall in subdivisions of monthly wise

3.2.2 Data Cleaning Data cleaning is the subpart in data pre-processing. Under data cleaning, some of the operations had been applied to handle unnecessary data like duplicates, outliers, and missing values. From Figures 6 & 7, we notice that how much percentage of values is missing in all features. Fill the null values with a mean of that corresponding that features.

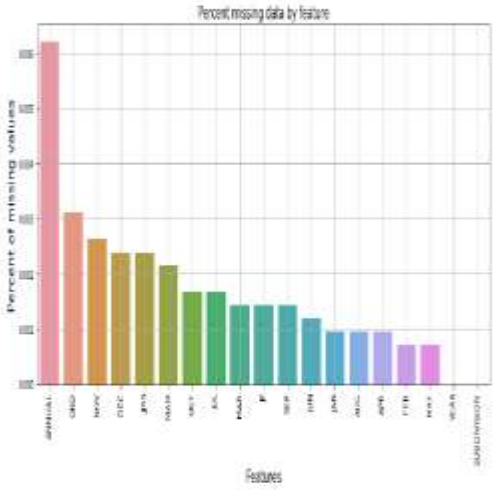

Figure 6. Percentage of missing values by features

	Total	Percent
ANNUAL	28	0.006209
OND	13	0.003104
NOV	11	0.002927
DEC	10	0.002388
JJAS	10	0.002388
MAM	9	0.002149
OCT	7	0.001671
JUL	7	0.001671
MAR	6	0.001433
JF	6	0.001433
SEP	6	0.001433
JUN	5	0.001194
JAN	4	0.000955
AUG	4	0.000955
APR	4	0.000955
FEB	3	0.000716
MAY	3	0.000716
YEAR	0	0.000000

Figure 7. Percentage of Missing values with the total number of values in features

4.1) Linear Models

There are three exclusive linear regression models, which are Lasso, Ridge, and Elastic Net regression. The easiest method to forecast output by applying a linear function of input features.

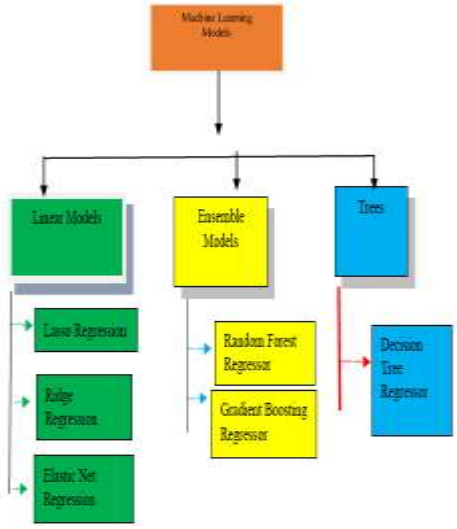

Figure 8. Regression of Machine Learning models there is an association between one or more independent or input features (X) and dependent or target feature (y) for simple Linear Regression (SLR). The regular equation for linear regression is assumed as yi= mixi+b. For multiple explanatory variables, where 'y' represents the target feature, and 'X' represents independent variable where i=0,1,2,.., n, indicates the explanatory or independent variables, 'm' termed as a slope. The process has been explained as Multiple Linear Regression (MLR)[10].

$$\hat{y} = m_0 x_0 + m_1 x_1 + \cdots + m_n x_n + b \qquad \text{----------(1)}$$

$$= \sum_{i=1}^{r} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{r} \left(y_i - \sum_{j=0}^{c} m_j \cdot x_{ij} \right)^2 \qquad \text{----------(2)}$$

The cost function for simple linear regression is defined in equation (2) from this equation; assume that there is being 'r' rows or instances and 'c' columns or features. The whole data set has been classified into a train and validation data set. Lasso and ridge regression models are used to minimize the complexity of the model and prevent over-fitting problems. *4.1.1) Ridge Regression:* Add the penalty to the square of the magnitude in the coefficient in the ridge regression.

