International Journal of Advanced Science and Technology
Vol. 29, No. 5, (2020), pp. 5727 - 5741

Performance Analysis and Evaluation of Machine Learning Algorithms in Rainfall Prediction

Prem Kumar.B[1], R. Lakshmi[2], Bichitrananda Behera[3]

[1,2,3]Department of Computer Science, Pondicherry University, Pondicherry, India.

[1] *premkumar.jones@gmail.com*, [2] *prof.rlakshmi@gmail.com*

[3]*bbehera19@gmail.com*

Abstract

Massive rainfall forecast is a significant problem for the meteorological department. This paper investigates the performance of the various Machine Learning (ML) models, namely Lasso regression, ridge regression, elastic net regression, random forest, gradient boosting and decision tree regressor. Those models performances have been calculated through the evaluation metrics such as R^2 score, Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). The objective of this study is to compare different machine learning regression algorithms in rainfall dataset. In this analysis, we conclude that the Lasso regression of the linear model is the best model among six ML models. Lasso model given more R^2 score is 99.21%, MAE is 13.68, MSE is 6432.41 and RMSE is 80.20 at 80 % training data set and 20% at test dataset.

Keywords: *Rainfall forecasting, machine learning algorithms, R^2 score, MAE, MSE, RMSE.*

1. 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 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, developingen 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 modelling 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 regressor is R^2, 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.

2. 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 modelling 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 regression 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.

3. 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.

3.1 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.

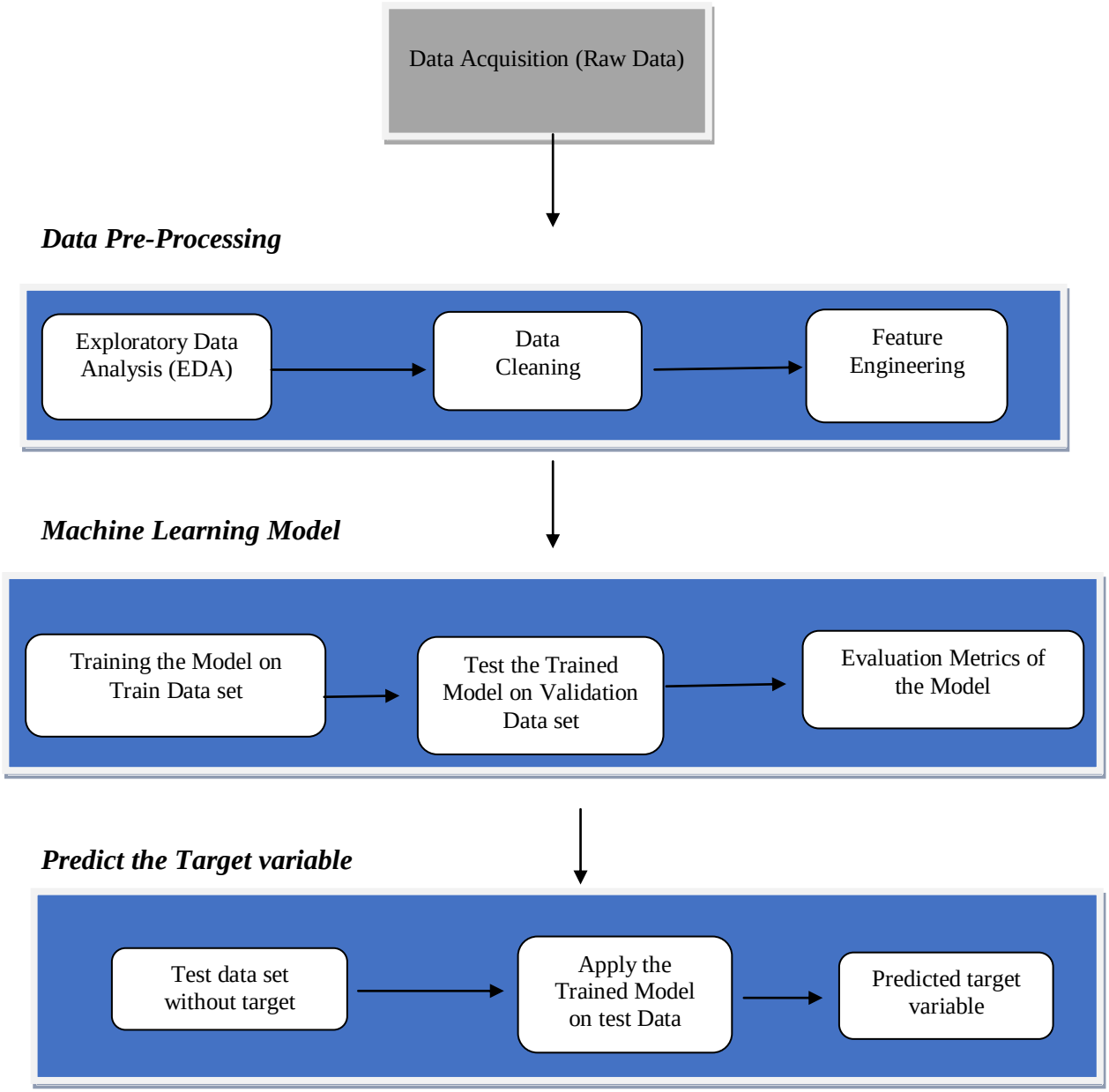

Figure 2.Framework for Machine learning Model

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	SUBDIVISI	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	JF	MAM	JJAS	OND
2	Andaman	1901	49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2	136.3	560.3	1696.3	980.3
3	Andaman	1902	0	159.8	12.2	0	446.1	537.1	228.9	753.7	666.2	197.2	359	160.5	3520.7	159.8	458.3	2185.9	716.7
4	Andaman	1903	12.7	144	0	1	235.1	479.9	728.4	326.7	339	181.2	284.4	225	2957.4	156.7	236.1	1874	690.6
5	Andaman	1904	9.4	14.7	0	202.4	304.5	495.1	502	160.1	820.4	222.2	308.7	40.1	3079.6	24.1	506.9	1977.6	571
6	Andaman	1905	1.3	0	3.3	26.9	279.5	628.7	368.7	330.5	297	260.7	25.4	344.7	2566.7	1.3	309.7	1624.9	630.8
7	Andaman	1906	36.6	0	0	0	556.1	733.3	247.7	320.5	164.3	267.8	128.9	79.2	2534.4	36.6	556.1	1465.8	475.9
8	Andaman	1907	110.7	0	113.3	21.6	616.3	305.2	443.9	377.6	200.4	264.4	648.9	245.6	3347.9	110.7	751.2	1327.1	1158.9
9	Andaman	1908	20.9	85.1	0	29	562	693.6	481.4	699.9	428.8	170.7	208.1	196.9	3576.4	106	591	2303.7	575.7
10	Andaman	1910	26.6	22.7	206.3	89.3	224.5	472.7	264.3	337.4	626.6	208.2	267.3	153.5	2899.4	49.3	520.1	1701	629
11	Andaman	1911	0	8.4	0	122.5	327.3	649	253	187.1	464.5	333.8	94.5	247.1	2687.2	8.4	449.8	1553.6	675.4
12	Andaman	1912	583.7	0.8	0	21.9	140.7	549.8	468.9	370.3	386.2	318.7	117.2	2.3	2960.5	584.5	162.6	1775.2	438.2
13	Andaman	1913	84.8	0.5	1.3	2.5	190.7	530	280.8	205.8	580.1	288.8	133	67.5	2365.8	85.3	194.5	1596.7	489.3
14	Andaman	1914	0	0	0	37.7	298.8	383.3	792.8	520.5	310.8	139.8	184.4	289.7	2957.8	0	336.5	2007.4	613.9
15	Andaman	1915	45	56.7	33.3	40.9	170.2	334.7	269	317.2	429.8	468.1	258.4	318	2741.3	101.7	244.4	1350.7	1044.5
16	Andaman	1916	0	0	0	0.5	487.4	450.1	317.3	425	561.2	369.7	192.6	133.7	2937.5	0	487.9	1753.6	696
17	Andaman	1917	8	3.6	112	4.5	295.9	301.1	394.8	437.4	471.8	238.1	108.3	236.9	2612.4	11.6	412.4	1605.1	583.3
18	Andaman	1918	77.4	6.9	11.4	10.7	729.3	710.8	200.9	455.4	303.3	227	366.9	175	3275	84.3	751.4	1670.4	768.9
19	Andaman	1919	10.2	18	0	35.5	283.9	542.5	246.5	259.8	170.7	186.2	340.4	258.4	2352.1	28.2	319.4	1219.5	785
20	Andaman	1920	122.3	7.4	3.1	13	237.4	546.9	294.4	467.4	505.4	397.5	262.9	85.5	2943.2	129.7	253.5	1814.1	745.9