$$= \sum_{i=1}^{r} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{r} \left(y_i - \sum_{j=0}^{c} m_j \cdot x_{ij} \right)^2 + \lambda \sum_{j=0}^{c} m_j^2 \qquad \text{----------(3)}$$

The above equation (3) is the cost function of Ridge regression. So, ridge regression had been set a constraint on the coefficients (m). [11] Factors had been regularized when we apply the penalty term (lambda (λ)), then the optimization function is penalized. So, ridge regression minimizes the coefficients, and it helps to decrease the model complication. The significant advantage of ridge regression is „coefficients shrinking" and reducing the „model complication." supposing, when putting λ=0,

the cost function of ridge regression becomes similar to the cost function of linear regression (eq.2).

4.1.2) Lasso Regression LASSO (Least Absolute Shrinkage and Selection Operator) regression [12] cost function can be written as

$$\sum_{i=1}^{r}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{r}\left(y_i - \sum_{j=0}^{c} m_j \cdot x_{ij}\right)^2 + \lambda \sum_{j=0}^{c} |m_j| \text{----------(4)}$$

The above equation (4) is the cost function for Lasso regression. So, coefficients of Lasso regression are similar to ridge Constraints on ridge regression coefficients. If λ=0, then equation 4 becomes equation 2, means Lasso regression becomes like cost function of simple linear regression. The difference between lasso and regression is the magnitude of coefficients. Some of the independent variables are removed from the dataset and select the most significant features for calculating the output. So, the main advantage of Lasso regression is to avoid over fitting and choose the best features.

Results and Discussion

Performance Measure In this section, we study the regression of machine learning algorithms. According to results of lasso, ridge and elastic net of linear models, random forest regress or, gradient descent regress or of ensemble models and decision trees of trees are explained before, and then we compare the results. As stated, in the paper total 4188 instances out of which 80% of data that is 3350 data samples for training and 20% of data that is 838 data samples are chosen for testing purpose. The results in this paper have been taken from test data hat is 838 data samples. The evaluation metrics for regression algorithms are R^2 score, Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

R Squared (R^2 or R^2 score) R^2 tells us, "How well a regression line predicts actual values." R-squared is the proportion of the target variable difference that is described by the linear model. The R-squared value lies between 0 and 100%. If the R-squared value is significant means about 100%, then the model properly fits data. If R-squared value is very fewer means about to '0', then the model not properly fits data and gives the wrong predictions. Here, y is the best fit line values; y is the mean of the actual values.

$$R^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

5.1.2) Mean Absolute Error (MAE) There are many ways of measuring the model's performance. MAE is one of the metrics for brief and evaluating the quality of a machine learning model. The error is calculated in MAE as an average of the absolute difference between the actual values and the predicted values. Where y_i is the real value and $\hat{y}_i$ is the predicted value.

$$\text{MAE} = \frac{1}{r}\sum_{i=1}^{r}|y_i - \hat{y}_i|$$

Analysis of Results

This section analyses the result of the extensive experiment conducted on different Machine learning (ML) algorithms such as Lasso, Ridge, Elastic Net of Linear Models, Random Forest, and Gradient Boosting Regress or of Ensemble Models and

Decision Tree Regress or for rainfall datasets. Table 1 shows the performance measurements of ML solutions on rainfall datasets. From Table 1, the lasso regression model provides better R^2 Score performance.