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 is 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).

Check the Shape (Rows & Columns) of DataFrame

```
df.shape
```

```
3]: (4188, 19)
```

```
print("Rows:",df.shape[0], "Columns:",df.shape[1])
```

```
Rows: 4188 Columns: 19
```

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 st age 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 a pplying machine learning techniques. EDA explains data by means of statistical and visua lization 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 w e notice that Subdivisions with highest annual rainfall are "COASTAL KARNATAKA," " ARUNACHAL PRADESH" and "KONKAN & GOA" with an approximate yearly rainfal l of 3403 mm, 3397 mm and 2987 mm respectively. Subdivisions with the lowest annual r ainfall are "HARYANA DELHI & CHANDIGARH," "SAURASHTRA & KUTCH," and "WEST RAJASTHAN" with an approximate annual rainfall of 528 mm, 496 mm and 29 4 mm respectively.

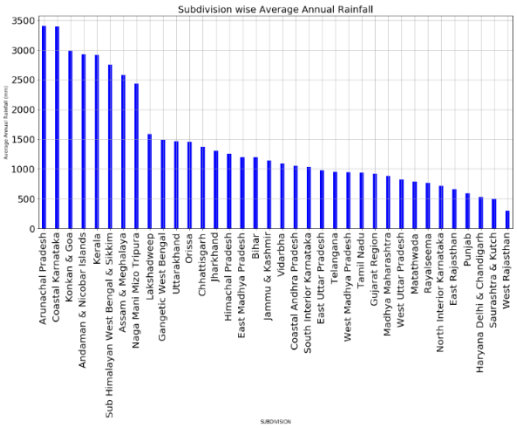

Figure 4. Subdivision wise highest and lowest rainfall

3.2.1.2 Rainfall in Subdivisions

From figure 5, we noticed that, majority of rainfall is received in the months of JUNE, JU LY, AUGUST, SEPTEMBER (JJAS) from Coastal Karnataka, Arunachal Pradesh, Konka n Goa, and Kerala and which are receiving the highest rainfall.

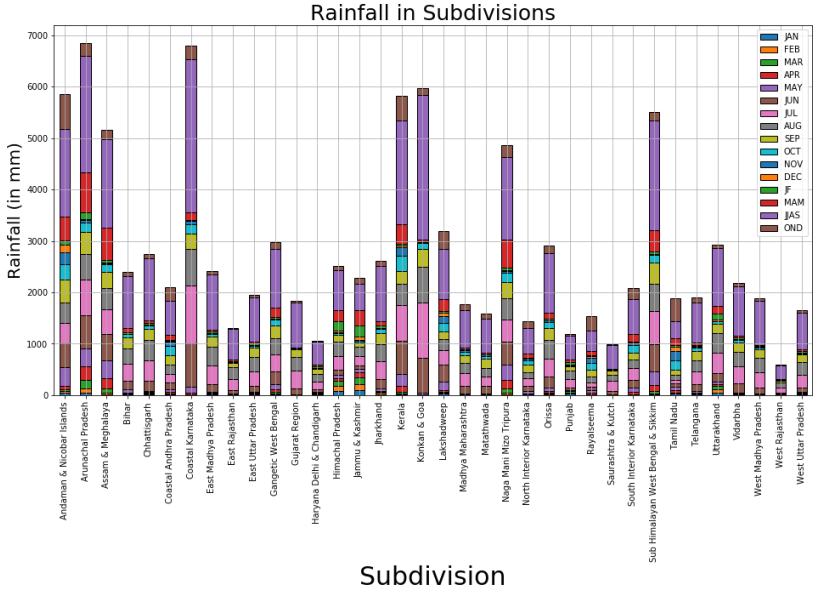

Figure 5. Rainfall in subdivisions of monthly wise

3.2.2 Data Cleaning

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

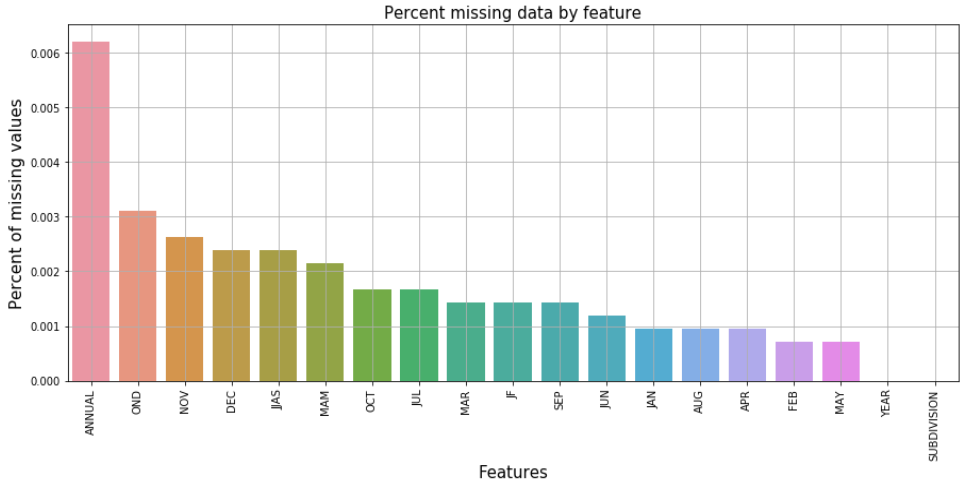

Figure 6. percentage of missing values by features

	Total	Percent
ANNUAL	26	0.006208
OND	13	0.003104
NOV	11	0.002627
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

3.2.3 Feature Engineering

In the feature engineering apply the scaling methods like standard scaler for bringing the a ll data into the same magnitude.

3.3. Build the machine learning model

After pre-processing and feature engineering, that data set had been splitting into train and validation data set at 80-20%. Fit the model on an 80% train dataset. Check the performa nce of the trained model on a 20% validation data set.

3.4. predict the target variable

After training, the model predicts the target variable on a test data set. Test datasets are al ways not having a target variable.

4. Machine Learning Algorithms

Machine learning or the predictive modelling filed is mainly concerned with minimizing the error of a model. All the machine learning models for regression can be categorized into Linear, Ensemble and trees, as shown in fig.8.

Machine Learning (ML) models like Lasso, Ridge, and Elastic Net Regressors, which have their own basic machine learning structure, are known as Linear models. The machine learning models like Random Forest and Gradient Descent Regressors, which is built by assembling basic machine learning models, are called Ensemble Models, and Decision Tree regressor belongs to trees models. The short description of those ML algorithms is explained as follows.

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.