Table 1.*Regression Performance of six-ML Algorithms*

Models	Train & Test (%)	R^2 Score	MAE	MSE	RMSE
lasso	70-30	98.44	10.80	12673.84	112.57
ridge	70-30	98.48	10.92	12339.40	111.08
enet	70-30	98.42	13.01	12819.53	113.22
rf	70-30	97.87	45.78	17312.12	131.57
gb	70-30	97.97	42.54	16468.26	128.32
dtr	70-30	96.63	83.30	27387.45	165.49
lasso	75-25	98.14	11.30	14847.94	121.85
ridge	75-25	98.18	11.37	14573.74	120.72
enet	75-25	98.11	13.62	15112.67	122.93
rf	75-25	97.37	45.76	21012.06	144.95
GB	75-25	97.84	37.10	17292.54	131.50
dtr	75-25	95.55	88.30	35759.71	189.10
lasso	**80-20**	**99.21**	**13.68**	**6432.41**	**80.20**
ridge	80-20	99.10	16.67	7307.13	85.48
enet	80-20	99.13	15.58	7110.44	84.323
rf	80-20	98.58	45.19	11602.47	107.71
gb	80-20	98.76	40.20	10136.62	100.68
dtr	80-20	95.91	82.50	33521.73	183.08
lasso	90-10	98.65	15.32	11025.61	105.00
ridge	90-10	98.60	17.80	11441.76	106.96
enet	90-10	98.58	16.71	11571.53	107.57
rf	90-10	98.04	47.19	16032.62	126.62
gb	90-10	98.11	41.34	15450.79	124.30
dtr	90-10	96.84	83.05	25817.36	160.67

Here, lasso=Lasso Regression, ridge=Ridge Regression, enter=Elastic Net Regression, fro=Random Forest Regression, gab= Gradient Boosting, dry=Decision Tree Regressor. The comparison of R^2 Score for different ML models is graphically presented in Fig.10. Among six ML models, and Lasso regression model has the highest R^2 Score with 99.21% compared to the remaining ML models at 80% train data set and 20% test data set.

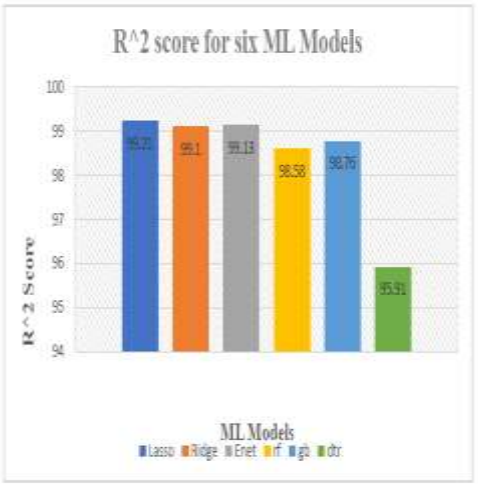

Figure 10. Comparison of R^2 Score (%) for ML Models

The comparison of Mean Absolute Error (MAE) for different ML models is graphically presented in Fig.11. Among six ML models, the Lasso regression model has the lowest MAE value, with 13.68 compare to remaining ML models at 80% train data set and 20% test data set.

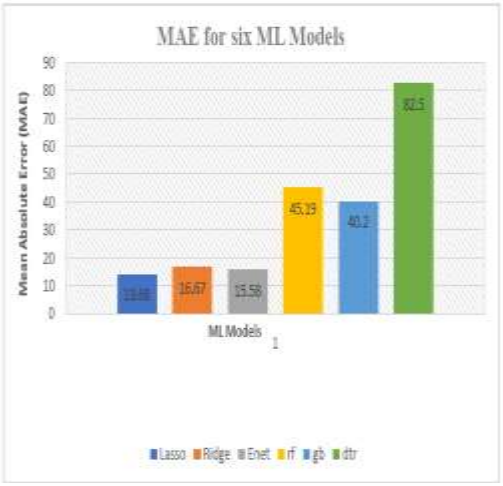

Figure 11. Comparison of MAE for ML Models The comparison of Mean Square Error (MSE) for different ML models is graphically presented in Fig.12. Among

the six ML models, the Lasso regression model has the lowest MSE value with 6432.41 compare to the remaining ML models at 80% train data set and 20% test data set.