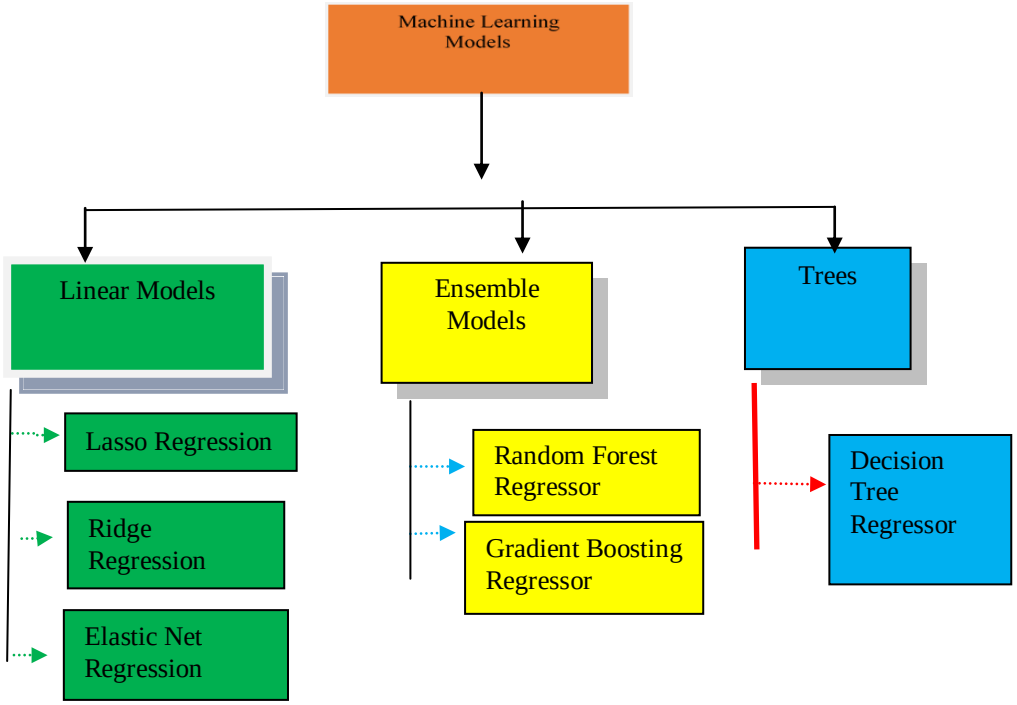

Figure 8. Regression of Machine Learning models

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 $y_i = m_i x_i + b$. 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].

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

The above equation (1) is the Linear Regression model with 'n' number of independent variables. Suppose only one independent feature 'x' with slope 'm', and 'b' will indicate simple linear regression. Linear regression appearances for adjusting the coefficients like 'm' and 'b' . The cost function can be written as

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

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

The cost function for simple linear regression is defined in equation (2) from this equation, assume that there are 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 minimise 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.

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

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

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

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

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

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 overfitting and choose the best features.

4.1.3) Elastic Net Regression

Elastic net regression typically performs fine when we have a large dataset. There are two parameters alpha and L1_ratio. The elastic net is a combination of both L1 and L2 regularization. Apply the Elastic net with both L1 and L2 penalty terms to the linear regression model [13]

$$= \sum_{i=1}^{r} \left(y_i - \sum_{j=0}^{C} m_j \cdot x_{ij} \right)^2 + \dot{\lambda} \sum_{j=0}^{c} |m_j| + \dot{\lambda} \sum_{J=0}^{c} m_j^2 \text{----------(5)}$$

Really, we have alpha=u+v and L1 ratio=u/(u+v). Where 'u' and 'v' are weights assigned to L1 & L2, respectively. So, when we modify the values of alpha and L1 ratio, m & n are set accordingly. Such that, they control the trade-off between L1 and L2 as= u*L1+v*L2.

4.2) Ensemble Models

Ensemble models help to reduce these factors (except noise, which is a fundamental error). Ensemble methods are machine learning algorithms that merge some base models to create one optimal predictive model, which gives a more accurate prediction score and improve the overall performance [14]. A model encompassed of some models like bagging and boosting is known as an Ensemble model.
Bagging: Parallelly to training a group of distinct ML models. Some parts of the dataset had been used to train each ML model.
Boosting: Sequentially, to training a group of different ML models. Each model train from mistakes which were made by the earlier model.

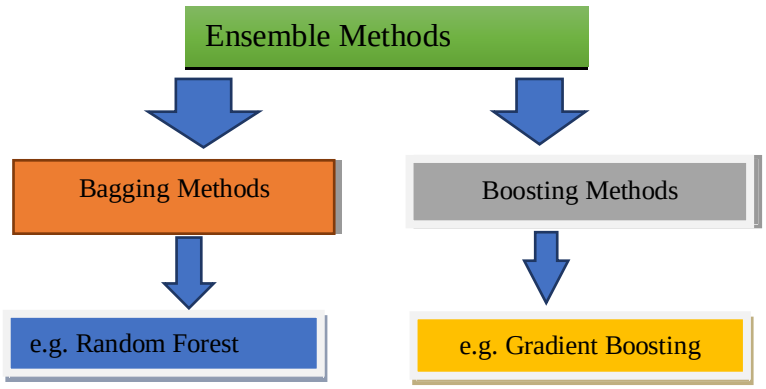

Figure 9. Ensemble Methods

4.2.1) Random Forest Regressor

The set of decision trees is known as random forest which are executed parallel and each decision tree is distinct. Gini index is applied for dividing attributes in the decision tree [15]. There is no interaction among trees while constructing trees. Random forest is combing the results of different predictions. It performs effectively on a vast dataset and control number of input variables without deleting variables. Advantages of Random Forest are, it execute proficiently on huge data sets and It assist an useful method for predicting missing data, [16].

4.2.2) Gradient Boosting

Gradient boosting is a powerful machine-learning technique that have applied in aextensive range of real-world applications [17]. Boosting method to transform weak learners to the dynamic learner. The gradient boosting method can be applied to regression and classification kind of problems.

4.3) Trees

4.3.1) Decision Tree Regressor

Decision tree mimics the human-level thinking. In the decision, each node denotes a feature, each link (branch) denotes decision rule, and each leaf denotes an outcome. There are two types of algorithms to build a decision tree. The first algorithm is CART (Classification And Regression Trees) in which apply the Gini index as a metric. The second type of algorithm is ID3(Iterative Dichotomiser
3), which uses two metrics, Decision tree, which are entropy and information gain [12].

5. Results and Discussion

5.1) 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 regressor, gradient descent regressor 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).

5.1.1) 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, $\hat{y}$ is the best fit line values, $\bar{y}$ 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.

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

5.1.3) Mean Squared Error (MSE)

MSE is the simple metric for regression kind of problems. The following equation defines MSE.

$$\text{MSE}=\frac{1}{r}\sum_{i=1}^{r}(y_i - y_i)^2$$

"Mean squared Error" is the most standard metric applied for regression problems. It mostly calculates the average squared error between the actual and predicted values for each point. If the average squared error values are large value, then the model is worse.

5.1.4) Root Mean Squared Error (RMSE)

RMSE is the square root of MSE. The square root is initiated to make the magnitude of the errors to be the same as the magnitude of targets.

$$\text{RMSE}=\sqrt{MSE} = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(y_i - y_i)^2}$$

5.2) 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 Regressor of Ensemble Models and Decision Tree Regressor 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	95.53	88.30	35759.71	189.10
lasso	**80-20**	**99.21**	**13.68**	**6432.41**	**80.20**
ridge	80-20	99.10	16.67	7307.15	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, enet=Elastic Net Regression, rf=Random Forest Regression, gb= Gradient Boosting, dtr=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.

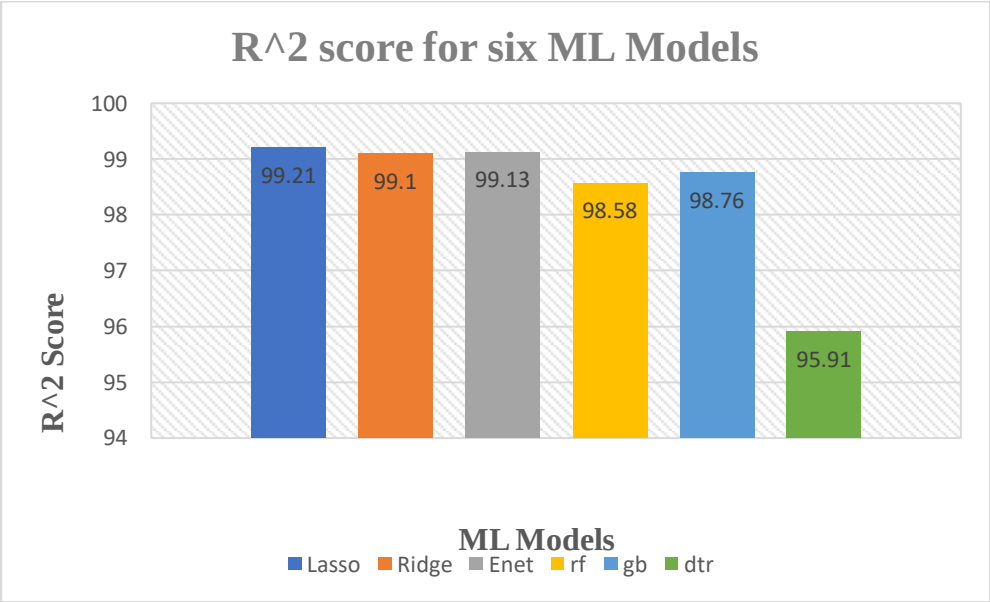

Figure 10. Comparison of R^2 Score (%) for ML Models

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.