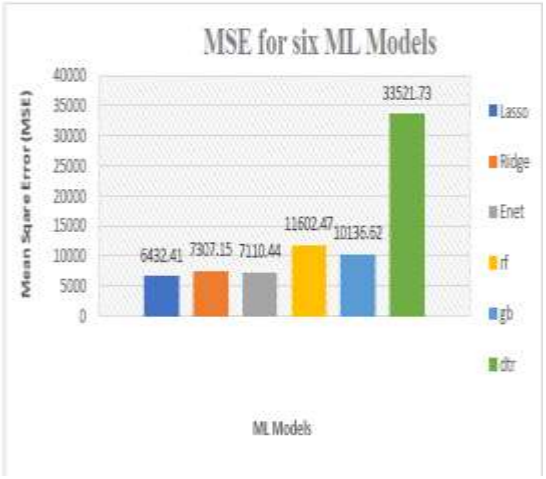

Figure .12. Comparison of MSE for ML Models

The correlation of Root Mean Square Error (RMSE) [z] for different ML models is graphically presented in Fig.13. Among six ML models, the Lasso regression model has the lowest RMSE value, with 80.2 compared to the remaining ML models at 80% train data set and 20% test data set.

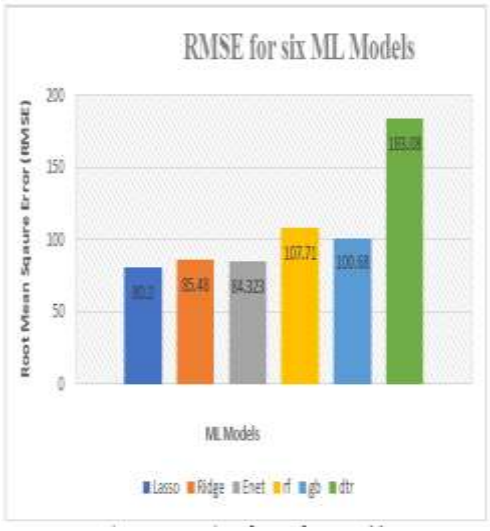

Figure 13. Comparison of RMSE for ML Models

Conclusions and Future Works

Many ML algorithms have been successfully applied for the automatic regression of rainfall. This research paper summarizes and exemplifies the working logic of the six ML algorithms and empirically evaluates the regression performance of all the ML algorithms to the benchmark rainfall dataset. Among the six algorithms, lasso regression got the highest R^2 score of 99.21% at 80-20% of training and validation dataset. Apart from this, the performance of all ML algorithms is evaluated and compared to the actual target values with predicted values. In the feature, we can apply the regression algorithms and improve accuracy.

References

[1]. N. Ghana Ankara, E. Ramada, "A Multiple Linear Regression Model to Predict Rainfall Using Indian Meteorological Data", International Journal of Advanced Science and Technology (IJAST) Vol. 29, No. 8s, (2020), pp. 746-758.

[2]. Irishman Alcntara-Ayala. Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries. Geomorphology, 47(24):107–124, October 2002.

[3]. Neville Nicholls. Atmospheric and Climatic Hazards: Improved Monitoring and

Prediction for Disaster Mitigation. Natural Hazards, 23(2-3):137–155, March 2001

[4]. Puente Sharma and Nadir Chitty, "Machine Learning-Based Modeling of Human Panther Interactions in Ravalli Hills of Southern Rajasthan", Indian Journal of Ecology 46(1): 126-131.

[5]. A.El-shafie, M.Mukhlisin, Ali A. Rajah and M.R. Tasha, "Performance of artificial neural network and regression techniques for rainfall-runoff prediction",

International Journal of the Physical Science vole 6(8), 18 April 2011.

[6]. Nikhil Seth et al., "Exploiting Data Mining Technique for Rainfall Prediction" in International Journal of Computer Science and Information Technologies ISSN: 09759646 Vol. 5 (3), pp. 3982-3984, 2014.

[7]. Ms Ashbin Mondale, Mrs. Jadhawar B.A, "Weather Forecast Prediction: A Data Mining Application", International Journal of Engineering Research and General Science Volume 3, Issue 2, March, April 2015, ISSN 2091-2730.

[8]. J.N.K. Liu, B. N. L. Li, and T. S. Dillon. An improved naive Bayesian classifier technique coupled with a novel input solution method [rainfall prediction]. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, (2):249 –256, 2001.

[9]. https://data.gov.in/resources/sub-divisional-monthly-rainfall-1901-2017.

[10]. Shen Rong, Zhang Bao-wen, The research of regression model in machine learning field MATEC Web of Conferences 176, 01033 (2018).