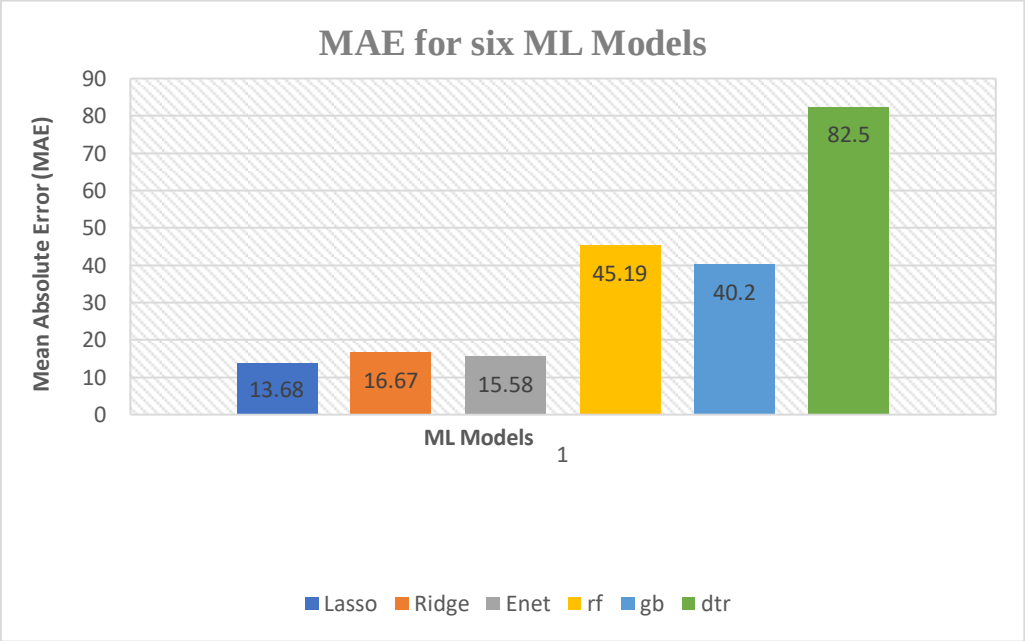

Figure 11. Comparison of MAE for ML Models

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.

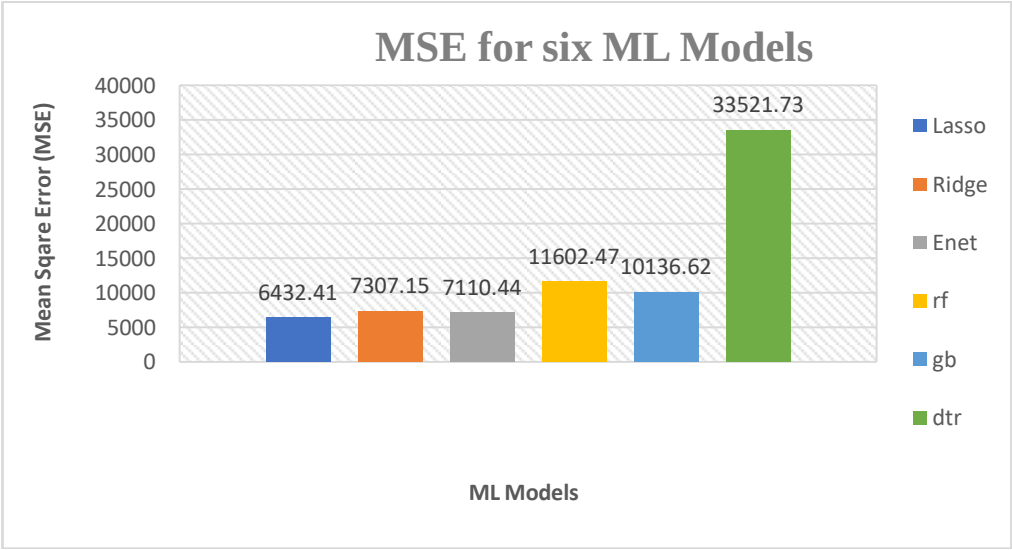

Figure .12. Comparison of MSE for ML Models

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.

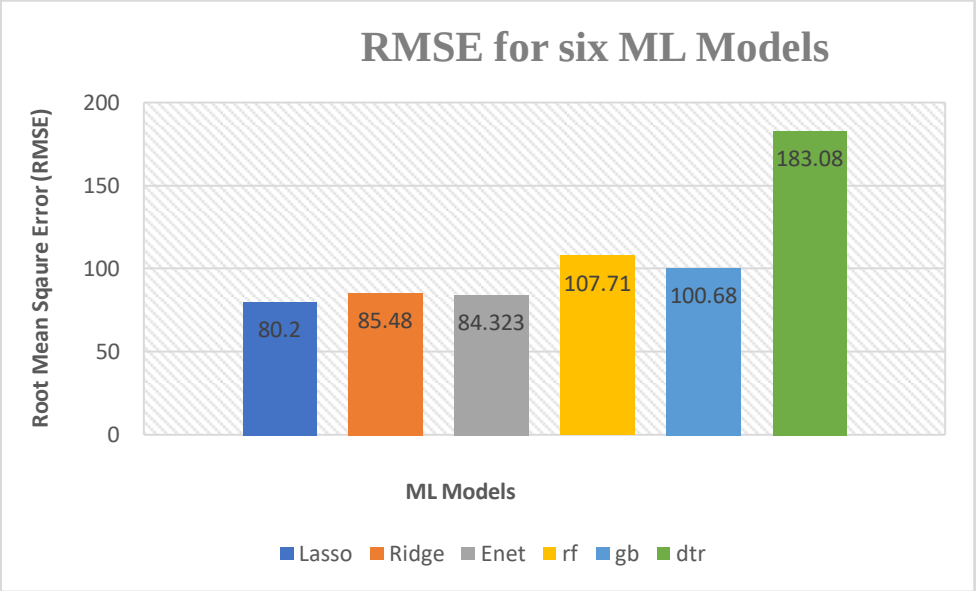

Figure 13. Comparison of RMSE for ML Models

6. 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. Gnana Sankaran, E. Ramaraj, “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]. Irasema 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]. Puneet Sharma and Nadim Chishty, "Machine Learning-Based Modelling of Human Panther Interactions in Aravalli Hills of Southern Rajasthan”, Indian Journal of Ecology 46(1): 126-131.
[5]. A.El-shafie, M.Mukhlisin, Ali A. Najah and M.R. Taha, " Performance of artificial neural network and regression techniques for rainfall-runoff prediction", International Journal of the Physical Science vol 6(8), 18 April 2011.

[6]. Nikhil Sethi 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 Ashwini Mandale, 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).
[11]. https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b.
[12]. sudheer kumar, s.d. attri and k.k. singh," Comparison of Lasso and stepwise Regression technique for wheat yield prediction", Journal of Agrometeorology 21(2): 188-192 (June 2019) 188 Comparison of Lasso and stepwise regres.
[13]. K.Lavanya,2K.Harika,3D. Monica,4K.Sreshta, Additive Tuning Lasso (AT-Lasso): A Proposed Smoothing Regularization technique forShopping Sale Price Prediction, International Journal of Advanced Science and TechnologyVol. 29, No. 05, (2020), pp. 878-886.
[14]. https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725.
[15]. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[16]. Lin Zhua, Dafeng Qiua, Daji Ergua, Cai Yinga, Kuiyi Liub," A study on predicting loan default based on the random forest algorithm", ITQM 2019, Procedia Computer Science 162 (2019) 503–513.
[17]. Alexey Natekin, and Alois Knoll 2013. Gradient boosting machines,atutorial Frontiers in Neurorobotics